SCIENCE CHINA
Information Sciences

• REVIEW •

September 2025, Vol. 68, Iss. 9, 190301:1–190301:73
https://doi.org/10.1007/s11432-024-4408-1

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

Distributed satellite information networks: architecture, enabling technologies, and trends†

Qinyu ZHANG[1,2], Liang XU[1], Jianhao HUANG[1,2], Tao YANG[1], Jian JIAO[1,2*], Ye WANG[2], Yao SHI[1,2], Chiya ZHANG[1,2], Xingjian ZHANG[1,2], Ke ZHANG[2], Yupeng GONG[2], Na DENG[3,4], Nan ZHAO[3], Zhen GAO[5], Shuai WANG[5], Shujun HAN[6], Xiaodong XU[2,7], Li YOU[8,9], Dongming WANG[8,9], Shan JIANG[9], Dixian ZHAO[8,9], Nan ZHANG[10,11], Liujun HU[10,11], Xiongwen HE[12], Yonghui LI[13], Xiqi GAO[8,9] & Xiaohu YOU[8,9]

[1]*Guangdong Provincial Key Laboratory of Aerospace Communication and Networking Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China*
[2]*Pengcheng Laboratory, Shenzhen 518055, China*
[3]*School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China*
[4]*State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China*
[5]*State Key Laboratory of CNS/ATM, Beijing Institute of Technology, Beijing 100081, China*
[6]*National Engineering Laboratory for Mobile Network Technologies, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[7]*State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[8]*National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China*
[9]*Purple Mountain Laboratories, Nanjing 211111, China*
[10]*ZTE Corporation Algorithm Department, Shenzhen 518057, China*
[11]*State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China*
[12]*Beijing Institute of Spacecraft System Engineering, Beijing 100076, China*
[13]*School of Electrical and Information Engineering, The University of Sydney, Sydney NSW 2006, Australia*

Received 16 December 2024/Revised 20 February 2025/Accepted 23 April 2025/Published online 14 August 2025

Abstract Driven by the vision of ubiquitous connectivity and wireless intelligence, the evolution of ultra-dense constellation-based satellite-integrated Internet is underway, now taking preliminary shape. Nevertheless, the entrenched institutional silos and limited, nonrenewable heterogeneous network resources leave current satellite systems struggling to accommodate the escalating demands of next-generation intelligent applications. In this context, the distributed satellite information networks (DSIN), exemplified by the cohesive clustered satellites (CCS) system, have emerged as an innovative architecture, bridging information gaps across diverse satellite systems, such as communication, navigation, and remote sensing, and establishing a unified, open information network paradigm to support resilient space information services. This survey first provides a profound discussion about innovative network architectures of DSIN, encompassing distributed regenerative satellite network architecture, distributed satellite computing network architecture, and reconfigurable satellite formation flying, to enable flexible and scalable communication, computing and control, fundamentally enhancing network resilience. The DSIN faces challenges from network heterogeneity, unpredictable channel dynamics, sparse resources, and decentralized collaboration frameworks. To address these issues, a series of enabling technologies is identified, including channel modeling and estimation, cloud-native distributed MIMO cooperation, new waveform design, grant-free massive access, non-orthogonal multicast, distributed phased array antennas, high-speed inter-satellite communication, network routing, and the proper combination of all these diversity techniques. Furthermore, to heighten the overall resource efficiency, the cross-layer optimization techniques are further developed to meet upper-layer deterministic, adaptive and secure information services requirements. In addition, emerging research directions and new opportunities are highlighted on the way to achieving the DSIN vision.

Keywords distributed satellite information networks, cohesive clustered satellites system, distributed regenerative satellite, network resource virtualization, semantic communications, direct satellite-to-device communications

Citation Zhang Q Y, Xu L, Huang J H, et al. Distributed satellite information networks: architecture, enabling technologies, and trends. Sci China Inf Sci, 2025, 68(9): 190301, https://doi.org/10.1007/s11432-024-4408-1

* Corresponding author (email: jiaojian@hit.edu.cn)
† All authors contributed equally to this work.

1 Introduction

Driven by the space-air-ground integrated network (SAGIN), satellite Internet of Things (S-IoT) and satellite-integrated Internet initiatives, the satellite communication (SatCom) system has witnessed a remarkable evolution, transitioning from a standalone large geostationary satellite serving mode to ultra-dense mega-constellation networks such as Starlink, Xingyun and OneWeb, catering to the vision of seamless global coverage and ubiquitous connectivity in the forthcoming sixth generation (6G) [1]. Nevertheless, as the demand for diverse intelligent applications in 6G integration continues to rise, the issue of information barriers inherent in existing independent satellite networks has become increasingly pronounced, rendering these networks incapable of providing globally open, cohesive, and unified information services. This has raised widespread concern in academia and industry regarding the distributed satellite information networks (DSIN) represented by the cohesive clustered satellites (CCS) system [2,3].

1.1 Limitations of current satellite network

To ensure interoperability and efficient utilization of SatCom, various key technologies such as advanced multi-satellite multi-beam collaborations, software-defined networking (SDN), on-orbit autonomy, on-orbit computing and intelligent resource management have been proposed to achieve the goal of 6G-enabled DSIN. In addition, the standardization efforts for SatCom and non-terrestrial networks (NTN) are gaining momentum, with the active participation of several standardization organizations, including the European Telecommunications Standards Institute (ETSI) and the 3rd Generation Partnership Project (3GPP) [4]. However, current SatCom networks will not meet all the requirements of the future 6G-enabled DSIN. One of the main distinguishing features of the current SatCom systems is the prevalence of using a single satellite to provide emergency communication or Earth observations (EO) services, which often struggle to offer continuous and robust communication services due to their singular nature and susceptibility to regional disruptions. Further, the proliferation of diverse satellite systems, including those dedicated to communication, navigation, and remote sensing, has led to a siloed approach where various departments often construct independent ground stations (GSs) to serve their specific domain applications. This fragmented infrastructure hinders the timely sharing and integrated utilization of massive satellite network business information across different satellite systems. In response, the future of the satellite networking paradigm is pivoting towards the multi-satellite collaborative shifts, i.e., the CCS system, aiming to create a cohesive platform for integrated acquisition, processing, storage, transmission, and distribution of space-based information. Last but not least, the integration of communication, sensing, and computing capabilities for distributed homogeneous or heterogeneous payloads is becoming increasingly critical. Existing satellite networks are grappling with the challenge of limited orbital and spectral resources, which are unable to scale in line with the growing complexity of space missions and the burgeoning demands of 6G-enabled DSIN. The weak functional synergy among diverse payloads, limited perception and computing capabilities, and poor dynamic coordination of heterogeneous resources impede the network from providing task/goal-oriented, information-centric and automation-level intelligent services, which is beyond what current satellite networks can deliver.

The above limitations have spurred the development of DSIN, especially highlighted by the maturation of micro and nanosatellite manufacturing, along with advancements in multi-satellite common orbit control, high-speed inter-satellite communication, and multi-load collaboration technologies. The typical architecture of DSIN encompasses: (1) the satellite constellations of hundreds to thousands of homogeneous satellites lacking inter-satellite control mechanisms, achieving near-real-time large-scale connectivity and seamless integration; (2) the CCS system of dozens to pairs of homogeneous or heterogeneous satellites in close proximity to form a virtual satellite, built upon orbital control, self-organizing networking and payload synergy technologies, tailored towards various complex space missions to provide flexible, reconfigurable, and resilient space-based information services [2,3]. These innovative DSIN architectures have significantly enhanced the self-organization and self-healing capabilities of current satellite networks, which have been advanced in notable research projects around the world, i.e., NASA's Earth Observing-1, SpaceX's Starlink, ONION funded by European Union's Horizon 2020, and the fast, flexible, fractionated, free-flying (F6) launched by the Defense Advanced Research Projects Agency (DARPA) [5]. In this regard, the development of DSIN is no longer a mere concept but an inevitable trend in the evolution of space information technology and 6G networks.

1.2 Motivations and contributions

The DSIN is anticipated to provide cohesive sharing of heterogeneous resources and ultra-autonomous multi-satellite collaborative networking to handle complex on-orbit operations with unprecedented flexibility and scalability. To achieve this, the following technologies are crucial and will be the main focus of our survey.

(1) Innovative network architectures are imperative. Among these, current satellite networks, predominantly utilizing transparent satellite mode, reveal inherent limitations in adaptability and operational flexibility. In contrast, advancing research into distributed regenerative satellite network architecture, particularly within DSIN in our survey, is crucial for unlocking enhanced scalability and dynamic traffic management, thus laying the groundwork for next-generation robust and adaptable space information infrastructures. Nevertheless, this distributed regenerative network architecture poses significant challenges in efficiently scheduling computational resources across widely dispersed satellites and in facilitating effective collaboration among them. In this context, the distributed satellite computing network architecture emerges as a promising, pivotal technology, poised to address these challenges and fully unlock the potential of regenerative satellite networks. Moreover, the task-driven, high-volume data generated within DSIN often necessitate dependable satellite formation control and inter-satellite communication amidst complex and variable spatial environments to ensure continuous, deterministic information delivery. Thus, reconfigurable satellite formation flying and collaborative control mechanisms are indispensable for ensuring the success of various formation-flying missions and are discussed in this survey.

(2) New air interface and transmission technologies are essential for achieving high spectrum efficiency and energy efficiency. This includes advancements in channel modeling and estimation, cloud-native distributed MIMO cooperation and signal processing, new waveforms, multiple access approaches, multicasting mechanisms, channel coding schemes, phased array antennas, erasure transmission, high-speed inter-satellite communication, network routing, and the proper combination of all these diversity techniques. Further, another key area of focus is the development of cross-layer optimization techniques that leverage the dynamic nature of satellite networks, the characteristics of DSIN, and the diverse needs of various intelligent services. These techniques, including mobility management, resource management, secure communications and testbeds, are crucial for guaranteeing high-deterministic, energy-efficient, ultra-secure and verifiable on-demand services.

(3) A suite of emerging research directions and challenges toward realizing the vision of DSIN are discussed in this survey. To provide adaptive resource utilization across dynamically changing network environments, network resource virtualization (NRV) technologies combined with resource pooling will be fully exploited. Distributed AI inference and on-orbit information processing technologies will be efficiently combined with multi-satellite collaboration to achieve shared learning, fast inference and optimal decision-making, thus sustaining deterministic and continuous communication in the CCS system. Moreover, to support the emergent paradigmatic transition from bit-level to semantic-level communication, semantic communications are anticipated to engender intelligent and concise information-centric services for DSIN. Further, goal-oriented integrated sensing, communication, and computation framework holds the potential to redefine how we utilize limited heterogeneous space-based resources to provide on-demand services for diverse tasks, bridging gaps between sensing, communication, and computation mechanisms to create autonomous and resilient DSIN. Last but not least, the emerging direction is direct satellite-to-device communications, which represents a fundamental milestone in the realization of an integrated satellite-terrestrial framework within DSIN. By employing innovative on-orbit base station (BS) architecture and sophisticated phased array technology, this approach harmoniously merges SatCom with terrestrial mobile networks and extends a spectrum of communication services such as voice, messaging, and broadband Internet access.

The organization of this survey is shown in Figure 1. The limitations of existing satellite networks and the requirements of future DSIN are introduced in Section 1. In Section 2, the new DSIN network architectures are presented. The enabling technologies, including the air interface and transmission technologies, are given in Section 3. The cross-layer optimization techniques are presented in Section 4. The emerging directions are presented in Section 5. The conclusion is drawn in Section 6. The abbreviations in this survey can be found in Appendix A.

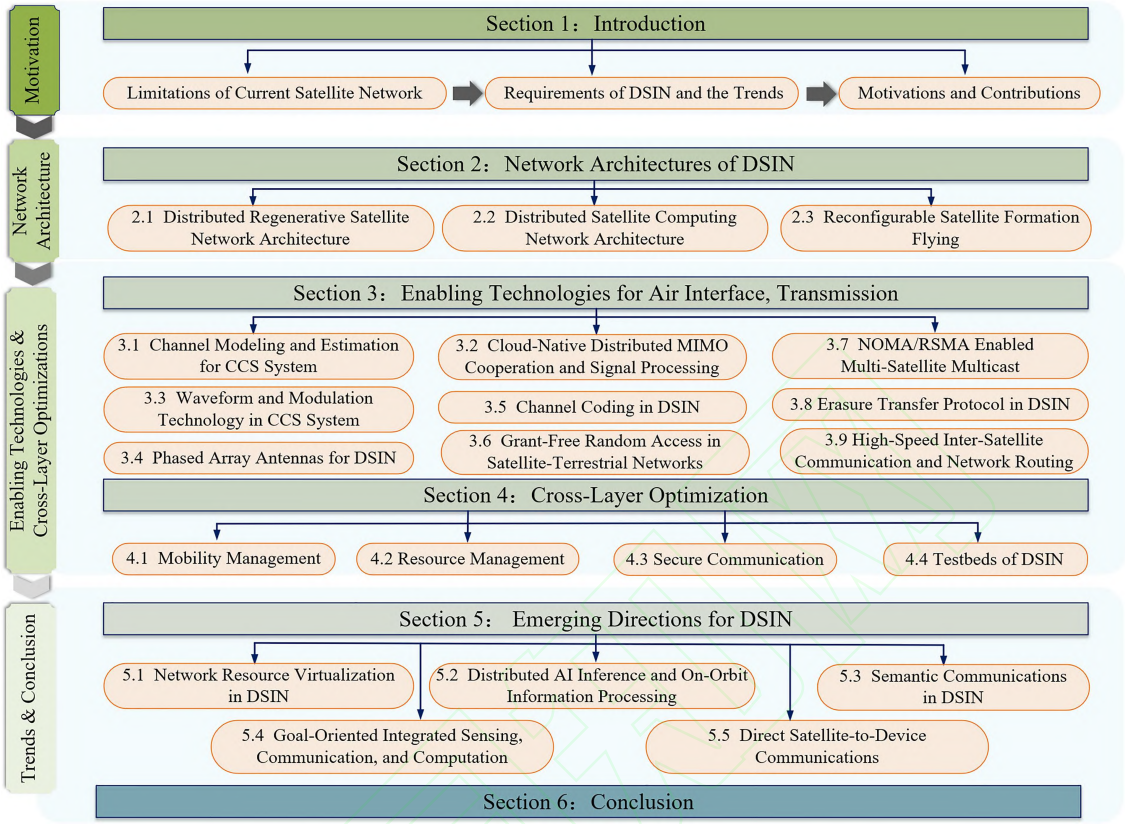

Figure 1 (Color online) Organization of this survey.

2 Network architecture

To better realize the requirements of DSIN, novel network architectures are studied and applied according to network characteristics and specifications in this section, including the distributed regenerative network architecture, the distributed satellite computing network architecture and reconfigurable satellite formation flying.

2.1 Distributed regenerative satellite network architecture

2.1.1 *From transparent to controllable/regenerative payload architecture*

The satellite with a regenerative payload on board is an emerging trend for space-borne communication. Various system architectures are investigated to enable the flexible network topology (e.g., centralized or distributed) for advanced satellite networks.

Architecture-1: from transparent to controllable satellite. In legacy satellite networks, given the restriction of the satellite platform, the transparent payload for communication is widely deployed, especially for the GEO and MEO systems with the following restrictions, e.g.,: (1) limited processing capability on satellites, which is usually dedicated to satellite control instead of communication payload; (2) the controlling of RF components, e.g., beam and frequency reuse mode, is limited and via the dedicated feeder link along with other information, e.g., satellite movement/gesture adjustment. Considering the hardware restrictions of satellites, the regenerative satellite payload may have a partial or full BS function onboard during its evolution. Although the satellites with higher capability or lower orbits are popular for the construction of the new satellite networks, re-framing the legacy on-orbit system is still essential considering the additional aspects, e.g., cost and capability of satellite manufacture.

As shown in Figure 2, the concept of network controllable repeater (NCR) [6] is introduced to enable the evolution of legacy systems with limited processing capability onboard. For example, the legacy transparent satellite (i.e., NCR-Fwd), still only implements frequency conversion, a radio frequency (RF) amplifier, and beam hopping in both up and down links. However, additional control information will be received by the NCR-MT via the control link from the ground BS. In this way, the coordination direction of

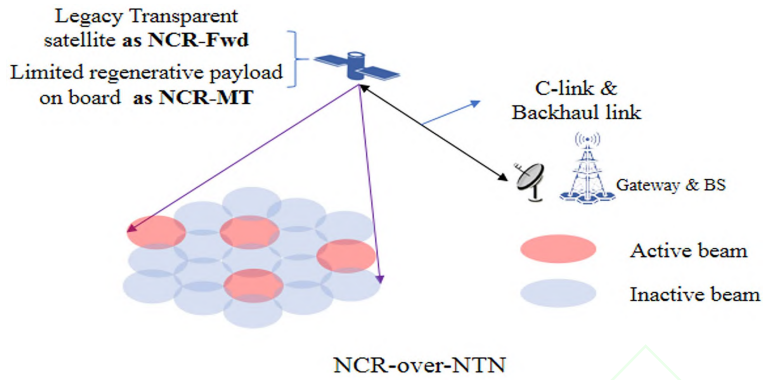

Figure 2 (Color online) Illustration of controllable payload as evolution of transparent satellite.

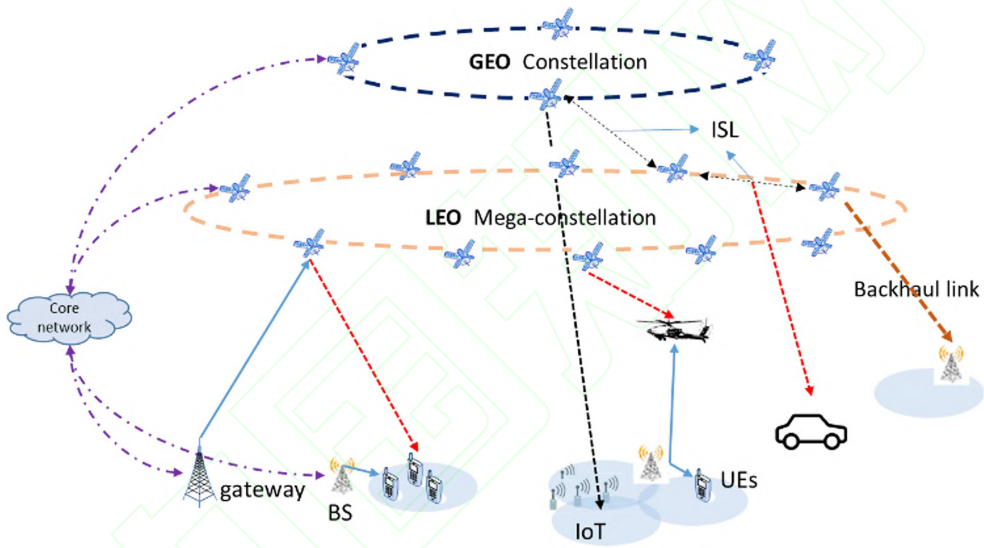

Figure 3 (Color online) Illustration of flexible network topology based on the regenerative payload.

the communication system can be well achieved via the centralized control center to optimize the network performance according to the dynamic traffic and scheduling per UEs/Beam footprint. Moreover, this architecture is more friendly to enable the joint deployment between SatCom and cellular systems, e.g., frequency sharing via dynamic turn-on/off the beam or updating the frequency used for each beam.

Architecture-2: from regenerative payload to full-gNB/CORE onboard. Along with the development of mega constellation for LEO satellites, the regenerative satellite has evolved to support more functionalities onboard, which enables more flexible topologies, e.g., CU-DU, full BS onboard with inter-satellite links (ISL) or CORE network onboard.

As the exemplified network shown in Figure 3, with more capable satellites, the system can enable additional applications. For example, with the ISL, the distributed satellites in DSIN can be deployed to enable complicated coordination onboard. For instance, enabling certain satellites to function as computation nodes for processing information, such as images, is expected to further reduce latency and alleviate traffic loads during information exchange between satellites and gateways.

Additionally, in [7], a group of satellites work together to provide services to the UEs with high reliability since multiple satellites in the CCS system can provide the service for the same area, along with backup in case of a satellite failure. Furthermore, several small and lightweight satellites in the CCS system, e.g., CubeSats, equipped with a commercial low gain patch antenna are tuned to create a large equivalent aperture providing a huge gain and a narrow beam. The CCS system can be arranged in a free-flying configuration (i.e., wireless connected) or tethered configuration (i.e., optically connected). The simulation results presented in [8] demonstrate that the performance is comparable with the classical and distributed implementations of antenna arrays.

2.1.2 *CU-DU split network architecture*

The ongoing work on 3GPP Rel-19 marks a milestone in the integration of satellite technologies into 5G networks as defined by 3GPP [9], particularly with the inclusion of NTN that feature regenerative, or packet-processing payloads. The regenerative satellite-based payloads offer greater flexibility and performance by hosting partial or complete functionality of a gNB, an NR BS, and supporting the Xn inter-gNB interface for ISL [10]. Nevertheless, due to the limitations in payload, power supply, and heat dissipation on satellite platforms, the processing capability of a single satellite is limited [11]. To deal with this issue, it is crucial to research and develop a distributed, open and intelligent DSIN architecture that can provide more efficient, flexible and smart solutions for a variety of applications, including worldwide communication, electromagnetic spectrum detection, remote sensing applications, and a multitude of other critical missions. As a promising option, the integration of open radio access networks (O-RAN) with regenerative payloads aligns with the distributed architecture that separates the central unit (CU) and distributed unit (DU) [12]. As a result, a cutting-edge CU-DU split [13] network architecture can be implemented. It originates from terrestrial 5G networks, where the CU is tasked with managing non-real-time protocols and services, including radio resource control (RRC) and packet data convergence protocol (PDCP). Typically situated in data centers or regional hubs, the CU facilitates efficient resource management and coordination. Meanwhile, the DU is charged with real-time processing of wireless access layer protocols, such as medium access control (MAC) and radio link control (RLC). It is connected to the radio unit (RU), often located in close proximity or co-located with the RU to minimize latency. The CU-DU Split not only fosters the sharing of baseband resources but also paves the way for the slicing and cloudification of wireless access. This innovative approach ensures effective collaboration between sites in complex DSIN. In addition, this separation is pivotal for unlocking the full potential of 6G networks in space, where the NTN gateway at the end of the feeder link can function as a router to the core network. The ISLs are indispensable for regenerative payloads, particularly in non-geostationary satellite orbits (NGSO) such as LEO and MEO, where the satellites move significantly faster than the Earth's rotation. This rapid movement necessitates seamless handovers of the feeder link to alternative gateways and the maintenance of service continuity by different space-borne platforms.

For satellite systems, the satellite centralized unit (SAT-CU) serves as the centralized processing unit, enabling efficient sharing of processing capabilities. By centralizing resource management and complex data processing in the SAT-CU, the system can dynamically allocate resources and quickly adjust task configurations to meet changing demands, while coordinating the collaborative work of multiple satellite distributed units (SAT-DU) in the CCS system to enhance the overall efficiency of DSIN. The SAT-CU, in its role as the control plane and user data anchor, can reduce data interruption delays caused by handovers, further improving the user experience. Meanwhile, the SAT-DU is responsible for preliminary data processing and filtering with high real-time requirements, reducing data transmission delays and improving response speed. Since SAT-DU only needs to provide RF and limited baseband processing capabilities, with user data being collected and processed centrally by SAT-CU without the need for independent core network connections, the weight, power consumption, and complexity of SAT-DU satellites can be effectively reduced, thereby further reducing the comprehensive cost of satellite network construction and deployment. This architecture also improves the robustness and reliability of the CCS system, better isolates and handles faults, and reduces the risk of single-point failures.

(1) SAT-CU/SAT-DU split options. When it comes to delineating the roles of the SAT-CU and SAT-DU, it is essential to explore different CU/DU functional partitioning methods tailored to diverse use cases. Factors to consider include network topology and coverage, latency and bandwidth requirements, computing and storage resources, network load and traffic patterns, security and privacy, cost and complexity, and the extent of collaboration (such as joint scheduling, joint reception, and joint transmission). Specific partitioning can refer to 3GPP TS 38.401 [14] and TR 38.801 [15]. 3GPP provides 8 functional split options as shown in Figure 4, varying based on the distribution of responsibilities between the CU and DU. Here is an overview of the common CU-DU split options.

• High-layer split. In high-layer split configurations, the CU manages upper-layer functions such as RRC and PDCP, while the DU oversees lower-layer tasks like RLC, MAC, and physical (PHY) layers. This setup effectively centralizes key control functions within CU, reducing the processing requirements at DU. High-layer splits are advantageous for scenarios where the DU has limited processing power, or when centralized control and streamlined maintenance are priorities. However, the reliance on a centralized control structure can lead to increased latency, making high-layer splits less suitable for applications

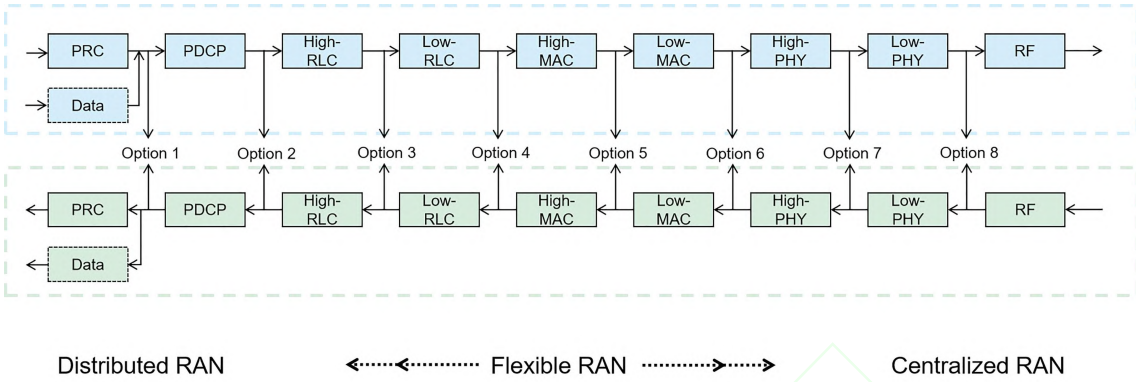

Figure 4 (Color online) CU-DU split options.

that require real-time responsiveness. Additionally, this configuration may impose significant bandwidth demands on the fronthaul link, potentially impacting efficiency in bandwidth-constrained environments.

• Mid-layer split. Mid-layer splits, which divide RLC functions between the CU and DU, offer a balanced approach to functional partitioning. By handling the PDCP in CU while assigning RLC, MAC, and PHY to DU, mid-layer splits manage latency more effectively than high-layer splits, while maintaining efficient control and resource allocation. Certain mid-layer variations, such as Options 3a and 3x, further refine the division of RLC tasks, with retransmission functionalities assigned differently to meet specific performance objectives. This distribution provides flexibility in controlling retransmission and segmentation operations, improving error handling close to the user end. While mid-layer splits can optimize latency and processing, they introduce complexity in managing distributed RLC functionalities, requiring careful coordination to ensure seamless operation. The variable processing load across CU and DU in dynamic DSIN conditions also adds to the challenges of mid-layer configurations.

• Low-layer split. Low-layer splits assign the most high-layer functionalities, such as MAC scheduling and forward error correction (FEC), to the DU, effectively decentralizing processing tasks. By placing these real-time functions closer to the user, low-layer splits significantly reduce end-to-end latency, making them ideal for latency-sensitive applications. For instance, mission-critical applications such as satellite-based disaster response and real-time navigation services benefit from this approach, as it allows immediate data processing at DU, minimizing the end-to-end latency. Additionally, low-layer splits facilitate scalability, enabling decentralized processing across broad coverage areas. However, these configurations may require DUs with substantial processing power and advanced capabilities, increasing deployment costs and power consumption. The power constraints are a concern in satellite networks, especially in remote locations where low-layer splits could be less feasible.

Selecting the optimal SAT-CU/SAT-DU split configuration requires a nuanced understanding of the tradeoffs between latency, bandwidth, and processing resources [16]. High-layer splits are often well-suited for S-IoT applications where low-cost DU units can efficiently manage large-scale device connectivity without the need for real-time response. In contrast, mobile satellite services, which require a balance between latency and data integrity, may benefit from mid-layer splits. By distributing control and error-correction functions across the CU and DU, mid-layer configurations can offer enhanced reliability and efficiency, making them a practical choice for applications where moderate latency and bandwidth demands coexist. For critical applications requiring near-instantaneous response, low-layer splits present the most viable option. By handling MAC and FEC functionalities in the DU, low-layer splits provide the freshness information needed for mission-critical applications, albeit with higher infrastructure demands.

(2) Interfaces. In future DSIN architecture, seamless communication between network elements is crucial to achieving efficient data flow and resource management. High flexibility in positioning network elements introduces challenges in interface design and standardization. The F1 and E1 interfaces, which connect the CU and DU, play a central role in managing these connections. The F1 interface is divided into F1-C (control plane) and F1-U (user plane) components, allowing the CU and DU to manage control signalling and data transmission independently. This division ensures efficient separation of control and data functions, enabling adaptive scaling of control and user data processing according to network conditions.

The E1 interface, connecting CU-CP and CU-UP, allows for further modularization within the CU. By

enabling control and user plane functions to optimize independently, the E1 interface ensures flexibility in managing network signalling and data flow, enhancing resource utilization. Designing these interfaces to handle transmission delays between satellite and ground elements is critical. Given that satellite networks often experience varying levels of latency and intermittent network coverage, interface protocols must accommodate these dynamics while ensuring high interoperability and compatibility between components. The 3GPP standards provide a baseline for interface design in CU/DU split systems, promoting consistency in functionality and data transmission even as network elements vary.

In conclusion, by employing functional, service-based, and physical splits, network operators can optimize latency, bandwidth, and processing resources based on specific service requirements. The design of interfaces such as F1 and E1 further supports seamless integration, ensuring compatibility across distributed elements and enhancing the overall performance of DSIN. As SatCom continues to evolve, the SAT-CU/SAT-DU architecture will play a crucial role in enabling high-performance, adaptable satellite networks capable of meeting the growing demands of global connectivity. Future research could explore adaptive split configurations that respond dynamically to changes in network conditions, pushing the boundaries of efficiency and responsiveness in DSIN.

2.2 Distributed satellite computing network architecture

With the development of satellite technology, the available computing, communication, sensing and storage resources of satellites are increasingly abundant. To reduce the data transmitted back to the ground for processing and improve information delivery efficiency, particularly in tasks such as pre-processing and target recognition of remote sensing images, the requirements of intensive computation tasks executed at satellites have become more urgent. The consensus among industry and academics is that by leveraging the regenerative capabilities of satellites, deploying computational power nodes on satellites represents a promising solution [17]. In this context, distributed satellite computing has become a common approach for onboard large-scale and massive computation tasks, offering benefits such as high reliability, fault tolerance, scalability, and fast computation. However, distributed satellite computing faces significant challenges. On the one hand, the computing nodes in the distributed satellite network must exchange numerous intermediate results with each other to compute the final result, which greatly increases communication overhead, especially in high-dynamics, high-mobility LEO satellite networks. On the other hand, distributed satellite computing is executed by a large number of computing nodes, which may have varying computing and networking resources. As a result, there can be straggling nodes-computing nodes that run slower than others, unintentionally increasing the overall time required to complete the computing tasks. To address these challenges, the concept of coded distributed computing, which combines coding techniques and distributed computing, has been proposed by [18] and has recently garnered significant attention. By employing coding-theoretic techniques, the distributed computing framework encodes the subtasks for the computing nodes, enabling the master node to recover the final result from partially finished nodes, thus mitigating the effects of stragglers [19]. Furthermore, coded distributed computing is highly compatible with NFV, mobile edge computing [20], and collaborative deep reinforcement learning (DRL) [21], which facilitates unified resource and service management within SAGIN. The network topology in DSIN is characterised by strong dynamic properties, with asymmetric node computing capabilities, making it difficult to apply AI methods designed for terrestrial networks to DSIN. To address this, Ref. [21] proposed a collaborative DRL paradigm for satellite-terrestrial networks to achieve intelligent multi-orbit spectrum and computing resource management. This paradigm employs a DRL algorithm combining graph convolutional networks and federated learning (FL) to analyze and infer network topology features, guiding the communication and computing resource management of distributed satellites. This enables efficient adaptive computing offloading while avoiding model retraining due to node mobility or failures. Additionally, the integration of data centers into satellite networks has also attracted attention [3] for achieving low latency, flexible networking, and efficient resource utilisation. Moreover, by deeply integrating cloud computing, the cloud-native satellite computing network can shape DSIN into a distributed, elastic, and horizontally scalable satellite system composed of interrelated microservices, isolating state in a minimal number of stateful components [22]. However, how to efficiently schedule the heterogeneous resources at distributed satellites and make those satellites located in a wide area collaborative is still an urgent problem that needs to be solved.

To solve the above issues, we propose a distributed satellite computing network architecture as shown in Figure 1. Exploiting the characteristics of GEO, MEO, LEO, and terminals, we design a cloud-

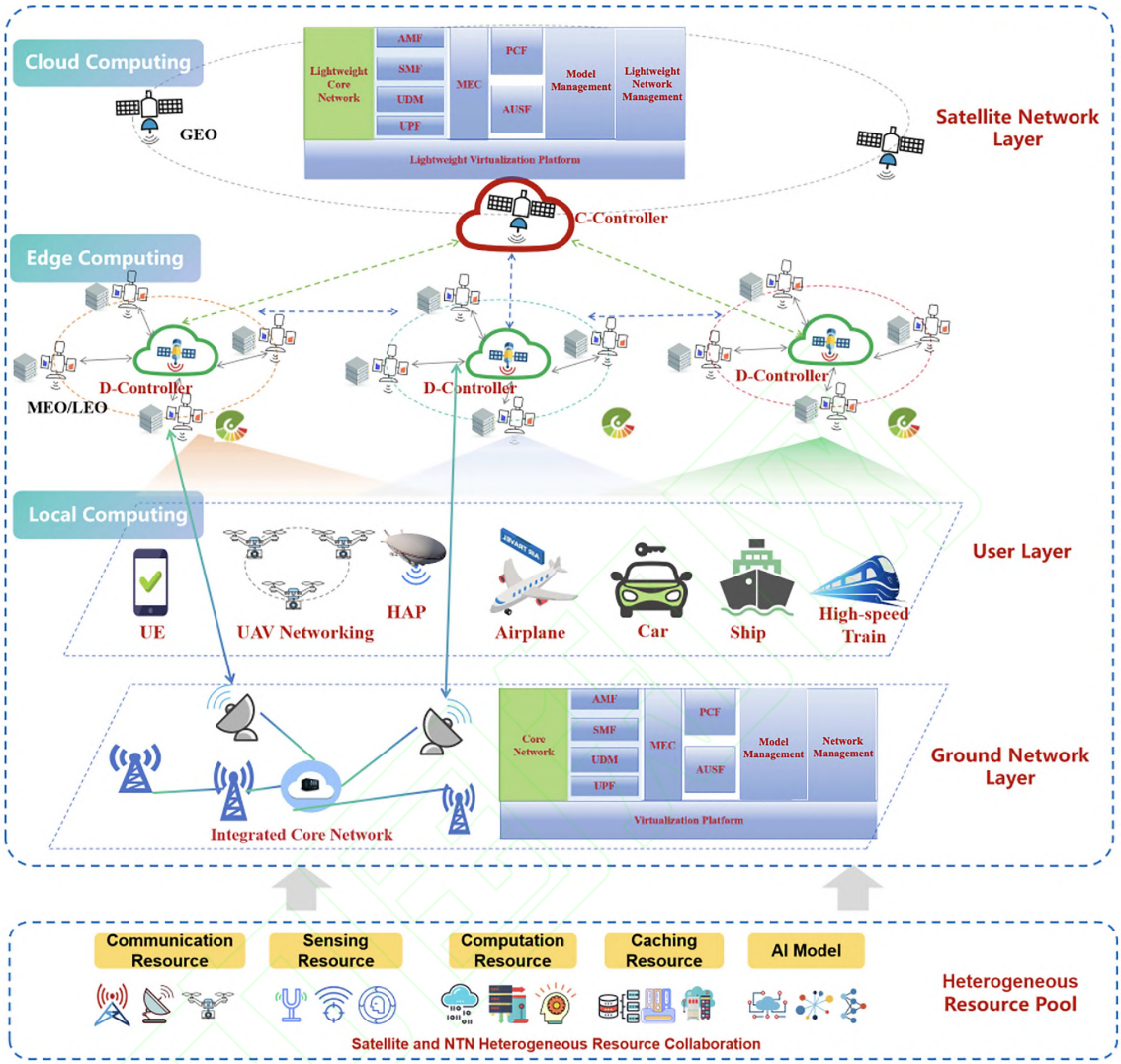

Figure 5 (Color online) Distributed satellite computing network architecture for DSIN.

edge-local collaborative computing architecture. In the satellite network layer, the GEO provides cloud computing services due to its abundant computing power and stability, MEO and LEO provide edge computing services with their advantages in quantity and distance to the Earth's surface. The terminals in the user layer provide local computing services using their various computing power capabilities. To enhance network performance through LEOs' cooperative transmissions, satellite clusters are proposed in [23]. Inspired by [23], we consider that some MEOs/LEOs are forming a CCS system, in which the leader satellite plays the role of the distributed controller (D-controller) in the MEO/LEO satellite layer, while the GEO works as the center controller (C-controller), and the follower satellites cooperate with transmission and computation. To make the onboard service forwarding or migrate between different satellites, deploying the CORE network function on the satellite network layer is a promising solution. Therefore, we consider an integrated core network function to be entirely deployed at the ground networks layer, and the lightweight core network function is deployed at GEO in the satellite network layer as shown in Figure 5.

2.2.1 *Key technologies of satellite computing deployment*

• Lightweight virtualization technology of satellite. Virtualization technology abstracts core network functions from distributed satellite hardware, allowing these generated core network functions to operate independently of physical infrastructure. Lightweight techniques such as network element function trimming, function fusion and reconstruction, and interface protocols are lightweight and employed to

achieve satellite-based lightweight core networks [24]. The deployment set of satellite network element functions can include access and mobility management function (AMF), session management function (SMF), unified data management (UDM), user plane function (UPF), policy control function (PCF), and authentication server function (AUSF). Moreover, the functions of MEC, model management, and lightweight network management are considered to better process intelligent tasks. Due to the rapid changes in satellite-ground topology, satellite-based computing nodes face significant time delays when making decisions. Integrating the computational resources of satellite-based distributed nodes through virtualization technology, computation tasks in the network are split into several microservices, and then executed on the computing units of the CCS system, efficiently reducing the computational burden on the individual satellites.

• Edge computing function of satellite division strategy. Structured functional division of satellite onboard edge computing is essential for optimizing space-based data processing capabilities, enabling modular and collaborative task execution in satellite edge computing. To improve the capability of satellite in-orbit service and to organize the intelligent satellites' edge computing nodes, the functional division of onboard edge computing is strategically organized into a layered architecture, and segmented into 4 distinct layers [4, 17]: resource layer, virtual abstraction layer, system service layer, and application layer. In the 4-layer architecture, services and applications with varying requirements demand close collaboration among system components [17]. Vertical collaboration involves coordination across hierarchical levels, harnessing the unique capabilities of each to optimize resources and enhance system performance. Horizontal collaboration facilitates the distribution of computing tasks and service composition among computing modules at the same level, which is essential in multi-service scenarios or computing fusion. Integrated collaboration combines both vertical and horizontal approaches, integrating computing resources, data, and services across layers and entities. This comprehensive model is crucial for achieving maximum efficiency and functionality in DSIN, and enabling the development of more robust and intelligent satellite computing power technologies [25].

• Onboard intelligent fusion of multi-source satellite. With the evolution of satellite technology, the received EO data are more precise and have a wider range of information with improved spatial and temporal resolution. Multi-source satellite data fusion can enable effective complementarity to EO data, eliminate conflicts and uncertainties between data, and obtain more accurate and reliable information than a single data source. Pre-processing, feature extraction, and fusion algorithm are three key steps critical for achieving high-quality fusion results. Based on the information abstraction level, the fusion algorithm can be classified into three levels: data, features, and decisions [26]. Meanwhile, AI has the potential to revolutionize the onboard information fusion processing method. Currently, AI-based methods have been proposed to correlate data from different sensors, DL has become increasingly prevalent in feature extraction and fusion algorithms, and FL has also been used to improve the performance of satellite data in-orbit fusion [27]. These AI-based algorithms adapt their strategies based on feedback, offering a dynamic approach to data fusion that is sensitive to the nuances of multi-source satellite data.

2.2.2 *Distributed computing power collaboration*

• Computing task migration mechanism. The collaboration of satellite and ground computing power can better assist in completing large-scale data processing tasks. On the one hand, in satellite remote sensing data processing, utilizing onboard computing power for preliminary data processing can reduce the amount of data that needs to be transmitted to ground stations. On the other hand, by employing task migration between the satellite network layer and the user layer, it becomes feasible to offload tasks from remote terrestrial devices to satellites. However, the limited payload capacity of satellites poses constraints on onboard computational resources, making it challenging to handle computationally intensive tasks effectively. The computing task migration mechanism mainly includes cloud-edge, edge-local, and edge-edge collaborations. In addition, imbalances in workloads among satellites result in higher queuing delays for tasks on heavily loaded satellites, while under-utilization of computing resources occurs on lightly loaded satellites. The evolution of large-scale satellite constellations has emphasized the importance of collaborative task processing among multiple satellites to formulate a CCS system to address the inadequacies in individual satellite computational resources [28]. Satellites engaged in satellite-ground computational migration not only participate in collaborative computations across the CCS system but also serve as pivotal scheduling decision units for task scheduling and resource allocation. Migrating tasks with high queuing delays to less loaded satellites benefits load balancing. Tasks partially

offloaded from devices can be seamlessly transferred within CCS systems, thereby enabling a more flexible task migration approach.

• Multi-dimensional and heterogeneous resource collaborative management. Currently, the independent networking of satellites and ground networks results in low flexibility in information interaction and significant differences in network characteristics, failing to meet various business requirements [29]. Additionally, different operations of satellite and ground networks often lead to isolated systems, causing inefficiency and resource waste. These issues underscore the urgent need for a collaborative management framework for multi-dimensional heterogeneous resources between satellites and ground networks to ensure optimal resource utilization and seamless interoperability, providing users with perception, communication, computing, and caching services [30]. Several key technologies, such as MDMA [31], can support the development of collaborative management of multi-dimensional heterogeneous resources between satellites and ground networks. However, collaborative management of heterogeneous resources must meet the requirements of abstraction, flexibility, and scalability. Abstraction simplifies diverse resources into manageable entities, achieving unified resource sharing and reducing overall costs despite underlying heterogeneity. Flexibility allows the system to adapt dynamically to varying workloads and resource requirements. Scalability ensures the system can expand and integrate new resource types or technologies as they develop. In addition, to reasonably allocating the entire network resources among DSIN and NTN networks, it is essential to take into account the complexity of computing tasks, and the computing and transmission capabilities of DSIN and NTN, guaranteeing the match between heterogeneous resources and diverse tasks regardless of their source or characteristics.

2.3 Reconfigurable satellite formation flying

The functionality of the CCS system relies on the cooperation among multiple satellites. The number of available satellites within a specific space determines the system's service performance in that area. To address sudden and intense functional demands, satellites need to remain clustered within a small area. When faced with widespread and uniform functional demands, a symmetric and regular satellite configuration is often required. Depending on the spatial scale and control method, existing satellite system configurations are generally categorized into three types: satellite formations, satellite swarms, and satellite constellations.

As a key advancement in DSIN, satellite formation flying (SFF) plays a pivotal role in modern space exploration and applications to maintain a targeted orbit configuration with desired relative separation and orientation between multiple spacecraft. The concept of SFF was first introduced in the 1970s to leverage multiple satellites to conduct interferometric infrared synthetic aperture imaging tasks traditionally handled by single large satellites [32]. A specific satellite formation is composed of multiple satellites distributed on the same or adjacent orbits, functioning as an extended "virtual spacecraft" and avoiding the technical and financial challenges of building one satellite of equivalent size. For a specific SFF, the member satellites are relatively close to each other, with fixed relationships or dynamical relationships that can be represented by linearized equations, and the space-relative positions are commonly expressed using Cartesian coordinate parameters. Early satellite formations are primarily employed in synthetic aperture radar (SAR) applications, such as the US Techsat-21 program, France's Cartwheel program, and Germany's Pendulum project [33]. Due to their configuration flexibility, satellite formations are widely utilized in Earth environment monitoring missions, such as the ESA Swarm, QB-50, and GRACE projects. In space science experiments, many projects such as TPF and LiSA have been launched, and satellite formations for gravitational wave detection projects like Tianqin and Taiji are under construction in China.

The SFF evolved from simple dual-satellite configurations to more complex architectures involving dozens or even hundreds of small satellites. A satellite swarm (or cluster) is a broad concept. In general, any satellite system in which multiple satellites collaborate to perform tasks can be referred to as a satellite cluster. To distinguish from other concepts, the "satellite swarm" discussed in this subsection typically consists of multiple satellites with similar orbital parameters and close spatial proximity [32]. The concept of "satellite clusters" emphasizes extending the functionality of the system on a spatial scale, without requiring the satellites to maintain stable relative positions. The functionality of a satellite swarm can be realized by ensuring a certain number of satellites within a specified spatial range. In remote sensing applications, projects like OLFAR, SWIFT, and KickSat propose the deployment of large numbers of satellites as observation arrays to improve remote sensing performance. Remote sensing

satellite swarms, such as FISC and Flock, enable frequent global observation updates.

A "satellite constellation" is a satellite system composed of multiple satellites distributed in several identically shaped orbits according to a specific pattern. Generally, satellites in a constellation are evenly distributed along their orbits, with nearly equal spacing between orbital planes [33]. The concept of satellite constellations emphasizes a balanced spatial distribution of satellite functionality. The relative positions of satellites within a constellation are typically characterized by orbit parameters, and the use of homologous orbits ensures a stable configuration, thereby reducing the need for frequent maintenance. Satellite constellations are commonly employed for global or latitudinally distributed communication and navigation services. Since the 1970s, various countries have progressively developed mid- and high-orbit navigation constellations such as GPS, BeiDou, GLONASS, and Galileo, low-orbit communication constellations such as Iridium and Global Star, as well as inclined elliptical orbit constellations such as Molniya and O3B. Large-scale LEO Internet constellations, represented by Starlink, OneWeb, Kuiper, and Telesat, have also been extensively deployed. The aforementioned satellite systems all use fully functional satellites as their basic units. Additionally, there is a distribution concept in which satellite subsystems are separated and perform formation flying in space.

To meet the demands of different scenarios, CCS systems require scalable and reconfigurable capabilities for configuration adjustments. In near-circular orbit formation flying scenarios, the relative motion of satellites can be described by the linearized Hill equation (also known as the C-W equation) [34]. Building on this, Ref. [35] proposed the T-H equation through variable substitution, which is a linear time-varying equation describing elliptical orbits via the eccentric anomaly. The aforementioned linearized equations are valid only for formation scenarios where the inter-satellite distance is less than 20 km. However, in non-near-circular orbits and conservative perturbation environments, obtaining the analytical solutions of these equations is challenging, rendering them unsuitable for relative configuration design or the construction of edge information hubs [36], but more suitable for motion control when combined with closed-loop control algorithms such as PID, sliding mode, optimization, or robust control methods. In addition, driven by the growing need for large-scale, the DSIN is envisioned to provide continuous, real-time services from deep space exploration to planetary orbit application fields [37], such as the high-orbit high-resolution optical EO, space target attachment, electromagnetic force formation and confederacy space system design. As such, the development of satellite formation configurations has evolved from simple relative position maintenance to complex multi-satellite cooperation, and further to intelligent and adaptive control. However, these expanded formations introduced new challenges in terms of communication constraints, incomplete information, and the need for advanced coordination strategies to maintain operational efficiency.

Various approaches to satellite formation control have been explored in the literature, each offering unique advantages and addressing specific challenges associated with maintaining and controlling satellite configurations in space, as shown in Figure 6. One of the foundational approaches to formation control is the leader-follower method. In this configuration, one or more satellites serve as "leaders", setting the trajectory for the rest of the satellites, or "followers", to track [38]. The followers adjust their position and orientation relative to the leader, thus achieving a coordinated formation. The leader-following methods can be implemented in various structures, including single-leader, multi-leader, and virtual-leader formations, which offer different levels of flexibility and robustness. However, a key limitation of the leader-following approach is its dependence on the leader. If the leader experiences a malfunction or significant perturbations, the entire formation may become destabilized, posing risks to the mission [39].

In constraint to centralized formation control, the behavior-based formation control method represents a more decentralized approach [40]. Instead of relying on a designated leader, this method assigns specific behaviors to each satellite, such as collision avoidance, formation keeping, and reconfiguration. Through these individual behaviors and local control rules, the overall formation is maintained in a coordinated manner. The core challenge in behavior-based control lies in designing effective behavior-coordination mechanisms that ensure global formation objectives are achieved. For example, satellites may have to balance conflicting requirements, such as maintaining formation while avoiding potential collisions [41]. Behavior-based methods are particularly advantageous for large formations with multiple interacting units, as they allow for greater adaptability and autonomy. However, the decentralized nature of this approach can make it difficult to achieve the high levels of precision required in certain missions, as the formation's overall stability depends on the interaction of individual satellite behaviors.

Moreover, the virtual structure formation method takes a different approach by treating the formation as a single "rigid body" [42]. Each satellite maintains a fixed position within a virtual structure, such

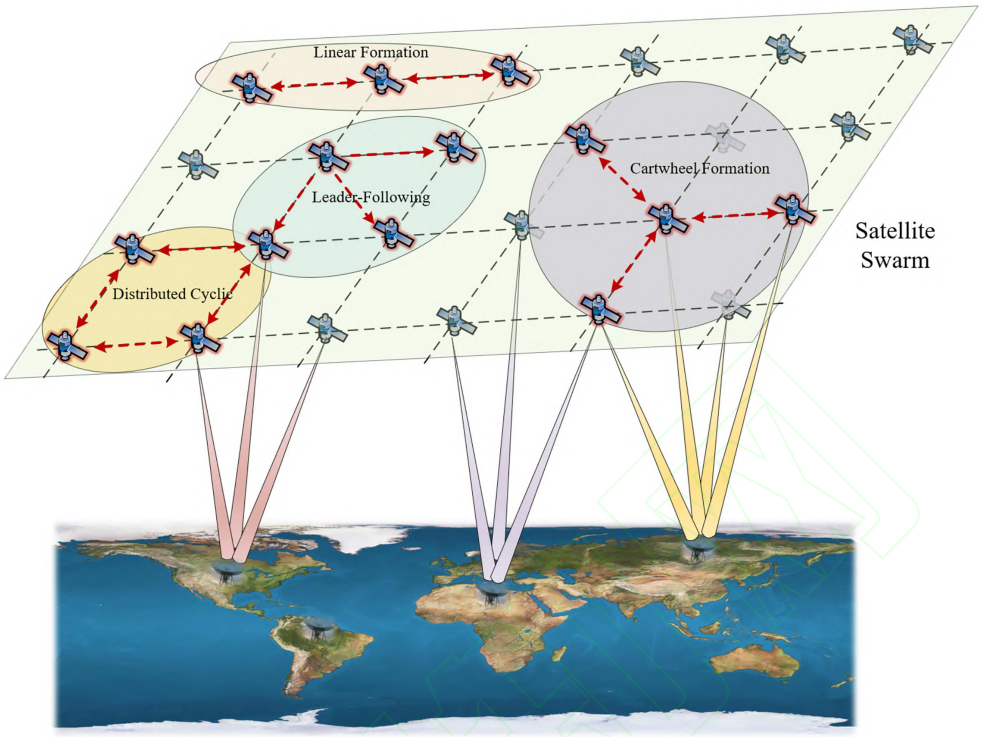

Figure 6 (Color online) Illustration of various satellite formations in satellite swarms.

as a wheel, a line, or a grid. This method is effective in achieving high precision in maintaining formation shape and alignment. For example, wheel formations involve satellites distributed around a circular or ring-like path, providing 360-degree coverage ideal for tasks like surveillance and communication relays. Other configurations include linear formations, which arrange satellites in a straight line, suitable for continuous EO and synchronized orbit tracking, and rectangular or grid formations, where satellites form a structured grid pattern for broad-area coverage. Triangular formations and cubic or spherical formations are also common in virtual structure control, offering three-dimensional coverage for complex space missions that require multi-directional observations or high spatial resolution. Additionally, honeycomb formations follow a hexagonal grid structure, maximizing spatial efficiency and are often used in communication networks and environmental monitoring due to their dense, uniform coverage. Arrow or V-formations are another variant, typically used in applications that require directional alignments, such as clustered navigation or coordinated maneuvering, as they reduce aerodynamic drag (for atmospheric applications) and optimize the communication link among satellites. Beyond these conventional configurations, recent research has also introduced optimization and predictive control techniques into satellite formation control. For instance, optimal control methods aim to minimize fuel consumption or time delay, which is essential for deep-space missions with limited resources. Model predictive control (MPC) allows satellites to adjust their trajectories based on predictions of future system states, providing adaptive and efficient control for complex formations with high maneuverability demands.

As satellite networks become more distributed, the importance of reconfigurable SFF (RSFF) cannot be overstated. Traditional formations are often designed as static structures, optimized for a specific set of mission parameters. However, in distributed satellite networks, each satellite may experience varying orbital constraints, mission objectives, and environmental conditions, requiring the formation to adapt dynamically to maintain operational effectiveness [43]. The RSFF allows satellites to perform signal synchronization, change relative positions, adjust their orientations, or even reallocate tasks among themselves, offering unparalleled flexibility for mission planning and execution [44, 45].

Several research efforts have focused on developing the necessary technologies to achieve RSFF. One of the most prominent approaches is the use of graph theory to model inter-satellite communication and control [46]. By representing satellites as nodes in a graph and their communication links as edges, researchers have developed algorithms to optimize formation control under various constraints, including limited communication bandwidth, time-varying network topologies, and external disturbances. Another

key area of research involves the development of formation control strategies based on artificial potential functions. These methods model satellites as particles moving within a potential field generated by their relative positions and target locations. By adjusting the forces acting on each satellite, the formation can be dynamically reconfigured to achieve the desired geometries while avoiding collisions or other hazards. This approach has proven effective in both simulation and experimental settings, particularly for missions involving close-proximity operations. Moreover, the control methods for swarm configurations share similarities with satellite formation control methods. However, for a CCS system already in the desired configuration, frequent control task execution leads to resource wastage. Therefore, an event-triggered mechanism can be designed to apply control only when configuration errors exceed the predefined threshold. Event-triggered control can effectively reduce both communication frequency among members and energy consumption for orbit maintenance. Currently, event-triggered control has been increasingly utilized for maintaining swarm system configurations, specifically in research areas such as first-/second-order integral multi-agent systems [47], directed/undirected topologies [48], centralized/distributed triggering mechanisms [49], fixed/switched topologies [47], linear/nonlinear systems [50], with/without time delays [51], and input saturation [52].

Further, developing decentralized control architectures that distribute decision-making across the DSIN is essential to improve resilience and flexibility. The DSIN should incorporate adaptive algorithms capable of responding to environmental changes and communication delays, enabling autonomous reconfiguration without relying on a central command unit. Moreover, hybrid centralized and decentralized control architectures use centralized control for high-precision tasks and decentralized strategies for adaptive and resilient reconfiguration, enabling the balance between precision and robustness. For inter-satellite distances exceeding 20 km, orbit maneuver requirements can be addressed based on the Lambert problem. The Lambert problem is a typical two-point boundary value problem that determines the maneuver trajectory given the spacecraft's initial and final positions and the transfer time [53]. Considering the univariate nature of the Lambert problem, recent studies have derived transfer time equations using parameters such as flight path angle [54], eccentricity [55], and terminal velocity [56], and proposed iterative solution methods. Furthermore, maneuver theory has been improved by incorporating factors such as environmental perturbations [57], thrust direction constraints [58], and multi-revolution orbit solutions [59]. The last but not least, researching algorithms that optimize energy consumption during formation reconfiguration is also crucial. These algorithms could focus on minimizing fuel use while considering the long-term sustainability of the mission. Techniques such as cooperative energy management and predictive energy budgeting can be explored to extend the operational lifespan of satellite constellations.

By adopting a phased configuration control method at different spatial scales, the CCS systems can be restructured according to service demands, choosing system configurations and control maintenance strategies that best fit the task requirements. This enables the design of highly flexible, scalable, and open network architectures for DSIN, along with the corresponding functional partitioning strategies. The future of RSFF lies in the continued integration of advanced control algorithms, high-precision sensors, and scalable communication networks. One promising avenue is the application of machine learning techniques to predict and adapt to environmental disturbances in real time. By incorporating predictive models into the control loop, satellite formations can proactively adjust their configurations to maintain optimal performance. Another important direction is adopting a phased configuration control method at different spatial scales, which allows the CCS system to be restructured according to service demands and to choose system configurations and control maintenance strategies that best fit the task requirements.

3 Enabling technologies

In order to achieve spectrum- and energy-efficient communication, as well as high timeliness and reliability, a variety of new air interface and transmission technologies will be used in the DSIN, especially in the CCS system. In this section, we will first introduce the air interface technologies at the physical and link layers, including channel modeling and estimation, new waveforms, distributed antennas, channel coding methods, multiple access approaches, and multicasting mechanisms, followed by network and transport layer technologies, involving erasure transfer protocols, distributed routing, and congestion control.

3.1 Channel modeling and estimation for CCS system

As the dawn of 6G communication technologies beckons, the crucial role of DSIN in achieving uninterrupted global connectivity has become more pronounced than ever [60]. Within this evolving landscape, the creation of precise channel models and the enhancement of channel estimation methods have become pivotal to the advancement of these systems. Channel models are indispensable for capturing the intricacies of satellite-ground communication links, offering a fundamental insight into their performance across diverse operational scenarios. Concurrently, channel estimation stands as the cornerstone technology, facilitating the real-time acquisition of channel state information, which is indispensable for fine-tuning signal transmission and guaranteeing the efficiency and dependability of DSIN [61]. Despite extensive research into terrestrial communication channel modeling and estimation, the unique challenges posed by satellite-to-ground channels in the CCS system demand a closer look. In this subsection, we delve into the channel modeling and estimation within the realm of the CCS system, and explore the latest advancements and present a comprehensive review of the current research landscape.

3.1.1 *Channel model*

In the CCS system, the channel model is essential to understand and predict the behavior of signals propagating between satellite-to-satellite and satellite-to-user links. We delve into the two primary categories of the channel model, the satellite-to-satellite and user-satellite channel model, focusing on the user-satellite channel model [62].

The satellite-to-satellite channel model, which concerns the communication links between satellites, is relatively straightforward compared to user-satellite modeling. This simplicity arises because space-to-space communication does not involve the complexities of the Earth's atmosphere or terrestrial obstacles. As shown in Figure 7, in satellite-to-satellite channel models, the primary considerations include the relative positions and movements of the satellites, the distances between them, and the characteristics of the transmission medium, which is typically the vacuum of space. The main challenges in the satellite-to-satellite channel model involve accounting for the dynamic nature of the CCS system, including the effects of gravitational forces, satellite orbit perturbations, and the need for precise alignment of communication beams. However, the relative ease of this model allows researchers to focus on optimizing the performance of ISL without the added complexities of atmospheric and terrestrial interference. The user-satellite channel model is where the real challenge lies. This type of model ensures reliable communication between satellites and ground stations [63]. The user-satellite channel model is more challenging than the satellite-satellite channel model, primarily due to many factors, including the influence of the Earth's atmosphere, interference from terrestrial obstacles, and the dynamic nature of the communication environment. User-satellite channel model methods mainly include geometric stochastic models and machine learning (ML)-based models.

The geometric stochastic model is designed with rational assumptions regarding the distribution of obstacles and scatterers near the receiving user. It abstracts the effective scatterers within the environment into one or more models characterized by distinct geometric forms, systematically distributing them across the environment's abstracted geometric structure. In [64], the studies undertook the channel model of a multi-satellite communication system, postulating that the satellites are uniformly dispersed across the surface of a sphere at a specified altitude above the Earth. Using the geometric relationship between the satellite and the Earth, the research deduced the correlations among the elevation angle, azimuthal angle, and the distance between the satellite and the Earth. Although the geometric stochastic model has high universality and low computational complexity, it has low modeling accuracy in specific scenarios. To have a more accurate channel model of the CCS system, much research is needed to consider more factors.

In CCS wireless communication environments, the channel conditions are more complex, the operating frequency bands are higher, and the mobility of the terminals is further enhanced. In such scenarios, ML's self-learning and predictive capabilities are particularly suitable, as they can extract critical features from highly complex channel data. Moreover, by thoroughly training the channel model, we can enable the model to better adapt to the rapid changes in channel states in high-dynamic scenarios, thereby significantly improving the generalization ability and performance of the model. The authors in [65] integrate ML with traditional channel modeling techniques, presenting a novel approach for channel modeling in LEO satellite systems. It underscores the critical role of radio channel forecasting in improving link

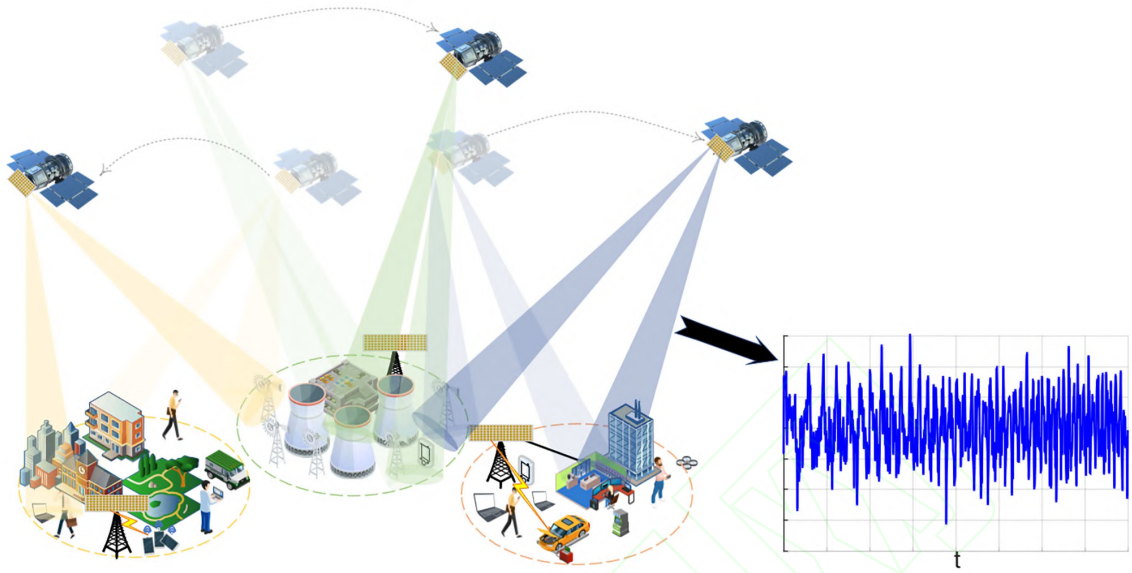

Figure 7 (Color online) Illustration of channel model in CCS system.

quality in the face of growing atmospheric impairments. In recent years, machine learning-based satellite-Earth channel modeling has seen rapid advancement, with many models being proposed, showcasing the technology's robust learning capacity and proficiency in handling vast datasets. The long short-term memory (LSTM)-based modeling has garnered significant attention, with superior prediction accuracy demonstrated in several studies. However, the predictive accuracy of current ML-based models is contingent upon the richness of the training dataset and demands substantial computational resources, which limits the applicability of many models to specific scenarios or frequency bands, as noted in [66].

3.1.2 *Channel estimation*

Channel estimation is an essential aspect of CCS systems, and is a critical issue that directly affects the performance and reliability of the communication link. This entails estimating the impact of the channel through which a transmitted signal traverses in a wireless environment. Conventionally, the channel effect is encapsulated in a channel state information (CSI) block in modern communication systems. While conventional methods like minimum mean square error (MMSE) are employed for CSI estimation, they often entail high computational costs and may not always align with the demands of real networks. Furthermore, obtaining timely CSI information gets more challenging due to extended propagation delays and fast-changing propagation environments in CCS systems [67].

The long propagation delays and high mobility inherent in SatCom channels contribute to increased channel estimation errors, while the large number of antennas adds significant overhead. In CCS systems, utilizing statistical CSI (sCSI) has proven to be a more practical approach [68]. This method effectively addresses the challenges of obtaining instantaneous CSI (iCSI) and significantly reduces the computational burden on satellite payloads by allowing for fewer updates to transmission strategies. Additionally, incorporating an effective pilot design can further enhance channel estimation, providing a robust solution to these challenges [69].

Various techniques have been developed to measure SatCom channels. In [70], Wang et al. introduced a method known as adaptive random-selected multi-beamforming estimation, which focuses on estimating geometric-based millimeter-wave multiple-input multiple-output (MIMO) CSI. They create a geometry-based channel model for multi-satellite applications within high-throughput satellite systems. A notable feature of this method is its application of compressive sensing, which estimates the combination of the transmit beamformer on the satellite side and the receiving combiner on the user side, organized randomly across several time slots. Taking advantage of the sparsity in the angle domain, this approach significantly reduces the number of measurements needed for precise channel estimation.

ML-based methods are increasingly being adopted as a promising alternative for channel estimation. This channel estimation can be framed as a supervised learning problem by using various channel features as inputs, including distance, time delay, received power, azimuth angles of arrival (AoA) and departure

(AoD), elevation angle, root mean square (RMS) delay spread, and frequency, with CSI serving as the output labels. In [71], the reciprocity property of downlink and uplink channels in time division duplexing (TDD) systems is explored, allowing the downlink channel to be estimated from uplink CSI using an LSTM-based deep learning model. Lu et al. [72] proposed a channel estimation method utilizing a fundamental deep learning architecture in multi-satellite, multi-ground environments. In [73], an ML-based CSI prediction method is applied in a massive MIMO communication scenario, using convolutional neural networks to extract temporal channel correlation features. In [74], an LSTM-based predictor is used to address the problem of channel ageing in LEO satellite communication systems. Meanwhile, several DL-based joint CSI prediction methods are incorporated in downlink precoding schemes [75, 76]. A DL-based quantized phase hybrid precoder is introduced to enhance spectral efficiency in [75]. Furthermore, taking channel prediction errors into account, Ref. [76] proposed a deep learning (DL)-based joint channel prediction and multibeam precoding scheme, which achieves significant gains in robust CSI acquisition and uplink transmission performance, even under high Doppler shifts and long propagation delays.

Although AI-enabled channel estimation methods provide a favourable trade-off between generalisation and performance-complexity, they lack interpretability in decision-making. To address this issue, Ref. [77] proposed a novel explainable AI (XAI)-based channel estimation scheme to offer detailed and reasonable interpretability of deep learning (DL) models. In summary, AI-enabled channel estimation methods have demonstrated unique advantages in enhancing the reliability and robustness of satellite-terrestrial communications. However, the significant mismatch between the learning capabilities of existing AI models and the constraints of onboard satellite systems remains a challenging issue. Therefore, guiding AI-enabled channel estimation methods toward lightweight and universally reconfigurable solutions will be a key focus for future CCS system.

3.2 Cloud-native distributed MIMO cooperation and coordinated signal processing

The deployment of large-scale satellite constellations, comprising thousands or even tens of thousands of satellites, has garnered considerable attention from industry and academia. In such systems, multi-satellite MIMO can reduce the reliance on a more significant number of antennas while improving data throughput [78]. One approach utilizes multiple non-collaborating satellites to independently apply MIMO techniques. Alternatively, cooperation between satellites, through data exchange or CSI, enables joint tasks such as transmission and resource allocation [61, 79]. This cooperation approach maximizes the benefits of the MIMO technique, significantly enhancing overall system performance. In addition, in the CCS system, collaborative reception and transmission among multiple satellite nodes can achieve uplink reception diversity gain and downlink coherent beamforming gain, respectively. In theory, joint transceiver processing can achieve optimal performance. Scalable distributed cooperative baseband signal processing has been extensively studied in cell-free massive MIMO systems [80]. The authors in [81] applied the concept of cell-free massive MIMO to satellite communications, and both theoretical and simulation results demonstrate that this approach can significantly enhance system capacity. However, the cell-free massive MIMO in [81] requires the deployment of a centralized processing unit in space, which is highly challenging to implement.

Based on Option 7 in Figure 4, the authors in [82] proposed a new low-layer physical splitting scheme to enable scalable distributed baseband signal processing, further evolving into a cell-free wireless access network. Introducing the new low-layer physical layer splitting scheme into the CCS architecture enables cloud-native collaboration of baseband computing power, thereby reducing the complexity of implementing cooperative transmission. As shown in Figure 8, the edge distributed units (Sat-EDU) implement distributed baseband transceiver processing, the Sat-DU performs high-layer physical layer processing and upper-layer processing for anchor users, while the Sat-CU handles anchor user selection, user-to-node and beam association, and other control functions.

However, the implementation of distributed MIMO cooperation and on-board coordinated baseband signal processing in CCS system faces significant technical challenges, primarily involving the realization of distributed transceivers, as well as synchronization and calibration among nodes. Specifically, when the clock or local oscillator signals are generated locally at each satellite, attaining precise synchronization and calibration in terms of absolute phase, frequency, and time becomes extremely challenging [45]. Also, distributed cooperative baseband signal processing is frequently hampered by the limitations of on-board processing capabilities and inter-satellite interaction capacity. Consequently, there is an urgent need to develop robust, efficient, and resource-conserving synchronization and calibration techniques to

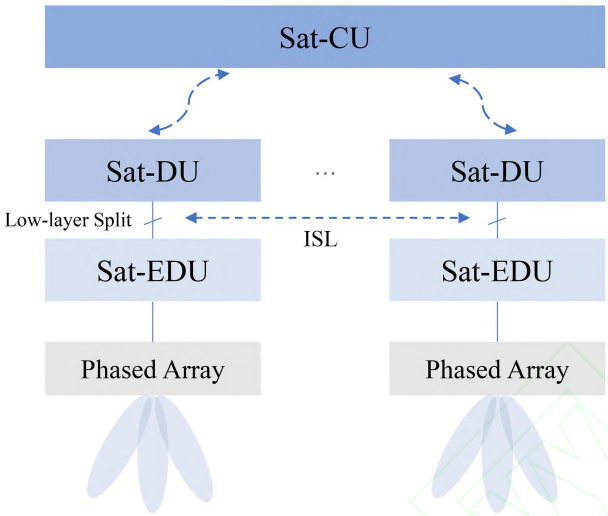

Figure 8 (Color online) Cloud-native baseband coordination.

bolster the functionality of multi-satellite MIMO systems. In addition to the above challenges, distributed MIMO proves to be of vital importance for the effective management of the vast amounts of data generated by such systems. Through the utilization of decentralized processing nodes, each satellite is enabled to independently process signals and subsequently share the results with neighboring satellites. This approach not only diminishes the overreliance on centralized data processing, but also mitigates communication overhead. Moreover, cloud-native architectures take the distributed MIMO cooperation and coordinated baseband signal processing to a new level by furnishing scalable and flexible capabilities that can dynamically adapt to the ever-evolving needs of the network [83].

3.2.1 *Uplink distributed receiver*

When applying the cell-free wireless access network of reference [82] to LEO networks, each node is equipped with a large-scale phased array, and both Sat-EDU and Sat-DU are deployed on LEO nodes. The system adopts a user-centric approach for node association, and the nodes serving users include a primary service anchor node and multiple secondary service nodes. For uplink reception, the receiver of each node independently performs multi-user detection [68]. The detected signals are sent to the primary service node Sat-DU of the user through the high-speed ISL for combined detection, and then perform higher-layer physical layer processing such as demodulation and decoding. As shown in Figure 8, uplink distributed cooperation relies on ISL. Compared to the link between Sat-DU and Sat-CU, the link between Sat-EDU and Sat-DU requires the exchange of detected modulated symbol data, placing higher demands on ISL capacity.

3.2.2 *Synchronization, calibration and downlink distributed beamforming*

Signal propagation delays, determined by satellite altitude and other parameters, typically range from tens to hundreds of milliseconds. This significant transmission delay imposes stricter timing management requirements for uplink and downlink signals. Moreover, signals from multiple satellites often cover the same area, enabling a single terminal to be served by multiple satellites simultaneously. However, the propagation delay differences and Doppler frequency offsets between various satellites and user terminals are significantly larger than those in terrestrial communication systems. This makes it challenging to achieve synchronization between satellites and terminals akin to synchronization and calibration in terrestrial systems. Managing synchronization and calibration between satellites and terminals to avoid interference between downlink transmissions from satellites and uplink transmissions from terminals is a critical challenge in multi-satellite cooperative transmission.

The terrestrial synchronization and calibration methods are mainly divided into hardware calibration and over-the-air (OTA) calibration. Hardware calibration requires additional reference antennas, which have been extensively studied in TDD massive MIMO [84]. OTA calibration requires no additional hardware and is realized through the transmission of reference signals between the remote radio unit

(RRU) or between RRU and UE, including self-calibration and UE auxiliary calibration. For example, the authors in [85] proposed a trunk-based calibration method, but the calibration time is long. The authors in [86] respectively adopted different synchronization methods in WiFi distributed MIMO experiments. The authors in [87] implemented the system based on the OpenAirInterface (OAI) platform and proposed a fast calibration method.

For downlink transmission, the primary serving node distributes information to the secondary nodes, and each node independently performs beamforming [68]. As mentioned above, multi-node coherent cooperative transmission relies on phase synchronization between nodes. Time-frequency synchronization between multiple satellite communication nodes is the core issue for downlink coherent transmission. Phase synchronization between analog beams of multiple distributed phased arrays is affected by the following factors: (1) time-frequency synchronization between multiple nodes; (2) consistency calibration between multiple analog beams of multiple nodes. Unlike terrestrial systems, satellite nodes typically direct their beams toward the ground, and phased arrays often lack an OTA link, making it challenging to achieve inter-satellite self-synchronization. A feasible method involves ground terminal-assisted phase synchronization and tracking. For TDD systems, terminal-assisted calibration and phase tracking can support coherent cooperative transmission among multiple phased arrays. For frequency-division duplex (FDD) systems, phase synchronization among multiple phased arrays can be achieved through terminal measurement, feedback, and phase tracking.

Terminal-assisted phase synchronization involves periodic measurement and feedback. However, the significant delays and Doppler frequency offsets between the terminal and multiple nodes result in substantial time overhead for measurement and feedback. Therefore, achieving high-precision phase synchronization imposes higher requirements on the local oscillator precision of the phased arrays and the accuracy of Doppler frequency offset estimation. For downlink coherent cooperative transmission, phase synchronization errors among the analog beams of multiple phased array nodes are unavoidable. Thus, designing robust distributed cooperative digital precoding is essential [88, 89].

From the above analysis, it can be observed that satellite systems differ significantly from terrestrial systems, rendering the synchronization and calibration technologies of terrestrial systems inapplicable to CCS systems. Satellite synchronization continues to face numerous challenges, such as achieving precise synchronization and calibration in terms of absolute phase, frequency, and time, especially when the clock or local oscillator signals are locally generated in each satellite. Moreover, the constrained resources of satellite systems further complicate these issues. It is necessary to develop more robust, efficient and resource-saving synchronization and calibration techniques to support MIMO cooperation and coordinated baseband signal processing in the CCS system.

3.2.3 *Cloud-native distributed MIMO*

In DSIN, LEO satellites move at high speed along a predetermined orbit, which continuously changes their position relative to the UTs. As a result, UTs must frequently switch links between different LEO satellites to maintain a stable network connection. This complex switching process spans two critical layers: the link and network layers. At the link layer, the primary goal is to seamlessly transfer the communication link from one satellite to another within the UTs line of sight, akin to constructing an invisible yet indispensable bridge. The network layer, however, poses a more significant challenge. During a satellite handover, when the UT connects to a new "home" satellite network, higher-level protocols such as transmission control protocol (TCP) and user datagram protocol (UDP) must migrate swiftly and seamlessly to the UTs' new Internet protocol (IP) address. This ensures that ongoing data transmissions remain uninterrupted. Unfortunately, the visible window of any single LEO satellite to a UT lasts only a few minutes, necessitating frequent handovers. This results in high signaling overhead and many issues, including reduced throughput, processing delays, data forwarding congestion, and lagging location updates. These challenges severely degrade network performance, compromise spectrum utilization, and negatively impact the user quality of service (QoS).

Distributed MIMO technology, which has demonstrated remarkable success in terrestrial communications, offers promising solutions for LEO satellite networks. In terrestrial-based scenarios, distributed MIMO leverages multiple access points in a collaborative, cellular-free manner to achieve high spectral efficiency, superior power efficiency, and excellent network flexibility [89]. Now, this cutting-edge technology is poised to revolutionize LEO satellite networks. By capitalizing on ultra-dense satellite constellations, ultra-fast ISLs, and line-of-sight (LoS) connectivity with terrestrial UTs, distributed MIMO is expected

to unlock unprecedented communication performance. In parallel, integrating cloud-native with LEO satellite networks has injected new energy into the development of distributed MIMO systems [81].

Cloud-native refers to a set of technologies that decompose applications into microservices and package them into lightweight containers for deployment and orchestration across various servers [22]. The cloud-native satellite cluster can be seen as a distributed, elastic, and horizontally scalable system, serving as a "cloud brain" for satellite networks [83]. The cloud-native satellite cluster consists of interrelated on-board microservices that isolate state in a minimal number of stateful components. As a cloud-native CCS system, it adopts practices such as microservices, containerisation, and orchestration to enable agility, scalability, and rapid development and deployment of satellite-integrated Internet applications. To set up a cloud-native distributed satellite cluster, the monolithic system must first be decomposed into self-deployable, function-specific microservices [22], which can communicate with one another through lightweight messaging protocols over ISL. Subsequently, each microservice is packaged into a container using virtualisation technologies such as Docker, KubeEdge, and Kubernetes. Each container is then orchestrated into an integrated system for functionality, with automatic configuration. Through the pooling and unified orchestration of computing power at satellite edge nodes, the cloud-native CCS system can dynamically allocate resources to each node for distributed signal processing and collaborative computation [90]. In this regard, cloud-native distributed MIMO efficiently leverages fragmented satellite node resources for cluster node status monitoring and information awareness. Based on the information gathered, it facilitates distributed collaborative transmission [91]. A cloud-based cell-free distributed massive MIMO system is investigated by [92], addressing challenges related to synchronization, calibration, and real-time baseband processing in 5G NR. General-purpose multi-core CPUs are employed for baseband signal processing.

In addition, when LEO satellite clusters perform collaborative transmission toward the ground, frequent connection handovers occur, which slow down the cluster service restart process and reduce satellite node resource recovery efficiency. On one hand, the cloud-native CCS system ensures consistency of satellite cluster information across nodes through a distributed communication protocol, while periodically updating the health status of cluster nodes to ensure rapid dissemination and proactive exclusion of faulty node information. On the other hand, by relying on the service migration and rapid resource release and recovery capabilities of the cloud-native CCS system, the cloud-native distributed MIMO can quickly rebuild container services for new satellite nodes and initialize satellite cluster information, thereby enabling flexible and robust signal processing and satellite-terrestrial transmission.

3.3 Waveform and modulation technology in CCS system

The waveform is the shape of signal as it propagates through the physical medium, typically represented by its distribution over time or space. It can also be abstracted to other domains such as frequency, code, or index. Since the waveform directly affects the effective signal-to-noise ratio (SNR) and spectral efficiency (SE) of CCS systems, flexible and efficient waveform design, along with its corresponding modulation techniques, is a key focus in DSIN.

3.3.1 *Current waveform and modulation technology*

(1) Orthogonal frequency-division multiplexing. Orthogonal frequency-division multiplexing (OFDM) is one of the most widely used modulation technology, which has been adopted by many wireless standards such as IEEE 802.16, 4G long-term evolution (LTE), 5G new radio (NR) and Wi-Fi [93]. In particular, OFDM transforms a wideband fading channel into a set of parallel narrowband flat-fading sub-channels. By appropriately selecting the number of subcarriers, it ensures that the bandwidth of each sub-channel is smaller than the coherent bandwidth of the transmission channel, thus preventing frequency-selective fading and avoiding inter-symbol interference (ISI). However, conventional OFDM with rectangular shaping suffers from high sidelobes in frequency-domain and stringent criteria of orthogonality between subcarriers [94], which limits the flexibility of its waveform configuration and makes it difficult to adapt to the diverse services emerging in 5G.

In response, although OFDM is still used in 5G, several variants are proposed during the standardization process to achieve overlapping orthogonal design. These schemes are mainly categorized into subband filtering and subcarrier filtering in multicarrier systems. On the one hand, subband filtering includes universal filtered multi-carrier (UFMC) [95] and filtered OFDM [96]. The former yields a good frequency-domain localization but may suffer from severe ISI due to the absence of cyclic prefix (CP). In

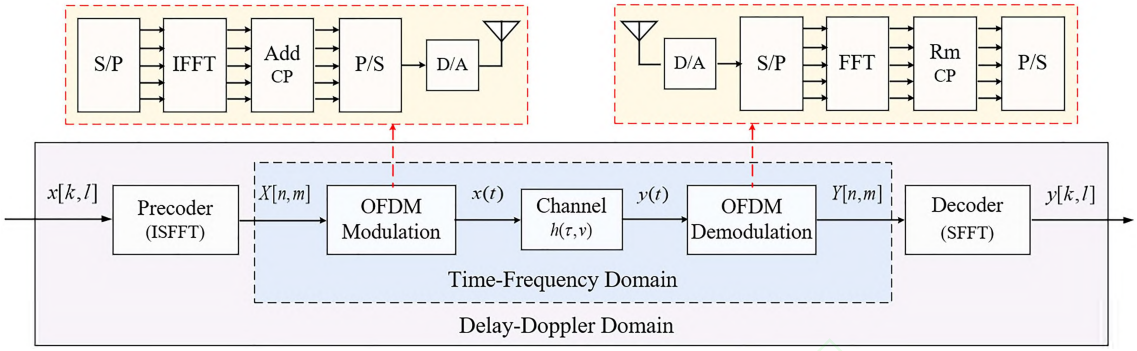

Figure 9 (Color online) Illustration of OTFS architecture compatible with OFDM system.

contrast, the latter offers subband filtering in a flexible manner to meet user requirements and uses the CP to mitigate ISI. On the other hand, subcarrier filtering involves filter bank multi-carrier (FBMC) [97] and generalized frequency division multiplexing (GFDM) [98]. The former provides a certain degree of flexibility through the selection of arbitrary pulse shaping filters and their parameters, whereas the absence of CP leads to similar issues as UFMC. The latter, based on independent block modulation, offers a flexible frame structure but has higher requirements for time synchronization.

In addition to filtering, windowing also plays a crucial role in waveform shaping. For example, a windowed OFDM scheme proposed in [99] can smooth the transitions between adjacent symbols by adding additional prefixes and suffixes, thereby reducing the out-of-band radiation caused by the rectangular pulse shaping. Overall, these techniques effectively prevent potential inter-carrier interference (ICI), ISI, and adjacent channel interference (ACI), at the cost of increased transceiver complexity, delay, and reduced SE.

(2) Orthogonal time frequency space. Current OFDM and its variants can only support high data rates and service quality in low- and moderate-speed mobility environments, while their performance is limited in LEO satellites of DSIN. The reason is that the OFDM-based modulation process occurs in the time-frequency (TF) domain, making it highly susceptible to the Doppler effect caused by the high mobility of LEO satellites, which can easily disrupt the orthogonality of the OFDM waveform and lead to severe ISI and ICI [100]. To address this issue, a direct solution is to increase the CP ratio in OFDM to preserve orthogonality, although it results in a significant decrease in SE. Another approach is to predict the Doppler shift based on the ephemeris and perform pre-compensation, but this method requires tracking each OFDM subcarrier individually, which brings significant complexity and hardware costs, further burdening the limited payload of LEO satellites [101].

Recently, a novel waveform modulation technique for high-speed mobility scenarios, called orthogonal time frequency space (OTFS), has been proposed [102,103]. OTFS is a two-dimensional modulation technique that represents transmitted signals in the delay-Doppler (DD) domain. By utilizing the symplectic finite Fourier transform (SFFT) and its inverse transform ISSFT, it converts the highly time-varying TF channel model into a sparse, slow-varying channel model in the DD domain, thus averaging out the rapid channel dynamics caused by satellite movement. This method takes full advantage of the complete diversity offered by time and frequency selective channels, resulting in much higher SE compared to OFDM, with smaller subcarrier spacing and reduced CP overhead. It also achieves a lower peak-to-average power ratio (PAPR), as the ISFFT spreading operation enhances the maximum energy per bit. As a result, OTFS is considered a promising waveform and modulation technology for 6G and has been included in discussions at the 3GPP meetings [104, 105].

The OTFS modulation also presents several significant advantages. First, OTFS can be implemented by embedding an ISFFT precoding and corresponding SFFT decoding module within the existing OFDM framework, allowing for seamless integration with current OFDM systems [106] and ensuring architectural compatibility with technologies like LTE, as shown in Figure 9. Second, the reduced time variability of channels in the DD domain enhances the practicality and robustness of OTFS, while also lowering the overhead and complexity associated with physical-layer adaptation. Moreover, the inherent sparsity of satellite-to-terrestrial channels is also beneficial for the design of low-complexity OTFS receivers [107]. Last but not least, the compact representation of DD channels in OTFS enables efficient packing of reference signals, which provides robust support for large-scale MIMO applications and facilitates the

integration with multiple access techniques, such as non-orthogonal multiple access (NOMA) [108]. In summary, OTFS is a promising technology for enabling coordinated waveform design within the CCS system in DSIN.

3.3.2 *Challenges and development of OTFS*

There are still some open issues and challenges in OTFS modulation that require further investigations, in addition to the advantages mentioned above.

(1) Interference management. If the time and frequency resolution of the signal is not inversely related to the channel's Doppler shift and delay in OTFS, it can result in a loss of channel sparsity and lead to ICI in the DD domain, also known as inter-Doppler interference (IDI). In response, researchers have proposed several solutions to enhance the sparsity of channels and mitigate the impact of IDI, such as performing cross-domain iterative detection and estimation jointly [109], executing windowing based on water-filling power allocation in the TF domain [110], and applying block-based joint detection and IDI elimination [111]. However, these schemes all come with the cost of increased receiver complexity. Unlike conventional OTFS receiver designs, recent studies have indicated that the successive interference cancellation (SIC)-based MMSE detector can achieve superior interference cancellation with lower complexity, even in the presence of imperfect CSI [112, 113].

Similarly, the RF impairments at the OTFS transmitter can also cause interference, with the presence of in-phase and quadrature imbalance (IQI) leading to mirror Doppler interference (MDI) in the DD domain [114, 115]. Different from the classical OFDM schemes, IQI does not cause saturation in the bit error rate (BER) performance of OTFS; instead, it only diminishes the channel diversity gain. Simulations in [116] evaluate the performance of OTFS under various RF impairments and demonstrate that the increase in pilot power and the number of Doppler bins, as well as the application of windowing, can enhance its interference resilience. Further, due to the difficulty of implementing complex nonlinear receivers on computing- and energy-constrained LEO satellites, machine learning techniques, leveraging the predictability of Doppler shift (i.e., satellite movement), can be used with well-trained agents to reduce the complexity of interference detection and cancellation.

(2) Channel estimation. Recent researchers have proposed several channel estimation algorithms for OTFS systems to fully leverage the sparsity of the DD domain, including methods based on impulse surrounded guard symbols [117], embedded pilots [118], modified compressive sensing (CS) matrix [119], and sparse Bayesian learning (SBL) frameworks [120]. In contrast to the above algorithms that only make use of pilot symbols, the schemes that incorporate data symbols can further enhance estimation performance. For example, the OTFS channel estimation method with superimposed pilots proposed in [121] utilizes the sum-product algorithm (SPA) for data detection, and then performs data-assisted channel estimation that achieves higher SE. Expanding on the use of data symbols as "virtual pilots", the authors in [122] further exploited the sparsity of the DD domain within an SBL framework by applying variational Bayesian inference (VBI) to estimate the DD channel vector, thereby achieving reduced complexity while maintaining channel estimation accuracy.

It is noteworthy that, although several studies have proved that collaborative downlink transmission in DSIN can substantially enhance the SE [123, 124], the parallel transmission of multiple data streams may cause a loss of channel sparsity in the DD domain, which requires more pilot resources to maintain channel estimation accuracy and significantly increases overhead. Therefore, the key lies in designing a novel channel estimation method that can strike a tradeoff between cost and accuracy for DSIN. The scheme proposed in [125] introduces a simultaneous pilot-based aggregate channel estimation algorithm to improve channel estimation in CCS systems, and designs a three-stage peak-searching correlation-based method to handle fractional Doppler estimation.

(3) MIMO-based OTFS. MIMO technology, with its extensive spatial degrees of freedom, low cost, and high integration, has accelerated research on OTFS beamforming and transmission efficiency, which enables further enhancements in the SE and diversity performance of current OTFS systems [126]. One of the main challenges for the MIMO system is the high processing complexity at the receiver. For instance, the message passing algorithm (MPA) proposed for MIMO-OTFS systems in [127] provides excellent performance but also incurs high computational complexity due to its nonlinear nature, especially in channels with high Doppler shift and high-order modulation. Moreover, spatial correlation can degrade the BER performance of MIMO-OTFS systems, particularly when receiving algorithms are designed without accounting for antenna correlation and inter-antenna interference (IAI) [128].

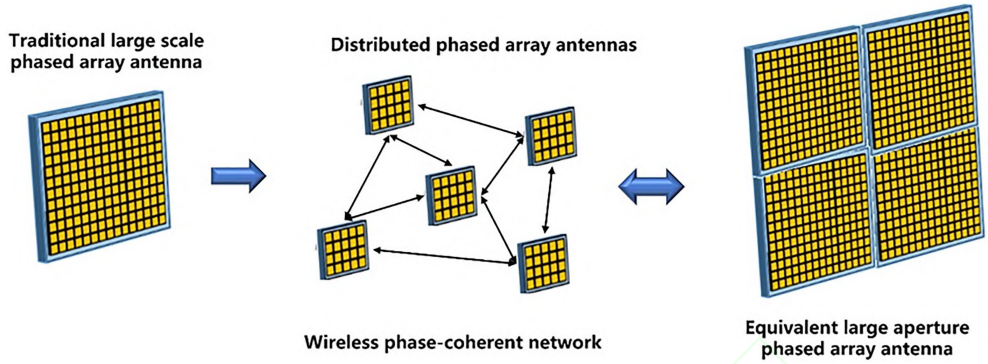

Figure 10 (Color online) Traditional large scale phased array antennas vs. distributed phased array.

To address these issues, recent studies have proposed several linear detection algorithms for MIMO-OTFS systems, based on maximum ratio combining (MRC) [129], MMSE [130], least squares minimum residual (LSMR) [131] and others, to fully exploit the two-dimensional sparsity in the DD domain and reduce computational complexity. Further, these schemes all utilize whitening transformation (WT), a technique commonly used in MIMO systems to modify its channel matrix, such as Cholesky decomposition, to mitigate the spatial correlation between transceivers. Considering that the methods mentioned above do not eliminate the spatial correlation at the receiver compared to the transmitter, the authors in [132] designed a whitening filter to further improve the BER performance of MIMO-OTFS systems with receiver correlation.

Furthermore, researchers have recently explored the integration of MIMO-OTFS with grant-free random access (GFRA) technology for more complex multi-user detection scenarios. The authors in [133] introduced a two-dimensional pattern coupling hierarchical prior in SBL, combined with a generalized approximate MP algorithm, to improve the utilization of 2D burst block sparsity in the GFRA channel matrix, which arises from the three-dimensional structural sparsity of MIMO-OTFS in the DD domain. In [134], the investigation is extended to multi-satellite MIMO-OTFS systems, where a joint scheme for device identification, channel estimation, and symbol detection in collaborative multi-satellite GFRA is proposed, and a distributed approach is adopted to offload computational tasks to edge satellites, thereby reducing the burden on individual satellites. In addition, it is important to note that OTFS modulation ensures inherent stability of the wireless channel in the DD domain [135, 136], which greatly facilitates the design of robust physical layer security schemes [137].

3.4 Phased array antennas for DSIN

The transition from large, single-satellite communication platforms to networks of relatively small, coordinated satellite platforms marks an important advancement in SatCom [138]. The distributed phased array antennas on these platforms can form a large equivalent aperture and enhanced flexibility in platform spacing, formation configurations, and network topology, enabling adaptable system configurations, which offer several advantages, including improved data throughput, and great resilience to interference [139]. By utilizing the collective capabilities of multiple small phased array antennas, the CCS system can dynamically adapt to mission changes and provide flexible coverage over vast areas, while minimizing the impact of individual platform failures [45, 140, 141]. Compared to traditional phased array antennas, CCS systems allow subarrays to be deployed across different coordinated satellite nodes, as shown in Figure 10. This approach mitigates the challenges of mounting large-scale phased array antennas on a single satellite, simplifying design and dramatically reducing the cost of satellite production and launch. To sum up, the distributed phased array antennas can enhance overall performance through phase-coherent collaboration in the CCS system. However, deploying distributed phased arrays introduces new design constraints and technical challenges, such as the limited payload capacity of small satellite platforms, precise synchronization between subarrays, multi-beam beamforming and fast scanning. Addressing these challenges is crucial for unlocking the full potential of distributed phased array systems in large-scale DSIN.

3.4.1 *High-performance silicon-based integrated phased array antennas*

The development of distributed phased array antennas involves two key considerations. First, the limited payload capacity of distributed satellite platforms necessitates highly integrated phased array antennas that deliver high output power, maintain low noise levels, and operate with low power consumption. Second, the demand for cost-effective designs to support the deployment of amounts of distributed phased array antennas across numerous satellites. The emergence of large-scale phased array antennas started in the late 1970s with the development of phased-array radars for missile early warning systems. Since then, phased array antenna technology research has been predominantly concentrated on military applications. Although the advantages of phased arrays in communications and radar are well-established, their use in industrial and commercial products has been limited due to the high cost. Over the years, notable efforts have been dedicated to reducing the cost, weight, and size of phased array systems. Recent advances in silicon-based CMOS process and PCB technologies, coupled with the increasing demand for millimeter-wave satellite communication applications, have spurred growing interest in developing low-cost, high-efficiency integrated phased array antennas [142–144]. Si-based CMOS process offers key advantages, including high yield and reliability, which, compared to compound semiconductors, significantly reduce the cost of active channels in phased arrays, making large-scale distributed systems feasible [145]. Furthermore, heterogeneous hybrid PCB technology allows the integration of large-scale microstrip antennas, RF circuits, control circuits, power supply circuits, and phased array chips onto a single PCB at millimeter-wave frequencies, as shown in Figure 11. This replaces traditional brick-and-tile phased array architectures, leading to enhanced antenna integration while reducing size, weight, and profile, thus enabling satellite-based phased array applications. The PCB-based approach also streamlines production by eliminating complex micro-assembly steps, increasing throughput, and reducing manufacturing costs [146,147]. However, Si-based CMOS technology still lags behind compound semiconductors in terms of key parameters, such as breakdown voltage, power density, electron mobility, and thermal conductivity. As a result, Si-based CMOS devices show inferior performance in RF applications, particularly in the millimeter-wave spectrum, in terms of output power and noise figure. To overcome these limitations and improve the performance of individual channels in the phased array system, hybrid CMOS-GaAs packaging is a promising solution [148].

3.4.2 *Distributed phased array synchronization*

The objective of synchronization in distributed phased arrays is to ensure that signals from multiple nodes reach the target location with precise phase alignment and accurate timing, thereby achieving constructive interference to maximize signal power [149]. For received signals, precise alignment of phase and timing across the distributed array is necessary to enable coherent processing. The techniques used to estimate and adjust these phase and timing discrepancies vary depending on whether the system is in receive-only mode or engaged in transmission. On the one hand, for receive-only configurations, data-driven methods can be used to estimate and correct phase alignment during post-processing, though this may introduce some latency. On the other hand, coherent distributed transmission is more complex, as it requires real-time alignment of the electrical states of each antenna element in the array. This process includes ensuring frequency synchronization across all nodes, calibrating internal phase delays, and correcting phase and timing differences caused by varying node positions [150].

3.4.3 *Multi-beam beamforming mechanism*

In CCS systems, the multi-beam capability of phased array antennas is crucial, as it enhances system spectral efficiency through beam diversity [151,152]. As shown in Figure 12, there are two implementation methods for multi-beam phased array systems: digital phased arrays and analog phased arrays [153, 154]. In a digital phased array, the signal received by each antenna element is amplified, filtered, down-converted, digitized at an intermediate frequency (IF), and digitally down-converted to create a zero-IF digital signal. Each antenna element's zero-IF digital signal is then sent to an array signal processing subsystem, where the desired multi-beams are formed. However, the digital phased array approach requires down-conversion and digitization of signals for each antenna element. Given that hundreds of channels may be needed to achieve the necessary antenna gain, this system requires a large amount of equipment, resulting in high power consumption and significant costs [155–157]. In an analog phased array, the signal received by each antenna element is amplified and split into multi-path, and each path

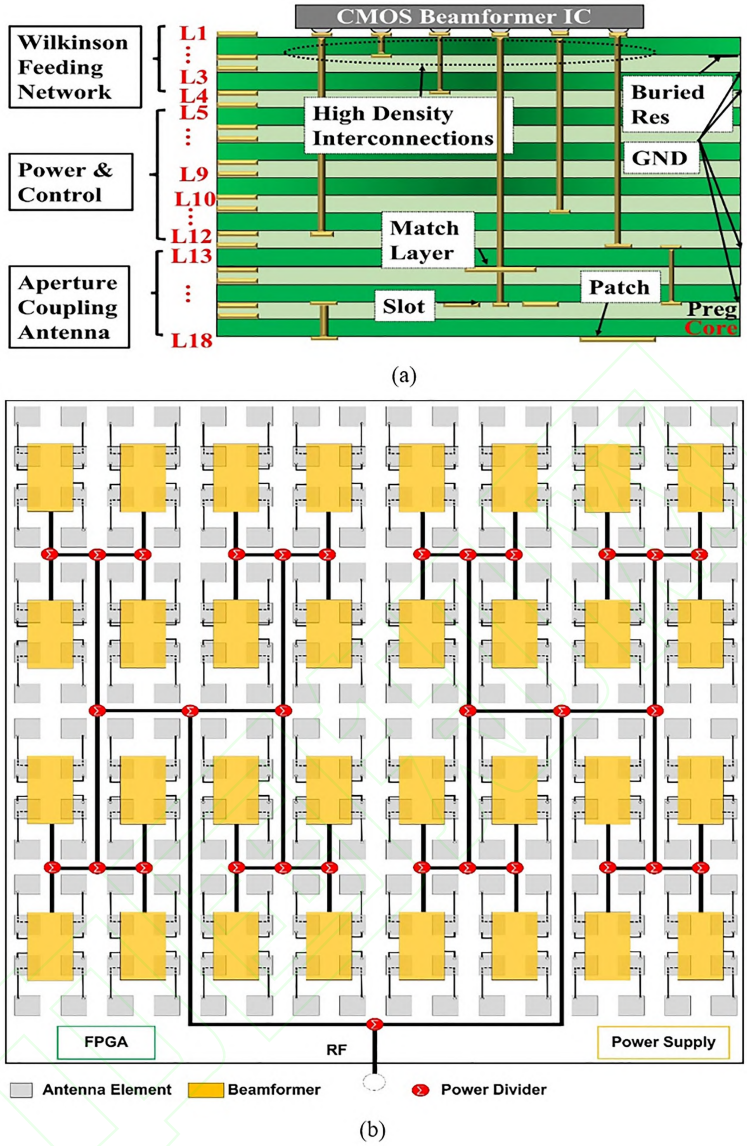

Figure 11 (Color online) (a) Typical multilayer hybrid PCB stack-ups; (b) block diagram of typical PCB phased-array.

is independently weighted in amplitude and phase in the analog domain, with delay compensation and synthesis to ultimately form the multi-beam. All beams are then down-converted and sent to the baseband processing subsystem [158]. The greatest advantage of the analog phased array lies in its greatly lower cost compared to digital phased arrays, as well as its high level of technical maturity. When the number of beams is four or fewer, multiple beams can be efficiently realized using standard analog phased array components, offering excellent cost-effectiveness [159]. To overcome the hardware limitations of fully digital beamforming, hybrid beamforming can also be utilized. In hybrid beamforming systems, a large number of antenna elements are connected to a limited number of RF chains through a network of phase shifters. In certain scenarios, hybrid beamforming can achieve spectral efficiency comparable to that of fully digital systems [160, 161].

3.5 Channel coding in DSIN

Channel coding is essential for ensuring the transmission reliability of our DSIN [162]. With the advancements in 5G and the potential applications of 6G, future communication scenarios will become increasingly diverse, presenting both challenges and opportunities for the design of channel encoding and decoding. For instance, the large- and moderate-length coding schemes with high throughput and reliability are more suitable for the enhanced mobile broadband (eMBB) in 5G, which tolerates higher delays

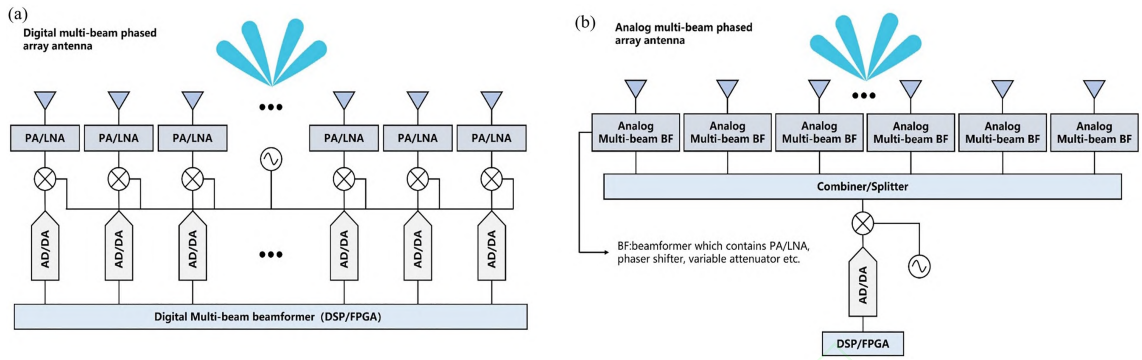

Figure 12 (Color online) Block diagrams of (a) digital multi-beam phased array antenna and (b) analog multi-beam phased array antenna.

Table 1 Channel coding used in air interface of DSIN.

State-of-the-art channel coding techniques			
Code	Decoder	Advantages	Disadvantages
TBCC codes	Viterbi decoding	Fast encoding and favorable rate performance with very low BLER at short block lengths	Low BLER requires high-order memory, but the decoding complexity increases exponentially with the memory order
Polar codes	SC decoding, SCL decoding, CA-SCL decoding	Fast encoding and decoding with simple algorithms, can achieve very low BLER at short block lengths with SCL decoding, no sign of error floor	The decoding with sequential structure introduce delay, a large list is required to achieve low BLER, the list decoding is not practical for low-power terminals
Turbo codes	Iterative soft decoding, MAP decoding	Fast encoding and decoding with parallel structure, can achieve near-Shannon limit performance at large block lengths	The iterative decoding introduce delay, exists an error floor at the high SNR region, poor performance at moderate and short block lengths
LDPC codes	BP decoding, MS decoding, IMS decoding	Fast and efficient decoding with parallel structure, excellent BLER performance at large block lengths	Exists error floor at short block lengths and low code rate, the BP decoder is sub-optimal at short block lengths, the modulation granularity is not flexible
Universal decoder design			
Encoder	Decoder	Advantages	Disadvantages
Any linear encoding scheme	GRAND algorithm	Can achieve very low BLER at short block lengths and high code rate with acceptable complexity	The decoding complexity is high at low code rate and moderate block lengths, the BLER performance is sensitive to noise patterns
	OSD algorithm	Can achieve near ML performance at short block lengths without an error floor, exhibits strong practical feasibility	The decoding complexity is increases exponentially with the order especially at moderate and large block lengths, the low-complexity improved methods introduce many adjustable parameters

but demands greater channel capacity. On the other hand, ultra-reliable and low latency communications (URLLC) require extremely low latency and higher reliability, making short- and moderate-length coding schemes that excel in low complexity, low delay, and ultra-high reliability more appropriate [163]. Table 1 illustrates a comparison between different channel coding and decoding techniques for our DSIN.

3.5.1 *State-of-the-art channel coding*

The challenges and potential coding technologies for various code lengths under the transmission requirements of our DSIN for future 6G networks are summarized as follows.

(1) Short-length coding schemes. In the short codes with lengths below 1024 bits, the competitive coding schemes include: the classic algebraic code such as Bose-Chaudhuri-Hocquenghem (BCH) and Reed-Solomon (RS) codes [164], the convolutional code [165] that can achieve efficient encoding and decoding through trellis structures, and the Polar code [166], which have been standardized for the control channel in 5G NR. Currently, various types of classic algebraic codes have been fully optimized, while convolutional codes, represented by tail-biting convolutional code (TBCC), have been replaced by the higher-performance Polar codes in the 5G standard. Thus, we focus on the development of Polar code, which can achieve near-capacity transmission by applying successive cancellation (SC) decoding to polarized sub-channels created through merging and splitting operations on i.i.d. binary memoryless channels, where some sub-channels approach noiselessness and others become fully noisy under complete polarization.

Motivated by the construction of the Polar code, it is evident that its code length is restricted to powers of two, making it challenging to meet the rate-adaptive requirements in various applications. To address

this limitation, a widely adopted solution is to use the original Polar code as a base code, and then adjust its length by removing certain codewords. Techniques like puncturing [167] and shortening [168] are typically employed, with puncturing being more suited for low-rate scenarios and shortening better matching high-rate requirements. As a result, the 3GPP adopts a rate-adaptive scheme for Polar code that integrates puncturing, shortening, and repetition in the current 5G standard [169].

(2) Large- and moderate-length coding schemes. In the large and moderate codes with lengths of over 1024 bits, notable coding schemes include the Turbo code [170] and the low-density parity-check (LDPC) code [171]. Thanks to the near-Shannon limit decoding performance at large block lengths and high-throughput parallel decoding capabilities, the LDPC code has already been standardized for 5G data channels. The Turbo code, however, has been replaced by LDPC codes in 5G due to the inherent error floor and relatively high decoding delay in iterative processes. Moreover, the concatenated BCH and LDPC coding scheme is used in the digital video broadcasting-satellite-second generation (DVB-S2) standard for satellite broadcasting, achieving near-Shannon capacity and excellent spectral efficiency in complex satellite-terrestrial channels. According to the Consultative Committee for Space Data Systems (CCSDS), the outer code in DVB-S2 employs the quasi-cyclic LDPC (QC-LDPC) code, which offers reduced computational complexity and effective parallel processing performance [172, 173].

To support the demand for higher data throughput in future networks, conventional serial coding faces high complexity at large block lengths, while parallel coding imposes significant hardware and power constraints. Consequently, coupled LDPC code [174] has become a key candidate scheme for handling ultra-long data stream transmission, constructed by introducing coupling constraints across multiple independent LDPC code blocks. Representative schemes involve spatially-coupled LDPC (SC-LDPC) code [175] and staircase LDPC code [176], each offering unique advantages for high-throughput streaming. SC-LDPC code can be constructed by coupling multiple independent LDPC code blocks into a chain using the matrix expansion-based method [177] or protograph-based method [178]. Staircase LDPC code is a type of product code with a generalized coupling structure. It introduces additional encoding constraints between adjacent code blocks to ensure that each row and column represents a codeword, which leads to a lower error floor. However, conventional staircase LDPC code with a fully coupled structure faces the issue of fixed re-encoded length and code rate once the component codes are determined. Thus, several studies proposed a partially coupled LDPC coding scheme to address this problem, allowing rate-adaptive code designs with minimal loss in coding gain [179, 180].

(3) Distributed coding within multi-satellite cooperation. The above coding schemes may encounter conflict resolution issues in multi-satellite cooperative transmission within the CCS system. In such cases, code-domain multiple access is a promising technology in the CCS system, which mainly includes T-fold multiple access [181] and lattice-code multiple access (LCMA) [182] schemes.

The T-fold multiple access scheme using concatenated codes can reliably recover a certain number of conflicting information with low complexity. Specifically, when the number of conflicting sources does not exceed T, the outer code can correctly decode the messages of each source from the superimposed codewords without errors. However, the decoding complexity of both the T-fold codebook and the outer code increases significantly with code length, especially when the outer code is Polar code rather than LDPC code, rendering it appropriate only for short block length transmissions [183].

In contrast, the lattice code maps messages onto lattice points with power constraints and achieves rate limits approaching $\log(1 + \mathrm{SNR})$ through the design of nested lattice structures [184]. Therefore, the LCMA scheme no longer distinguishes signals from interference. Instead, it approaches the limits of multiple access capacity and multiplexing rate by exploring the optimal mapping between multi-user aggregated signal structures and lattice structures, combined with the advantages of fast parallel processing and low-complexity single-user decoding, which also enables it to handle a wide range of block lengths. Nevertheless, the spreading sequences of lattice code cannot adapt to time-varying channels in real-time, and the modulation schemes must match the encoding alphabet of lattice code [185], which both limit its performance in diverse variable-length data services under complex channel constraints within our DSIN.

3.5.2 *Decoder design*

The decoder can be classified into two types: the dedicated decoder designed for a specific coding scheme, and the universal decoder that can adapt to any linear coding scheme.

(1) Dedicated decoder. Polar codes can be efficiently decoded using the SC algorithm, but their performance may degrade due to incomplete channel polarization at short and moderate block lengths. In response, the SC list (SCL) algorithm [186] is proposed to enhance decoding performance, which can achieve a block error rate (BLER) lower than 10^4 with a list size greater than 32 for 5G Polar code. Further enhancement in the SCL decoding performance can be obtained by integrating cyclic redundancy check (CRC) bits, but at the cost of increased decoding complexity [187]. Moreover, to address the extra delay introduced by the serial SC decoding, several studies suggest parallel processing of certain special nodes in SC decoding to reduce decoding delay, such as belief propagation (BP)-based parallel decoding for Polar code [188].

The BP decoding algorithm commonly used for LDPC code involves nonlinear functions in the computation of check nodes, resulting in high implementation complexity. Therefore, to meet the high-speed transmission requirements of 5G LDPC codes, several simplified alternative algorithms have been proposed, such as the widely used min-sum (MS) algorithm [189] and its various improvement schemes (collectively referred to as IMS) [190], which also come with a considerable performance loss, particularly in scenarios involving short and moderate block lengths and low code rate. Further reduction in the complexity of the LDPC decoder while ensuring decoding performance is still necessary to make it a realistic solution for the DSIN with Ubiquitous diversified services.

(2) Universal decoder. The maximum likelihood (ML) algorithm offers excellent BLER performance, but its huge decoding complexity makes it impractical for satellite platforms with limited payload. Similarly, most of the near-ML performance universal decoding algorithms, such as Guessing random additive noise decoding (GRAND) [191] and ordered statistics decoding (OSD) [192], also face these challenges. More precisely, while GRAND supports high-rate linear block codes effectively, it incurs very high decoding complexity for codes with moderate to low rates. Leveraging the assumption that errors are most likely to occur at the least reliable positions of received symbols, OSD identifies the most reliable basis (MRB) and flips a small number of bits within the MRB to regenerate a list of candidate codewords, from which it selects the best one as the output.

Researchers have recently proposed several methods to reduce their complexity. In the case of OSD algorithm, its decoding complexity mainly stems from the re-encoding and Gaussian elimination (GE) steps, which make its improvements broadly focus on two approaches: one involves avoiding the execution of GE in the OSD process [193], while the other introduces stopping or skipping criteria to reduce the number of test error patterns (TEP) evaluated during the search for the optimal candidate codewords [194]. There are still however challenges that require resolution. First, the above improved OSD algorithms help reduce decoding complexity in several ways, but they also require extensive parameter tuning, which complicates the process of achieving optimal BLER performance. Second, most OSD-based algorithms operate in iterations and lack effective methods for parallel decoding. For these issues, several studies suggest that the complexity can be further reduced through techniques such as the segmentation-discarding method [195], effective tree-based search algorithms [196], and cascaded BP decoding [197].

(3) Joint decoding of parallel multi-stream reception. Existing code-domain multi-satellite cooperative transmission technologies, including the MPA decoder for sparse code multiple access (SCMA) scheme [198], the BP-based q-ary decoder in LCMA scheme [199], and the joint decoder (JD) used in T-fold multiple access scheme [200], all suffer from high decoding complexity. For instance, the decoding complexity of JD is approximately $O(mT\log^2 T\log\log T)$, where m represents the message length. Moreover, since the signal quality is heavily influenced by the complicated and volatile satellite-terrestrial channels, further research on distributed decoding schemes that adapt to such communication conditions and achieve high-reliability decoding of superimposed coded signals in DSIN, especially in the CCS system is necessary.

3.6 Grant-free random access in satellite-terrestrial networks

Random access and multiple access technologies are essential to establish communication links in satellite-terrestrial networks [201]. The traditional four-step random access procedure is depicted in Figure 13(a). Initially, a user terminal randomly chooses a preamble from a set of sequences and transmits it as Msg1 over the physical random access channel to the BS. Upon detecting the transmitted preamble, the BS replies with a random access response (RAR), i.e., Msg2, which includes the random access preamble identifier (RAPID), a cell-radio network temporary identifier (C-RNTI), and the timing advance (TA)

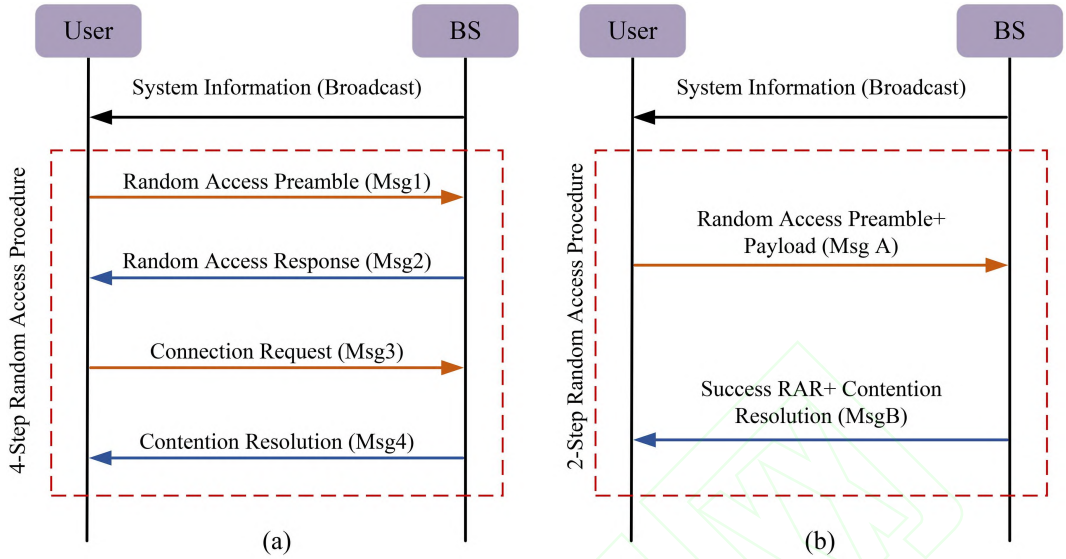

Figure 13 (Color online) (a) The four-step random access procedure in 4G LTE/5G NR; (b) the two-step random access procedure introduced in 5G NR.

information. Once the user receives Msg2 within a specified time window, it sends Msg3 to the BS, containing a connection request and terminal identifier for contention resolution. If multiple users employ the same preamble, the BS may not be able to decode Msg3 correctly. If successful, the BS transmits a contention resolution message, Msg4, which embeds the user's identifier. The terminal confirms successful access to the BS if it can correctly parse its identifier; otherwise, a new access schedule is attempted.

To decrease access delay, 3GPP Release 16 introduced a two-step random access process, illustrated in Figure 13(b). This approach merges the preamble sequence from Msg1 with the payload data from Msg3 into MsgA and merges the RAR from Msg2 with the contention resolution information from Msg4 into MsgB, reducing the interactions between the user and the BS. Nevertheless, both the four-step and two-step access methods depend on signaling interactions between the BS and terminals, classifying them as grant-based access. Besides, the existing 3GPP framework primarily uses orthogonal preamble sequences, which inevitably result in sequence collisions as the number of access terminals increases, reducing the probability of successful access.

In contrast to terrestrial cellular communication, terrestrial-satellite communication (TSC) typically experiences high propagation delays. Grant-free random access permits users to transmit pilot and data payload directly without BS coordination, thus satisfying fast access requirements in high-delay scenarios. Moreover, non-orthogonal pilot sequence allocation can reduce the pilot collision probability, thereby enhancing multi-user capacity within limited time-frequency resources. Therefore, grant-free NOMA presents an effective strategy for TSC.

In the context of grant-free access, the wide geographic coverage of satellite-based BSs leads to significant differential delays among users within the same service area, further complicated by high Doppler frequency shifts associated with relative movement. Consequently, time-frequency synchronization and pre-compensation are essential before transmission to ensure. Unlike grant-based access where the BS calculates and sends TAs to users based on preamble sequences, grant-free access permits multiple users to independently perform timing advances and frequency offset pre-compensation by utilizing the broadcast signal from the BS, which enhances efficiency. When terrestrial terminals have precise ephemeris information and positioning capabilities, they can compute the relative distance and velocity between the user and the satellite, facilitating time-frequency offset pre-compensation for synchronization. If user terminals lack positioning capabilities or cannot access ephemeris data, the method leveraging downlink synchronization signal block for time-frequency offset estimation, as proposed in [202], can convert differential time-frequency estimates into the locations of either user terminals or satellites to achieve synchronization.

In NOMA scenarios, compressive sensing-based techniques for active user detection and channel estimation can effectively leverage user activity sparsity to reduce the required pilot length. To ensure optimal sparse recovery performance, designing non-orthogonal pilot sets with low correlation is crucial

for minimizing user interference. A method for constructing pilot sets based on Golay sequences is proposed in [203], which successfully minimizes pilot correlation and the peak-to-average power ratio (PAPR) of spreading codes in multicarrier transmission systems. This approach provides lower PAPR while maintaining robust sparse recovery performance compared to random binary, Gaussian, pseudo-random, and Zadoff-Chu (ZC) sequences. Authors in [204] presented a deterministic scheme for non-orthogonal pilot construction using a discrete Fourier transform (DFT) matrix with a mask. This method can generate $\mathcal{O}(L^3)$ non-orthogonal sequences within a length-L sequence and offers theoretical guarantees on inter-sequence correlation bounded by $\mathcal{O}(\frac{1}{\sqrt{L}})$. This enlarges the pilot set size and outperforms Gaussian random pilots by reducing collision probability. In [205], a composite preamble construction method is proposed. By combining orthogonal ZC sequences with multiple ZC root sequences with different phase rotations, this approach extends the pilot set, thereby differentiating users selecting the same orthogonal pilot. An iteratively constructed pilot sequence is proposed in [206] using conjugate gradient descent and space projection, subject to the constraint on the PAPR. The pilot sequence proposed in [206] can achieve up to 47% and 28% lower coherence compared with the binary Golay sequence and the traditional ZC sequence, respectively. Moreover, a low-correlation-zone periodic sequence (LPS) is designed for the superimposed pilot structure in [207]. The average auto-correlation and cross-correlation of the LPS is $\mathcal{O}(1/L)$, which is $1/\sqrt{L}$ times lower than that of the multi-root ZC sequence, and results in a reduced decoding failure probability in satellite-based massive access systems. It is worth noting that the expansion of the non-orthogonal pilot sequence set not only alleviates access conflicts but can also be applied to pilot index modulation [208] and superimposed transmission [207], thus enhancing spectral efficiency.

Regarding active user detection and channel estimation for LEO satellite uplink access, authors in [209] addressed satellite-terrestrial link channel modeling, incorporating phase shifts and channel impairments. A Bernoulli-Rayleigh message-passing algorithm is proposed, which leverages the sparse activity of ground users to reduce pilot overhead. Acknowledging the wideband transmission needs of terminal devices and the high-delay, fast-varying characteristics of satellite-terrestrial links, Ref. [67] introduced an OFDM symbol repetition mechanism. This approach addresses residual time and frequency offsets in LEO satellites, proposing an enhanced variance state propagation algorithm for joint active user detection and channel estimation. To accommodate rapidly changing satellite-terrestrial channels, Ref. [210] applied OTFS modulation for massive multi-user access, delineating a two-stage process for joint active user detection and channel estimation followed by data detection. However, previous studies have primarily focused on global navigation satellite system (GNSS)-based solutions, which are unsuitable for power-limited time-sensitive communications (TSC). In response to this, Ref. [211] proposed a joint design for device identification, channel estimation, and symbol detection in LEO satellite-enabled GFRA systems, without relying on GNSS assistance. This approach uses OTFS modulation with a message-passing-based algorithm to handle large differential delays and Doppler shifts at the satellite receiver. Additionally, algorithms based on block coordinate descent, tensor Bayesian learning, and approximate message passing are also extensively employed for compressed sparse recovery problems and can be effectively utilized in uplink signal estimation and detection for TSC.

To meet the extensive data transmission demands of terrestrial users, implementing NOMA for overload transmission is an effective strategy to enhance multi-user capacity. According to [212], a code-domain NOMA scheme based on low-density signatures (LDS) is designed. This scheme exploits the structured sparsity of transmitted signals and user activity to develop a Gaussian-approximated message-passing-aided sparse Bayesian learning algorithm, achieving superior bit error rate and false alarm rate performance compared to existing approximate message-passing algorithms. In [213], an SCMA scheme using affine frequency division multiplexing is proposed. This approach modulates sparse codewords onto chirp subcarriers and discusses in detail the design of sparse codes, chirp rate selection, and receiver algorithms. The method demonstrates significantly better performance than OFDM-based SCMA in high mobility scenarios, underscoring its potential application in multi-user access scenarios for LEO satellites.

Another research branch of GRRA focuses on improving access timeliness. It is worth noting that large-scale user device access conflicts, along with access failures and the occurrence of multiple retransmissions, can lead to the accumulation of long queue buffers at IoT devices, significantly degrading the timeliness of information [214], which can be quantified using the age of information (AoI) metric [215]. In [216], the instantaneous AoI evolution of each device is traced through Markov analysis to derive the expression for the average AoI of the access system. A grant-free age-optimal random access protocol

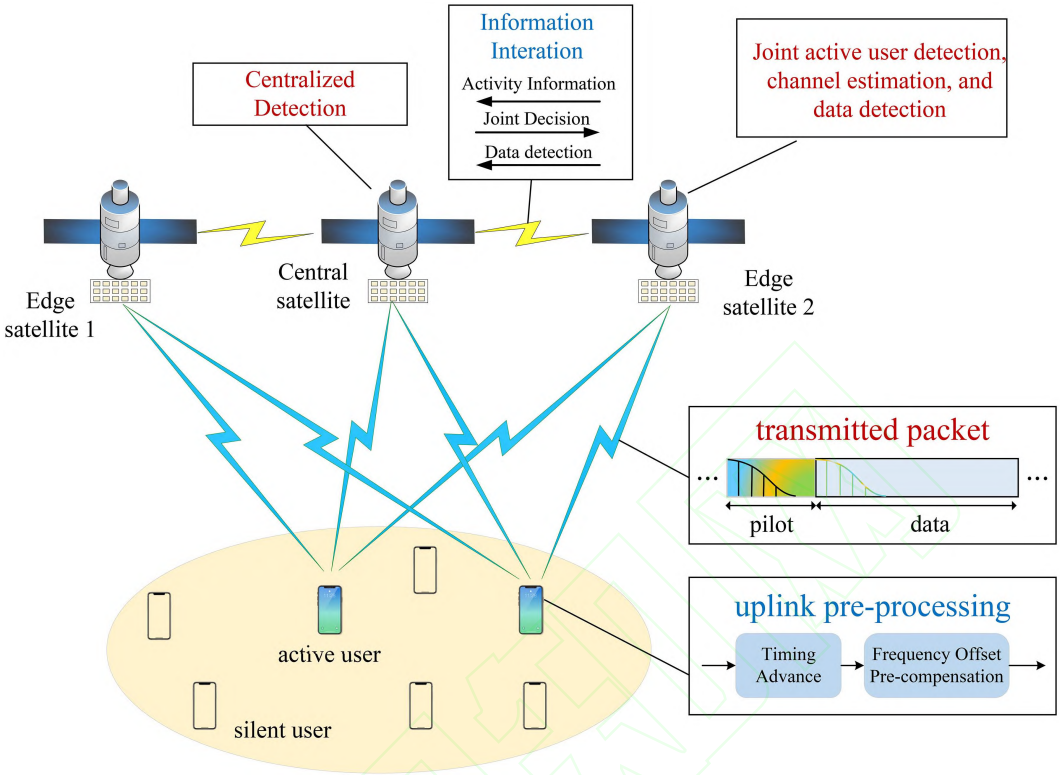

Figure 14 (Color online) Diagram of multi-satellite cooperative grant-free random access.

is then proposed to dynamically adjust the number of access slots, optimizing the system's average AoI while enhancing the maximum throughput. Furthermore, considering that satellite networks need to accommodate service devices with diverse timeliness requirements, Ref. [217] proposed the unequal timeliness protection massive access scheme for mission critical communications in S-IoT. This scheme aims to minimize the average AoI for user groups with different timeliness demands, ensuring reliable access under their respective reliability constraints.

In multi-user TSC, a constellation comprising multiple satellites extends coverage and design flexibility beyond a single satellite system. This configuration offers spatial diversity and multiplexing gains owing to multiple observations in the spatial domain. As depicted in Figure 14, in complex scenarios where multiple users simultaneously access multiple satellites, precise multi-channel estimation strategies, and multi-user detection mechanisms with multi-satellite collaborative data demodulation strategies are vital for achieving reliable massive access. The authors in [218] proposed a novel multi-satellite collaborative user detection scheme for traditional slotted ALOHA-like random access methods. By modeling the communications between multiple users and multiple satellites as an equivalent distributed MIMO system, the proposed multi-satellite cooperative MIMO detection algorithm significantly enhances the success detection probability of collided data packets when compared to the single-satellite scheme. The authors in [219] introduced an advanced multi-satellite cooperative scheme for active user detection and data detection. By capitalizing on the observation diversity gain provided by multiple satellites, this scheme substantially improves the accuracy of active user detection and effectively mitigates the performance degradation in data detection caused by spatial channel correlation among active users. In addition, Ref. [134] proposed a MIMO-OTFS modulation-based multi-satellite collaborative random access scheme. Initially, it derives users' initial channel estimates at each satellite through a message-passing algorithm and subsequently aggregates feedback information from multiple edge satellites to achieve enhanced activity detection, and data detection outcomes. Moreover, the proposed approximate expectation propagation method can be executed either centrally by a single satellite or distributed across multiple satellites. Notably, the distributed algorithm requires only two rounds of information exchange to achieve performance levels comparable to those of centralized processing.

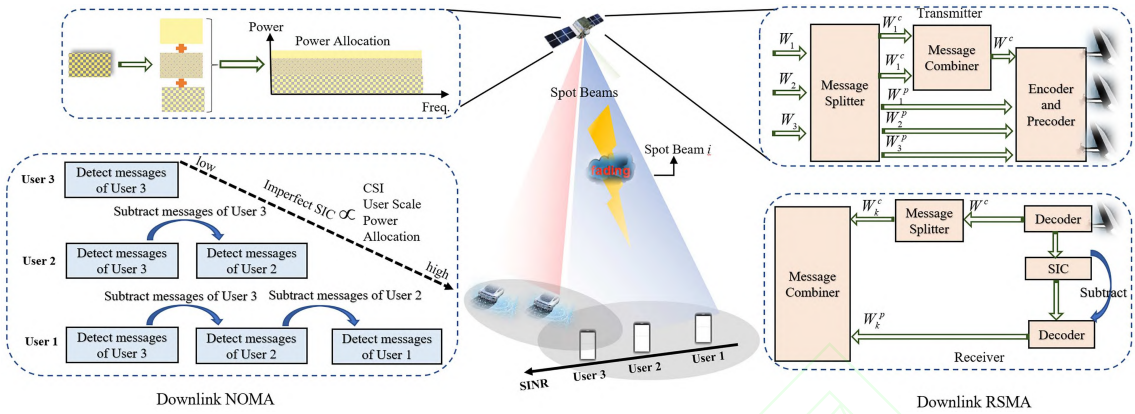

Figure 15 (Color online) Illustration of downlink NOMA and RSMA enabled satellite communication system.

3.7 NOMA/RSMA enabled multi-satellite multicast

3.7.1 *NOMA enabled multi-satellite multicast*

The rapid evolution of mobile networks towards 5G and beyond has driven the need for highly efficient multiple-access techniques capable of handling the ever-increasing demand for wireless connectivity. Among these techniques, the NOMA stands out as a promising solution due to its ability to support multiple users on the same frequency and time resources by exploiting power domain multiplexing [220]. Unlike traditional orthogonal access schemes such as TDMA and FDMA, NOMA enables simultaneous access for multiple users, making it particularly well-suited for scenarios where resources are scarce, such as SatCom. As NOMA evolves, the integration of NOMA into SatCom systems, particularly in downlink multicast communication of CCS system, presents a significant opportunity to enhance service quality, improve spectral efficiency, and support a larger number of users [221]. As shown in Figure 15, by enabling superposition transmission, the NOMA operates by allowing multiple users to share the same time-frequency resources, with user separation achieved through the power domain, thus enhancing the overall system capacity and the overall spectral efficiency of satellite networks. This is particularly beneficial for CCS system, which often serve a wide geographical area with diverse user demands.

Research on NOMA in satellite multicasting has seen significant progress. Early studies focus on the theoretical performance of NOMA in satellite channels, demonstrating its superiority over orthogonal multiple access (OMA) in terms of user capacity [222], spectral efficiency [223] and user fairness [221]. Subsequent studies delve into the practical aspects of implementing NOMA in satellite systems, including the design of precoding techniques to mitigate inter-beam interference [224], the development of user scheduling algorithms to optimize resource allocation [225], and the improvement of information freshness to support time-critical services [226]. The application of NOMA in satellite multicasting has also been explored in the context of high-throughput satellite (HTS) systems, where the combination of NOMA with advanced beamforming techniques has been shown to significantly improve the throughput performance [227]. Moreover, the potential of NOMA in cognitive radio-inspired satellite networks has been investigated, where NOMA principles are applied to enhance spectrum sharing between primary and secondary users [228]. Recent research has also addressed the challenges of limited feedback and imperfect CSI in NOMA-enabled SatCom systems, proposing robust designs to ensure reliable communication in the presence of channel uncertainties [229]. Additionally, the application of NOMA in multi-beam satellite systems has been explored, aiming to improve the sum capacity throughput by using the joint beamforming and power allocation design [230]. Other studies have examined the integration of NOMA with other advanced satellite technologies, such as cognitive radio [231] and DRL [232]. Cognitive radio enables dynamic spectrum sharing between satellites and terrestrial networks, while DRL can be used to optimize resource allocation in the NOMA-enabled CCS system. Over time, the focus shifted towards more complex satellite architectures, such as LEO constellations and CCS system. In particular, multi-satellite systems, where multiple satellites collaborate to provide downlink services to a common set of users, have become a key area for NOMA deployment. The transition to multi-satellite systems necessitates new approaches to address inter-satellite interference, and optimize decentralized resource allocation [233] under the constraint of asynchronous cooperative communication [234].

While NOMA has shown great potential in enhancing downlink multicast communication for DSIN, several challenges remain. One of the key challenges is managing the complexity of resource allocation in DSIN. As the number of satellites in a CCS system increases, the coordination of power allocation, beam hopping, and user scheduling becomes more complex. It is necessary to develop more efficient algorithms for managing these resources in large-scale NOMA-enabled DSIN. A DL-based MIMO-NOMA framework is proposed in [235] to maximize the sum data rate and energy efficiency. As an evolution of NOMA, hybrid NOMA offers superior compatibility, enabling effective coexistence between OMA and NOMA. The optimality of downlink hybrid NOMA in the two-user case has been invalidated in [236]. Furthermore, a scalable hybrid NOMA scheme is proposed in [237] to jointly control power allocation and bandwidth overlap in the presence of both imperfect CSI and SIC, thereby improving the sum rate in a spectrally efficient manner. Additionally, leveraging the spectral efficiency advantages of NOMA, reconfigurable intelligent surfaces (RIS)-aided NOMA systems have garnered significant attention for maximizing the communication success ratio, while meeting the diverse QoS requirements of users and the dynamic energy constraints at the RIS [238]. In particular, the RIS-aided NOMA scheme can significantly enhance the sum rate under finite-block length coding by jointly optimizing user power allocation and RIS beamforming [239]. Moreover, most of the existing PHY signal processing methods are limited to the study of two-user NOMA systems with limited satellites. Extending those solutions to large-scale networks with multiple satellites in a cluster or a swarm is of great importance. Another challenge is the management of inter-satellite interference. In CCS systems, satellites often operate in close proximity, leading to increased interference. Techniques such as multi-agent deep reinforcement learning (MADRL) enabled cooperative NOMA and beamforming can be seen as a good avenue for future research to mitigate the interference issue. In addition, as a common consumption in existing research work, perfect SIC processing may lead to overestimating the performance of the NOMA-enabled CCS system, which can be addressed in future research.

3.7.2 *RSMA enabled multi-satellite multicast*

The integration of rate-splitting multiple access (RSMA) with satellite communication systems has emerged as a promising approach to enhance spectral efficiency and energy efficiency, particularly in multi-satellite downlink broadcasting CCS system. The RSMA is a novel NOMA scheme that has been gaining traction in the SatCom domain. Unlike traditional NOMA, which focuses on power domain multiplexing, the RSMA splits the messages into common and private parts, allowing more flexible interference management as shown in Figure 15. This approach has been shown to outperform conventional schemes in terms of spectral efficiency and energy efficiency, especially in multi-beam, multi-group multicast scenarios.

The primary distinction between the RSMA and the aforementioned NOMA lies in the message splitting strategy. The RSMA and NOMA both aim to maximize spectrum efficiency but differ fundamentally in their approaches. While the NOMA relies on successive interference cancellation (SIC) to decode signals from multiple users, the RSMA treats part of the interference as noise and decodes another part via SIC, which results in greater flexibility and robustness. This ability to combine noise cancellation and SIC allows RSMA to outperform NOMA, especially in scenarios involving imperfect CSI or complex satellite networks where multiple downlink beams may overlap. The flexibility of RSMA in handling various levels of interference makes it superior to NOMA for CCS system, particularly in high-interference regimes where NOMA's performance degrades due to the complexity of interference management across multiple beams. An exhaustive survey of the RSMA combined terrestrial communication literature from both the information-theoretic and communication-theoretic perspectives is presented by [240].

Inspired by the advantages of RSMA-aided multigroup multicasting in terrestrial networks, the deployment of RSMA in the realm of DSIN is intriguing and shows great potential. Specifically, in multi-satellite downlink broadcasting, the RSMA enables managing inter-beam and inter-satellite interference effectively, which is a significant advantage over the NOMA. Several key technologies make RSMA a feasible option for the CCS system. (1) Massive MIMO: the RSMA leverages multi-antenna systems to deliver high spectral efficiency and robust communication links [241]. This is particularly useful in DSIN, where beamforming and spatial multiplexing are critical. The practical applicability of the RSMA in downlink multi-antenna communications is demonstrated by utilizing the physical layer design and link-level simulations [242]. (2) Intelligent reflecting surfaces (IRS): the RSMA can be combined with IRS to dynamically control signal propagation, improving coverage and reducing interference in DSIN. As a result,

the capacity of SatCom can be improved significantly by integrating the RSMA and IRS, as explored by [243]. (3) Precoding and beamforming: the RSMA requires sophisticated precoding techniques to optimize the transmission of common and private streams across satellite networks. This ensures minimal interference between beams and users, enhancing the overall system performance. The statistical beamforming and common stream separation based on RSMA are incorporated to maximize the minimum user rate under the total power budget of the transmitter [244], which demonstrates RSMA's robustness against the inaccuracy of CSIT and its superiority over SDMA in terms of explicit max-min fair rate gain. In addition, some studies have also begun to integrate RSMA with AI-driven optimization algorithms to manage complex satellite constellations dynamically, opening up new possibilities for automated satellite network management and enhanced performance under variable conditions [245]. Joint optimization of resource allocation and power control for RSMA-based LEO satellite-terrestrial networks is investigated in [246] by employing DRL to maximize energy efficiency. Further, the concept of distributed RSMA has opened new avenues for research in DSIN. To enhance spectral efficiency and manage interference in complex satellite networks, Ref. [247] introduced a distributed RSMA approach for multi-layer satellite systems. More recently, the RSMA-enabled multi-satellite multi-beam communication systems have been investigated. Ref. [248] addressed the enhancement of sum-rate in multi-beam satellite systems by employing rate-splitting precoding, enabling simultaneous transmission of public and private frames, and proposed a low-complexity design that demonstrates significant performance gains. The RSMA is also proven to be promising for addressing practical challenges such as feeder link interference and imperfect CSIT in multigateway multibeam satellite systems [249].

Despite its promising capabilities, several challenges remain for the deployment of RSMA in DSIN. (1) Complexity of SIC in large networks: while the RSMA simplifies SIC compared to the aforementioned NOMA technology, the complexity can still increase in large-scale satellite constellations, requiring further research into low-complexity decoding algorithms. (2) Interference management: managing interference across a CCS system with overlapping beams remains a significant challenge. Future work should focus on optimizing RSMA for highly dynamic satellite networks where user mobility and environmental factors can create varying interference patterns. (3) Integration with emerging technologies: the integration of RSMA with emerging technologies such as AI, removable antennas, and ultra-massive MIMO presents both opportunities and challenges. Research is needed to ensure these technologies can work seamlessly in a large DSIN.

3.8 Erasure transfer protocol in DSIN

In contrast to relatively stable terrestrial networks, the dynamic formation configurations, non-trivial round-trip propagation delays, and significant long-distance transmission losses in the DSIN substantially impact and constrain the performance of conventional terrestrial communication protocols, such as the widely used transmission control protocol/Internet protocol (TCP/IP), which may misinterpret the packet loss under poor CSI as an indication of network congestions [250, 251]. To tackle the issues that TCP/IP encounters in space communications, various solutions have been proposed at both the transport layer and other layers. The most widely adopted schemes are the CCSDS protocol suite represented by the space communication protocol standard (SCPS) [252], and delay/interrupt tolerant networking (DTN) protocol suite, represented by the Bundle protocol and Licklider transmission protocol (LTP) [253]. Currently, the SCPS protocol has been downgraded to historical status, while the LTP protocol, benefiting from its no-handshake operation and delayed acknowledgement-based retransmissions, is widely applied in the satellite-terrestrial and ISLs characterized by long-distance and high-interruption probability, and has been adopted within the CCSDS protocol suite [254].

Achieving freshness and reliable information delivery required by mission-critical applications with the LTP protocol has long been a core area of research in SatCom, where reliability serves as a critical prerequisite for ensuring timely delivery. However, the physical layer (PHY) forward error correction (FEC) approaching the Shannon limit still cannot fully ensure reliable packet-level transmission over satellite-to-terrestrial channels. Indeed, to maintain a low BER and even BLER level, these FEC schemes typically require a lower coding rate, which results in limited throughput and compromised timeliness [255, 256]. Consequently, link-layer error control methods, particularly retransmission and redundancy mechanisms, have been extensively explored as alternative solutions to achieve cost-effective error-free reception. The key techniques of these two mechanisms and their application in satellite transfer protocols are illustrated in Figure 16.

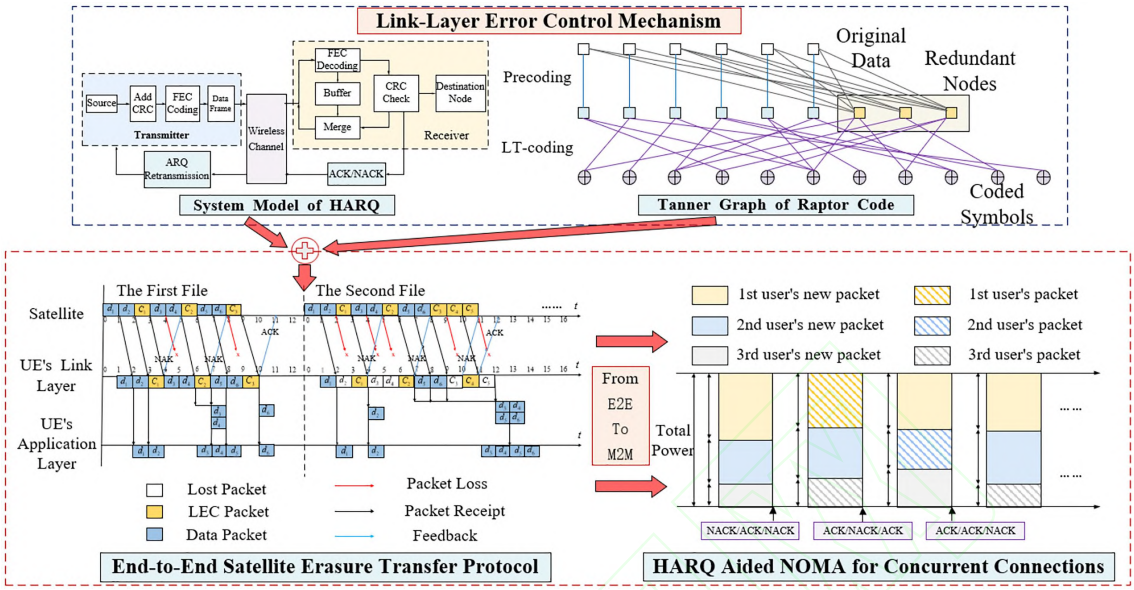

Figure 16 (Color online) Overview of the introduced satellite erasure transfer protocols.

3.8.1 *Current link-layer error control techniques*

(1) Low-delay retransmission mechanism. The classic retransmission mechanism of the LTP protocol, automatic repeat request (ARQ), enables retransmission of uncorrectable frames through feedback from the receiver to the transmitter, but multiple retransmissions can significantly amplify the high-delay effects of long-distance transmissions [257]. A more advanced error control method is hybrid ARQ (HARQ), which combines ARQ with FEC to improve reliability by enabling both frame correction and retransmission of uncorrectable frames. In contrast to Type I HARQ, which discards the uncorrectable received frames and performs memoryless decoding for each single encoded frame at the receiver, Type II HARQ employs frame combining to improve the utilization of transmitted frames and enhance detection performance. It can be further classified into two types: chase combining (CC) [258], which retransmits identical coded frames, and incremental redundancy (IR) [259], which retransmits additional parity bits.

Given that throughput, outage probability, and the number of average retransmissions for conventional HARQ have been extensively studied in terrestrial networks [260], this subsection focuses on the timeliness issues of HARQ in SatCom. More precisely, the feedback-based retransmission mechanism in standard HARQ renders it a high-delay scheme in the context of long-distance transmission. To minimize long delay, various enhancements to the HARQ protocol have been proposed, such as fast HARQ [261] and dynamic HARQ [262]. The fast HARQ reduces delay by omitting certain feedback signals and allowing the receiver to decode only when the estimated overall channel gain is sufficiently high, while the dynamic HARQ adaptively adjusts the number of retransmissions for each packet based on the decoding condition of the previous one. Further taking into account the delayed feedback, a novel HARQ protocol called early HARQ has been proposed to mitigate this issue by introducing predictive mechanisms [263]. Specifically, the receiver predicts the decoder outcome before actual decoding by comparing its BER estimate with an empirically calculated threshold and then generates early acknowledgment (ACK)/non-ACK (NACK) or uncertain feedback to reduce feedback delays. Moreover, the authors in [264] introduced machine learning to improve the early HARQ, and demonstrated that the enhanced decoding prediction can significantly reduce estimation errors and redundant retransmissions compared to the empirical threshold-based scheme.

(2) Link-layer erasure-based redundancy. The long erasure codes (LEC) specification [265] published by CCSDS is an effective erasure scheme, which utilizes a packet-level LEC to alleviate packet loss at the PHY, i.e., an opportunity to recover packets that failed to decode at the PHY. Typical LEC schemes, such as RS code, LDPC code, and fountain code can reconstruct the original information from any k received packets by adding m redundant erasure packets to k original data packets using their respective coding methods. Since link-layer RS codes are mainly used in storage systems and LDPC codes are more commonly used for PHY error correction, we focus on the fountain code with the advantages

of rate adaptability, no feedback retransmission required, and low encoding/decoding complexity [266]. From this perspective, fountain code is also a promising scheme for satellite communications featuring time-varying channels, long feedback delays, and limited payload of space segments.

Luby transform (LT) code was the first to implement the concept of fountain code [267], which generates an arbitrary number of encoded packets by randomly selecting and XORing data packets from the original data, and allows the receiver to decode the complete information once a sufficient number of encoded packets are received. To address the issue of unnecessary redundant packets generated by the random selection of data packets in LT code, which increases transmission overhead and reduces decoding efficiency, researchers have proposed Raptor code by adding concatenated precoding on top of LT code [268]. Compared to LT code, Raptor code not only reduces the number of encoded packets required for decoding but also decreases the decoding complexity at the receiver, and has been standard for DVB handheld and 3GPP multimedia broadcast/multicast service (MBMS) [269]. To improve the performance of Raptor code, the authors in [270] introduced the feedback mechanism of HARQ to optimize the degree distribution function, and proved that the feedback Raptor codes can achieve linear complexity with only a slight increase in decoding overhead. Moreover, the rate adaptability of Raptor code allows flexible allocation of redundant packets across different nodes or transmission paths, which means that even if some nodes fail, the original data can still be recovered [271]. Therefore, the Raptor code is highly adaptable for the DSIN, where the distributed Raptor code can greatly improve decoding probability and system fault tolerance, as well as reduce feedback delay.

3.8.2 *Satellite erasure transmission scheme*

Network coding (NC) applies a similar concept to LEC at the network layer, which enhances the transmission efficiency and reliability by forwarding data packets that are linear combinations of the original ones. Random linear NC (RLNC), which selects coding coefficients randomly over a finite field, can theoretically achieve the maximum network flow [272, 273]. To lower the encoding and decoding complexity of NC, the authors in [274] proposed an excellent RLNC scheme that integrates fountain code to achieve both low computational complexity and high transmission efficiency, in which the outer code is typically a matrix-based fountain code, and the inner code uses RLNC to encode data within the same block. Leveraging the advantages of the above LEC scheme, where the receiver can recover the k lost original packets with high probability and low decoding complexity when it successfully receives k NC packets, the authors in [275] proved that the NC aided HARQ (NC-HARQ) enables energy-efficient transmission in satellite-to-terrestrial downlink communication.

Further, recent studies have explored the timeliness of NC-HARQ erasure protocols, which further incorporate AoI as a measure of information freshness [215]. The authors in [276] investigated the NC-HARQ with limited or no feedback retransmission for satellite communications, showing that limited retransmission optimises delay, while no retransmission yields improved AoI. An excellent erasure scheme named adaptive NC-inserted HARQ was proposed in [277] to further improve timeliness, where NC packets are dynamically inserted into the data stream based on delayed CSI to accelerate packet recovery at the receiver. The above scheme was extended to a dual-hop satellite relay scenario in [278], aiming to achieve optimal AoI and delay through a combination of adaptive CSI-based relay modes with adjustable LEC packet insertion intervals. Further enhancement of SatCom can be obtained through cross-layer optimization of error control methods. For example, the authors in [279] proposed a dual-layer coding scheme that jointly optimizes link-layer LEC and physical-layer FEC. By balancing redundancy between the LEC and FEC, i.e., finding their optimal coding rate, this erasure transfer protocol enables age-critical and reliable transmission in satellite networks. Similarly, designing a cross-layer optimization framework between the link and transport layers is another promising research direction. The benefit of this approach is that the TCP variants and LTP protocols can regard the satellite-terrestrial channel as a transparent transmission path by leveraging link-layer error control to mitigate packet loss [280]. In this case, many congestion control algorithms can be applied to the satellite network effectively, further enhancing the actual transmission performance.

It is worth noting that the above schemes are designed for point-to-point transmission and are not applicable in DSIN with multiple points at least on one end. In response, a satellite transmission scheme that combines HARQ, cooperative communication, and NC is proposed in [281], which enables concurrent multi-stream data transmission through cooperation between multiple user antennas, thus reducing retransmissions and increasing throughput. In more complex scenarios involving multiple users or satel-

lites, recent studies suggest integrating HARQ with NOMA to achieve age-critical, high-throughput and reliable transmission in multi-node concurrent connections. The outage performance [282], diversity order [283], average throughput [284], and BLER performance [285], have been thoroughly analyzed in the finite block length for HARQ aided NOMA (HARQ-NOMA) system. The obtained closed-form expressions, approximations, or bounds of the above metrics show that the HARQ-NOMA system can significantly improve the performance of conventional NOMA schemes by adjusting the power level of users during retransmissions to reduce the number of attempts.

Recent studies have further analyzed and optimized the AoI of the HARQ-NOMA system, where the results showed that HARQ can mitigate information ageing due to retransmissions by enhancing the reliability of NOMA [286]. Moreover, to further improve transmission efficiency, the authors apply DRL for the HARQ-NOMA system in [287] to derive an age-optimal scheme integrating intelligent power allocations and retransmission decisions. In summary, most studies on HARQ-NOMA systems lack the application of link-layer LEC, and the erasure transfer protocols for DSIN still need further investigation.

3.9 High-speed inter-satellite communication and network routing

The proliferation of LEO satellite constellations has ushered in a new era of DSIN, promising extensive coverage, unprecedented connectivity and intelligent informative services. At the core of this paradigm shift is the inter-satellite high-speed communication technology, efficient routing of data and congestion control through the constellation, which demands innovative solutions to address the dynamic and complex nature of these networks.

3.9.1 *High-speed inter-satellite communication*

The DSIN requires high-speed inter-satellite communication to achieve global ubiquitous coverage, but the increasing demand for inter-satellite communication bandwidth, coupled with the instability of free-space channels, presents new challenges in DSIN. As the communication capacity required for satellites and various spacecraft grows exponentially, the current communications based on RF are becoming insufficient to meet the sharply rising demand for capacity [288]. Inter-satellite free space optical (FSO) communication has emerged as a promising alternative or complementary solution to traditional RF systems, owing to its vast unlicensed bandwidth and the ability to transmit data at exceptionally high rates over long distances [289]. Moreover, in contrast to conventional RF-based wireless systems, the narrow and directional nature of the laser beams used in FSO communications offers enhanced security, reduced power consumption, and immunity to electromagnetic interference. Constellation systems such as Starlink, OneWeb, Kuiper, and Telesat have already adopted inter-satellite FSO communication as one of their core transmission links. Nevertheless, despite the great potential of FSO communication for DSIN, its performance suffers from various limitations and challenges. On the one hand, the FSO communication is susceptible to the detrimental effects of atmospheric turbulence, including beam-wandering-induced pointing errors, beam scintillation, and attenuation caused by weather conditions such as haze, snow, fog, and clouds [288]. Specifically, pointing errors induced by beam wandering and misalignments between receivers and transmitters can severely impact the reliability of the link. To address these challenges, the hybrid FSO/RF scheme within satellite networks presents a promising solution, leveraging the complementary nature of both technologies to enhance link quality without causing interference between the FSO and RF links [290]. Relay-assisted transmission via high-altitude platforms (HAPs) offers another promising solution to enhance the usability of FSO links. For example, unmanned aerial vehicles (UAVs) can be deployed between satellites and GSs to establish a structured space-air-ground FSO network within DSIN, providing high reliability and high-speed transmissions [291]. This approach has already been explored in [292] and has been validated by NASA's laser communications relay demonstration (LCRD) mission, which successfully transmitted data from GEO at a rate of 1.2 Gb/s. However, the performance of FSO communication significantly deteriorates as the propagation distance increases, particularly in environments with strong turbulence.

On the other hand, the successful implementation of long-distance FSO communications in DSIN largely hinges on the performance of the pointing, acquisition, and tracking (PAT) system [293]. The PAT system is critical for ensuring accurate alignment and maintaining stable communication links over vast distances, especially in the presence of atmospheric disturbances and dynamic satellite movements. FSO communications can mitigate pointing errors, acquire incoming light signals, and maintain stable links during missions by continuously monitoring system-wide performance metrics, such as received

signal power and Strehl ratio, and dynamically adjusting correction elements like gimbals, mirrors, or adaptive optics. Pointing refers to the process of aligning the transmitter within the receiver's field-of-view (FOV), while signal acquisition involves aligning the receiver with the arrival direction of the beam [294]. Tracking ensures the ongoing maintenance of both pointing and signal acquisition throughout the optical communication link between satellites. To achieve high data rates through concentrated light intensity and the extended reach of narrow beams, various PAT mechanisms have been proposed, including gimbal-based, mirror-based, gimbal-mirror hybrid, adaptive optics, liquid crystal, RF-FSO hybrid, and other PAT approaches, as detailed in [293]. With the decreasing size of satellites, UAVs, and airborne micro-stations, PAT mechanisms are anticipated to become smaller and less complex, necessitating innovative ATP designs. For example, mirror-based PAT mechanisms struggle to maintain a stable optical link between a GS and an LEO satellite travelling at 7.5 km/s for more than 1 ms. In this context, developing an agile PAT system is both essential and challenging to ensure the stability and reliability of FSO links in DSIN, where network outages are either impermissible or must be anticipated proactively. To explore the feasibility of FSO communications in near-Earth 6G NTNs, a baseline PAT system for vertical FSO links equipped with conventional detectors and actuators has been introduced in [294]. This system demonstrates improved robustness and agility through the incorporation of AoA estimation and retroreflectors, showcasing significant performance gains. Moreover, current laser terminal control typically relies on a one-to-one mechanical beam-linking approach, which results in prolonged access times. Additionally, since the beam direction is fixed to a single trajectory, this method is limited to point-to-point communication and cannot establish links with multiple terminals simultaneously, falling short of the requirements for efficient inter-satellite networking. To address this limitation, researchers have proposed an optical spherical head array [295], which enables electronically programmable control of optical head directions to achieve rapid beam angle adjustments. This approach ensures swift angle switching with high resolution, presenting significant potential for advancing laser-based inter-satellite networking in the future.

Further, in DSIN, small satellites can significantly enhance information fusion and high-speed transmission capabilities through distributed cooperation in communication and computation. However, achieving precise synchronization in absolute phase, frequency, and time remains a formidable challenge when the clock or local oscillator signals are generated locally at each distributed node, which has received plenty of attention in [227, 296]. As previously discussed for distributed MIMO collaboration, the GSs must maintain synchronized clocks with sub-nanosecond accuracy to support bandwidths of several hundred MHz, thereby ensuring symbol-level synchronization. Moreover, the relative motion of satellites in DSIN, along with varying and long distances, further complicates signal synchronization in diverse space science tasks such as remote sensing, inter-satellite ranging, and relative positioning. A non-collaborative CFO estimation technique is proposed in [297] to address synchronization challenges in distributed mMIMO mobile communication systems. Additionally, methods such as closed-loop, open-loop, master-slave, and consensus-based approaches have been suggested for implementing frequency and phase synchronization, as detail reviewed in [45]. Furthermore, in recent years, ML techniques have garnered widespread academic interest in addressing various synchronization issues in end-to-end communication systems, including frame synchronization, compensation for sampling frequency/time offsets, and phase noise characterization. In this regard, the authors in [298] proposed a CNN-based synchronization model with a softmax activation function to detect the actual position of a frame header, while the authors in [299] employed multi-instance learning to solve frame synchronization problems across different frequency ranges without requiring additional modifications. However, due to the limited on-orbit computational power, ML models may struggle with poor convergence rates, especially under varying transmission conditions. Therefore, there is a pressing need to investigate optimal model synchronization technologies among small satellites in a DSIN to improve the accuracy of ML training models and expedite training times.

3.9.2 *Distributed network routing*

The pursuit of ubiquitous global Internet connectivity has driven the development of satellite constellations, which aim to provide seamless coverage across the globe. With the increasing number of satellites in orbit, the routing of data within these constellations has become a critical area of research. To achieve efficient routing in satellite networks, researchers in the space industry are delving into network routing design, predominantly decentralized and centralized, to offer choices for the multi-satellite cooperative backhaul. The centralized routing approaches are mostly computed on the ground and distributed to the

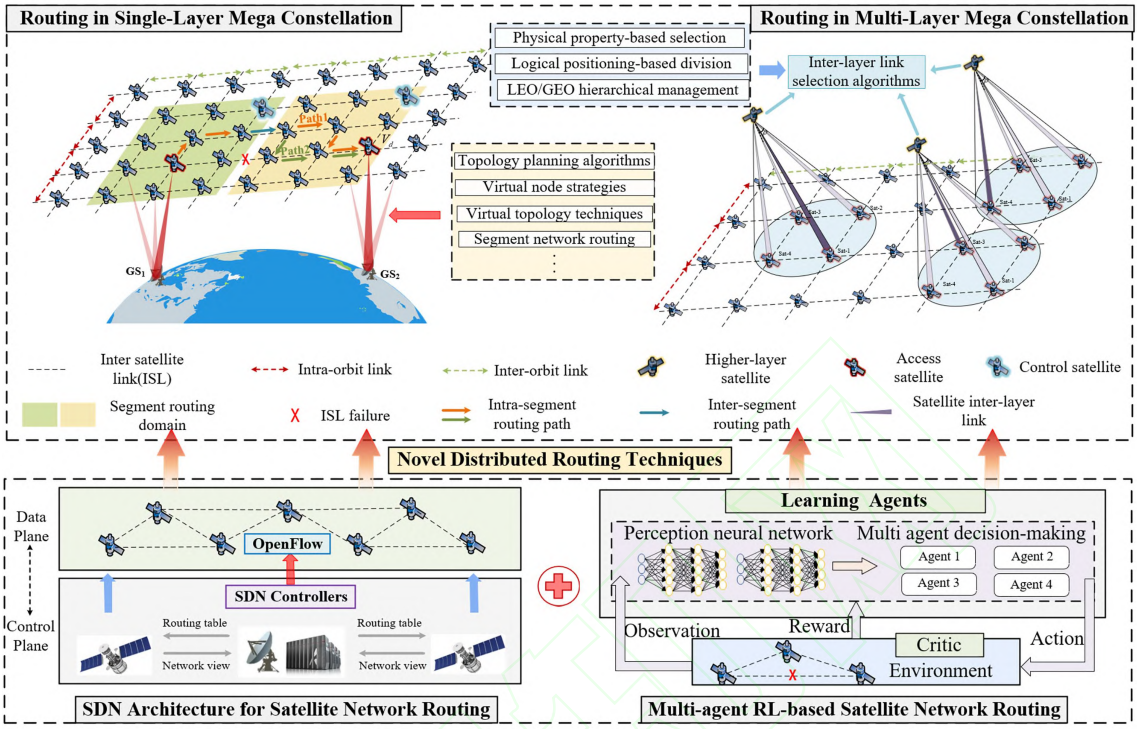

Figure 17 (Color online) Overview of distributed network routing and congestion control for DSIN.

satellites. Early routing algorithms, such as Dijkstra's shortest path algorithm, were adaptations of centralized routing algorithms and were not designed for the dynamic nature of satellite networks [227]. The stale information on the traffic status can result in suboptimal routing algorithms and valueless routing updates in CCS systems. The ongoing advancement of space-based networking technologies, including space-based routing and ISL, is critical for meeting the substantial data transmission requirements associated with S-IoT services. It also spurred the transition of satellite constellation routing from centralized to distributed approaches. As shown in Figure 17, the distributed routing enables to calculation of the routes in orbit as the satellites are typically equipped with onboard computation and storage resources, thus guaranteeing the timely utilization of the latest traffic queuing status. Distributed, heterogeneous satellite constellations that integrate communication, navigation, and remote sensing capabilities are now recognized as a pivotal development trajectory for satellite networks [300]. These networks, characterized by high-velocity LEO satellites and frequent ISL changes, are inherently dynamic, particularly within multi-layer LEO topologies. This dynamic nature presents significant challenges for fully distributed routing at the network layer, necessitating innovative inter-layer link establishment and maintenance strategies [301–303].

To address these complexities, researchers have proposed a variety of adaptive routing strategies that emphasize topological flexibility and low-complexity modeling. For instance, the inter-layer link algorithms which designed for inclined orbital constellations choose to prioritize link selection based on physical properties such as distance and angular velocity [300]. Virtual node strategies, which divide satellite constellations into geographically bounded virtual regions, further mitigate dynamic network topology issues by assigning domain controllers within each region [301]. By introducing domain controllers, virtual node methods enable to reducing of the impact of topological fluctuations and facilitating manageable sub-network structures. In addition to virtual nodes, virtual topology techniques provide a structured approach by dividing network timeliness into discrete snapshots, thereby maintaining a manageable database of topological states for each time slot [304]. This approach reduces the demand on satellite storage resources while enabling efficient, snapshot-based routing [305]. A notable example of such virtual topology applications includes the hybrid use of virtual topology and virtual node methodologies, which, despite their efficiency, can still present computational challenges when inter-layer node counts are substantial [303]. An alternative approach involves multi-hop route stability estimation, where link interruption probabilities are calculated using stochastic geometry to ensure robust connections

through multi-hop routes [305]. Routing in multi-layer satellite networks also benefits from algorithmic approaches that integrate logical positioning within LEO constellations, allowing each satellite to communicate exclusively with neighbouring layers. This cross-layer design, leveraging geographic divisions, minimizes signaling overhead and maintains load-balancing [306, 307]. Back-pressure routing strategies, such as distance-based back-pressure routing (DBPR), assign link weights according to distance, thereby favouring non-congested, shorter paths to the destination [308]. These strategies have proven effective in fulfilling key performance criteria, such as high throughput and low latency, which are often difficult to achieve concurrently with traditional algorithms.

As DSIN grows in scale and complexity, future routing strategies must be tailored to meet the needs of satellite-integrated Internet applications. Strategies grounded in SDN principles offer a promising framework for these networks by leveraging global network awareness to optimize routing paths and reduce the computational cost associated with link selection in real-time [309]. For example, the fybrrLink QoS-aware routing algorithm, designed for GEO/LEO satellite networks, uses SDN's global perspective to expedite inter-layer link selection by combining modified Bresenha and Dijkstra algorithms, thus significantly reducing optimal route computation times [309]. Further, AI techniques are being explored to predict network conditions and optimize routing dynamically. Recent research has focused on leveraging ML, particularly reinforcement learning (RL), to address the challenges of routing in LEO satellite constellations. Q-learning, a type of RL, has been proposed for distributed routing in LEO satellite constellations [310]. This approach models the routing problem as a multi-agent partially observable Markov decision process (POMDP), where each satellite interacts only with its neighbours. The proposed Q-learning solution demonstrates comparable delays to centralized algorithms under steady-state conditions, increased supported traffic load without congestion, and minimal signaling overhead among satellites. Another significant contribution is the work on robust beam-to-satellite routing strategies for mega-constellations [311]. Aiming to minimize end-to-end latency and maximize the supported traffic load, strategies are proposed to address the challenges of routing in the presence of highly imbalanced traffic and dynamic network topology.

Moving forward, researchers anticipate the need for multi-constraint, integrated sensing-computation-routing mechanisms tailored for partitioned satellite networks. These mechanisms are expected to facilitate adaptive, rapid re-routing by calculating optimal paths based on ISL attributes for each hop, enabling network state awareness and further enhancing resilience and adaptability in multi-layer, heterogeneous DSIN.

3.9.3 *Congestion control*

The dynamic nature of DSIN introduces unique challenges in traffic management and congestion control. On the one hand, in DSIN, a large amount of signaling backhaul will result in severe local network congestion and aggravate network control load, which further worsens the control delays [312]. On the other hand, the unprecedented demands of real-time, high-volume data transmission in DSIN necessitate advanced multi-hop relaying and satellite-based processing to optimize data flow. Characterized by high-bandwidth and long-latency ISL, these networks resemble "long fat networks" (LFNs) and require carefully tailored flow control protocols to effectively manage both congestion and link idleness. Current end-to-end flow control algorithms typically operate through two primary phases: link state assessment and rate adjustment [313, 314].

During link state assessment, protocols monitor status metrics such as packet loss and delay, enabling dynamic evaluation of network conditions. This real-time monitoring allows for adaptive decision-making, optimizing network performance and ensuring efficient data transmission by adjusting to fluctuating network conditions. Packet loss detection algorithms, including TCP Reno and TCP Cubic, perform well in stable link environments with low packet loss but face challenges in LEO satellite networks, where frequent handovers introduce elevated packet loss [315]. Conversely, delay-sensitive algorithms like TCP Vegas and Fast TCP respond to increases in round-trip time (RTT) by preemptively reducing transmission rates, mitigating packet loss through early congestion avoidance [316, 317]. The above approaches aim to maximise bandwidth utilisation without causing congestion, keeping the link in a near-threshold state upon stabilisation. However, as arrival rates approach service rates, high bandwidth utilisation can compromise data timeliness, with the effects being particularly pronounced in latency-sensitive applications [317].

Recent advances, inspired by the AoI-sensitive ACP+ algorithm [318, 319], indicate that leveraging

age of information (AoI) metrics can provide timely congestion management across large-scale satellite constellations [320]. By adjusting transmission rates based on end-to-end AoI variation, this approach enhances the timeliness and reliability of flow control in LFNs. Rate adjustment methods often employ additive increase multiplicative decrease (AIMD) mechanisms, which converge sources to comparable rates over time, dynamically balancing throughput across sources. While multiplicative increase multiplicative decrease (MIMD) methods allow faster adjustment, they are generally less fair [320]. In fact, each adjustment mechanism impacts network stability and convergence. For instance, AIMD protocols like TCP Reno struggle to quickly reach target congestion windows in LFNs, limiting convergence speed [321]. TCP Hybla attempts to address this by comparing actual RTT with a predefined standard to fine-tune window increments, though slow convergence remains a constraint under LFN conditions [316]. By contrast, TCP Cubic introduces a nonlinear window growth function to expedite rate stabilization, although its aggressive scaling risks incurring congestion [315]. Future research should prioritise the development of flow control protocols that strike a balance between data freshness and efficiency while adapting to the frequent link handovers and sudden traffic surges characteristic of mega-constellations. Such advancements are crucial for ensuring robust, real-time performance in next-generation satellite networks, enabling them to meet the demands of emerging, latency-sensitive applications.

4 Collaborative cross-layer optimization

4.1 Mobility management

Mobility management can be categorized into location management and handover management, where the former focuses on tracking and updating the real-time location information of mobile terminals (MTs), while the latter ensures uninterrupted service continuity as an MT switches between two access nodes. We then elaborate on the overall architecture of mobility management and its specific schemes for the above two components in the following.

4.1.1 *Mobility management architecture*

Consider that the dual mobility arising from the high-speed movement of the CCS system, and the uneven distribution of mobile terminals, along with the highly overlapping satellite coverage areas that cause frequent handovers, can significantly increase handover delay, signaling overhead, and decision-making burden in DSIN [322]. These issues are intensified in the real world by the limited number and fixed locations of GSs equipped with mobility management functions (MMFs), as well as the unavoidable non-trivial backhaul delays across satellite-terrestrial nodes. In such cases, a large amount of mobility management signaling must be relayed through inter-satellite links with many hops to a few GS, leading to severe network congestion, increased management delays, and excessive signaling overhead [323]. Moreover, in the NTN architecture proposed by 3GPP with gNB functional split that allows for a tradeoff between costs and payloads, the two main network functions involving MMFs, i.e., the gNB-CU-CP and access and mobility management function (AMF), are still deployed on GSs, which makes their feeder links struggle to meet the delay and capacity requirements necessary for the mobility management of a large number of satellites [324]. Therefore, a mobility management architecture with dynamic MMF configuration is crucial for ensuring service continuity and timely management in DSIN.

The conventional mobility management architectures of LEO satellite systems are primarily divided into two parts: pure ground-based deployment (PGD) [325] and space-based management function deployment (SMFD) [326]. In the PGD architecture, which relies on GSs for mobility management, the number of GSs is significantly insufficient compared to the satellite networks because of the geographical and policy factors, which make it challenging to improve or even ensure the delay and overhead associated with mobility management. The SMFD architecture performs mobility management through satellite controllers equipped with MMFs. However, the satellites are required to continuously interact to gather information from across the network, which not only consumes data transmission resources but also creates a heavy control load. In addition, the number of controllers needs to be dynamically adjusted to the network scale.

To address these shortcomings, a novel architecture named space-ground integrated mobility management (SGIMM) was proposed to integrate the advantages of the above two approaches [312, 327]. It introduces non-LEO satellites as management nodes in the space segment, characterized by wider cover-

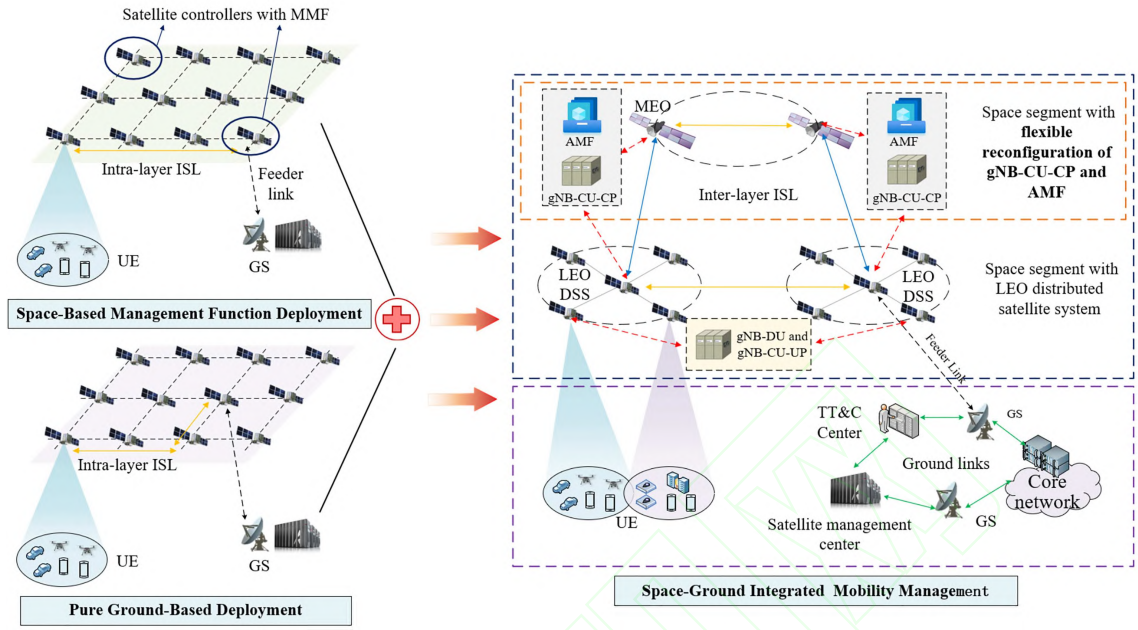

Figure 18 (Color online) Network view of SGIMM under the 3GPP NTN architecture with gNB functional split.

age, higher capacity, and fewer requirements for deployment quantity, and enables them to collaborate with GSs to perform mobility management, achieving a separation of transmission and management both physically and logically. On this basis, we can further apply the functions and interfaces defined by the 3GPP NTN architecture as shown in Figure 18, where the medium earth orbit (MEO) satellites and GSs are deployed with AMF and gNB-CU-CP for user location management and handover decision-making, respectively, while each LEO satellite is equipped with gNB-DU and gNB-CU-UP for accessing and signaling transmission [328]. This approach is equally applicable in DSIN, in which the management nodes of MTs can be further switched among the leader satellites in LEO parts of DSIN, the MEO satellites and GSs based on the tradeoff between delay and overhead.

Moreover, the SGIMM architecture will also cause lots of issues, such as the triple mobility due to MEO satellites, along with the additional delay and overhead caused by storage and content migration during the handovers among management nodes. In response, several studies suggest that distributed RL algorithms can be employed to address the relative mobility among MTs, GSs, and satellites, enabling ground MTs to intelligently select the optimal ground/satellite management node with minimal handover and migration delay [329]. Further, the authors in [330] proposed a flexible and distributed mobility management architecture (FDMMA) based on the SGIMM, including an overview of its network configuration, communication protocols, functional implementation, and handover procedures. The lightweight handover decision and distributed MMF configuration designed for the FDMMA offer improved handover delay and reduced signaling overhead.

Besides, further research is needed on the implementation of location management and handover strategies in our DSIN. On one hand, due to the time-varying topology and increased density of satellite systems, location management requires more frequent updates and paging for real-time tracking and recording, which imposes high demands on overhead and delay. On the other hand, the exponential growth of access requests, ultra-dense cellular coverage, and diverse types of services considerably complicate the handover triggering conditions and increase the selection of access nodes, straining the computational capabilities of MMFs and impeding the seamless handover [331].

4.1.2 *Handover and location management*

The feasible schemes for these two types of management techniques regarding the above challenges are discussed in this subsection.

(1) Distributed handover schemes. Handover schemes in mobility management fall into two main categories: beam handover and satellite handover, where the former is triggered when MTs move across boundaries between neighboring beams of a satellite, and the latter occurs as MTs move across the

coverage areas of adjacent satellites. Early research on handover schemes primarily focuses on beam handover since the number of satellites in conventional satellite systems is limited. Moreover, the channel allocation strategies [332] that ensure new channels are available during handovers and the handover guarantees [333] that prevent call drops or blockages throughout the handover process are the prime issues in managing beam handover requests. With the growing density of satellite networks, the frequency of handovers between satellites has increased considerably, prompting a shift in recent research toward the design of satellite handover schemes.

Graph theory is a typical approach to implement satellite handover, where each satellite's coverage period is treated as a node, and the handovers between satellites are represented as directed edges [334, 335]. In this framework, the handover process can be described as selecting an optimal path from all available handover paths. Further taking into account the MT request distribution, available channel state, satellite coverage time and other QoS requirements, the constructed graph can be utilized to find the optimal handover in diverse communication scenarios [336]. However, these handover schemes rely on centralized decision-making based on global state information, leading to excessive complexity and overhead in DSIN with numerous satellites and MTs.

Distributed RL, such as multi-agent Q-learning, is viewed as a potential solution to the above issues. By deploying well-trained agents at various nodes of DSIN, each satellite can derive localized strategies that simultaneously satisfy payload constraints and minimize handover costs with low computational complexity [337, 338]. Nevertheless, due to the limited state-action pairs stored in the Q-table, most RL-based handover schemes are only suitable for simple scenarios with a limited number of satellites. In more complex DSIN, it is necessary to employ more advanced DRL algorithms to design effective handover schemes. A MT-driven deep Q-network (DQN) based handover scheme was developed in [339], in which a centralized agent deployed at the management satellite node for training the DQN, and each MT individually makes its handover decisions based on the parameter disseminated from the trained node. Adopting a similar centralized-training and distributed execution approach, the authors in [340] proposed a multi-agent fingerprints-enhanced double deep Q-network to address the handover issues under burst traffic scenarios with delay constraint. Since the above handover schemes are all designed under the static propagation conditions, the authors in [341] developed a multi-agent successive hysteretic DQN algorithm to address the handover problem that involves the throughput requirement of MTs and load-balancing demand of satellites in a time-varying satellite-terrestrial channel.

(2) Dynamic location management. Location management (LM) primarily involves the location area (LA) (i.e., tracking area) design, location update and paging, where the size of LA is crucial, as it affects the frequency of signaling interactions during updates and paging. Specifically, a larger LA may result in delayed updates of MT location information, while a smaller one may lead to excessive updates and paging overhead [342, 343]. As a result, a dynamic adaptive LA scheme is proposed in [327] to achieve an optimal tradeoff between LA division and updates/paging overhead. The scheme enlarges the LA size for high-speed MTs to reduce overall LM overhead and shrinks the LA size in high call-traffic areas for faster user access while keeping the number of paged satellites manageable in large-scale satellite networks.

In an IP-based network, the Mobile IP version 6 (MIPv6) [344] is a classical management scheme introduced by the Internet engineering task force (IETF), where an MT only binds to a new IP address upon handover to maintain its TCP connection with an access point or GS and minimize the impact of MT location changes. As communication networks become more complex, the IETF introduces enhancements such as Proxy MIPv6 (PMIPv6), Fast Handover for MIPv6 (FMIPv6), and Hierarchical MIPv6 (HMIPv6) to strengthen the location management performance of IP-based network [345]. Moreover, since these centralized LM schemes face high update overhead and limited scalability due to frequent handovers with increasing MTs and satellites, the IETF is working on developing LM solutions for future satellite networks.

Seamless IP diversity-based generalized mobility architecture (SIGMA) is a promising LM solution for DSIN, where an MT can continue using its previous IP address while obtaining a new one, with the timing for switching to the new IP address and deleting the old one predicted based on the deterministic movement path of satellites [346]. However, SIGMA does not account for the significant signaling overhead induced by frequent satellite handovers under high mobility conditions. To mitigate the link management (LM) cost, two dynamic LM approaches have been proposed. The first approach is a dual-location area link management (LM) scheme [347], where a satellite location area (SLA) is introduced for the positioning of LEO satellites. By combining the SLA with the user location area (ULA), this scheme adapts to the varying mobility patterns of LEO satellites and mobile terminals (MTs), effectively reducing

the frequency of global updates. The second approach is a virtual attachment point (VAP)-based link management (LM) scheme [348], in which mobile terminals (MTs) first establish a connection to a fixed logical point, the VAP. From there, MTs access the satellite network via this designated point. In such a case, the AMF only tracks the MT's location relative to the VAP, along with the VAP's location relative to the satellite, which can mask the mobility of satellites from MTs.

Further, to enhance the scalability of the existing LM schemes, several researchers propose distributed solutions that offer advantages like near-optimal path selection, reduced workload distribution, and improved handover performance [349], which can be viewed as a potential approach for DSIN. Virtual MIPv6 (VMIPv6) is a typical distributed LM scheme [350], where a virtual agent cluster (VAC) comprising a group of LEO satellites within a specific LA collaboratively manages the mobility of MTs within its virtual agent domain (VAD). When MTs perform handovers within the same VAD, they only change their local addresses and retain their global addresses, which helps reduce LM overhead and delay.

4.2 Resource management

The proliferation of SatCom, particularly with the advent of satellite-integrated Internet and the anticipated 6G services, underscores the criticality of resource management and allocation mechanisms in DSCN. For instance, satellite constellations operating in LEO provide reduced latency and improved bandwidth for applications such as real-time telemedicine and autonomous vehicle coordination. However, the dynamic dual mobility nature of satellites and users, especially in LEO and medium earth orbit (MEO) based satellite networks, requires advanced resource allocation strategies to address both scalability and flexibility in resource distribution across hundreds or thousands of satellites [227]. Moreover, the aggregated resource management for satellite edge computing highlights the need for efficient resource allocation to support the computational and storage demands at the network edge [351]. This need underscores the critical role of resource management mechanisms in meeting the complex requirements of DSCN. In general, the DSCN faces the following core challenges in resource management.

(1) Scarcity and uneven distribution of resources. On the one hand, traditional resource management solutions do not fully exploit the regenerative capabilities of next-generation satellite systems, which offer flexibility in bandwidth, power control, and onboard data processing. To address this, research has ventured into areas such as dynamic bandwidth allocation, edge caching optimization, and onboard signal predistortion for digital transparent satellites [10]. On the other hand, exacerbated by the rapid growth in the number of satellites and the diverse nature of satellite missions, the scarcity of available frequency spectrum becomes a critical bottleneck [352]. Techniques such as cognitive radio (CR) have been proposed to enable dynamic spectrum sharing among satellite networks, minimizing interference and maximizing spectrum efficiency. However, implementing CR in the CCS system poses unique challenges due to the need for real-time interference management across multiple satellites. It is important to note that with the increasing density of satellite constellations or clusters, managing interference, both intra-system and inter-system, becomes a significant challenge. Hence, intelligent resource allocation algorithms that can predict and mitigate interference are essential for improving spectrum efficiency.

(2) Hysteresis effect. The DSCN exhibits a hysteresis effect, where resource scheduling lags behind the update of resource status information, resulting in low resource utilization efficiency. To tackle this, researchers have introduced resource scheduling algorithms based on DRL, which can adapt to complex and dynamic environments, addressing the mismatch problem of traditional scheduling methods. The DRL algorithm is widely used to address the joint allocation of sub-channels and power in multi-beam SatCom systems [353], offloading between terrestrial and satellite networks [354], demonstrating the innovative significance of DRL in heterogeneous terrestrial-satellite communication networks. Additionally, in the field of distributed resource management in DSIN, MADRL has been increasingly utilized to address the complex challenges of coordinating and optimizing resources across multiple agents (satellites). By employing MADRL, researchers have been able to develop innovative solutions that address the challenges of dynamic heterogeneous resource allocation, task offloading, and service allocation in satellite networks. The integration of MADRL with other techniques and frameworks has shown promise in enhancing the performance and efficiency of satellite communication systems, and it is expected to play a crucial role in DSIN. For instance, a hierarchical cross-domain satellite resource management framework is analyzed by employing MADRL to guide the collaborative work of multiple satellites within a domain [355], enhancing the scheduling capabilities of multi-domain satellite systems.

(3) Edge computing and network slicing. Edge computing has been introduced to reduce latency by bringing processing closer to the data source. Distributed caching can alleviate bandwidth demands by storing data at or near satellite locations. The utilisation of MADRL in optimizing resource allocation in complex satellite-terrestrial networks allows for enhancing the efficiency of edge computing in complex satellite-terrestrial networks [356]. Moreover, network slicing strategies for real-time applications in large-scale satellite networks are crucial for managing network resources efficiently in satellite networks [357]. Further, the use of digital twins to enhance resource slicing in LEO satellite networks is explored, emphasizing the importance of robust and adaptive resource management strategies for the dynamic nature of SatCom [358]. However, these technologies are still in their infancy within satellite networks, and their deployment across globally dispersed satellite constellations remains a challenging endeavor. Issues such as data synchronization, cache coherence, and global scalability hinder the full-scale adoption of these methods in real-world applications.

Looking ahead, the future DSIN will require innovative approaches to enhance resource management efficiency. One promising direction is the development of integrated resource management for terrestrial-satellite systems, which can optimize the use of resources across both domains. Additionally, research into intelligent resource management for satellite and terrestrial spectrum shared networking is essential [359], focusing on spectrum sensing, prediction, and allocation to improve spectrum efficiency. Moreover, in the distributed satellite network architecture, blockchain-based frameworks for decentralized resource management could enable autonomous resource allocation among the CCS system, enhancing flexibility and reducing the need for centralized control. By providing a secure, distributed ledger, blockchain can facilitate transparent and verifiable transactions, improving the reliability and scalability of resource sharing. Further, future DSIN would benefit from the development of adaptive algorithms capable of real-time resource allocation adjustments to handle dynamic demand fluctuations. These algorithms should be designed to operate in high-mobility environments, allowing networks to dynamically adjust to the changing conditions of satellite constellations without compromising efficiency. The last but not the least direction is multi-orbit satellite collaboration, where collaborative resource management strategies among satellites in different orbits, such as LEO, MEO, and GEO, can be obtained. This significantly enhances resource utilization by enabling satellites to share resources based on their specific capabilities and operational constraints, resulting in improved network resilience and performance.

4.3 Secure communications

With the rapid development of the communication industry, the exponential increase in confidential and sensitive data over wireless links yields increasingly prominent security concerns. Compared with terrestrial communication networks, the electromagnetic broadcast characteristics of satellite networks enable longer transmission distances and broader wireless coverage, making these links more susceptible to eavesdropping, jamming and unauthorized access. Moreover, the three-dimensional wide-area coverage of satellite networks allows malicious nodes to conduct illegal activities from any position in space, air or ground, and it is challenging to detect and cope with malicious nodes in the passive listening mode. Additionally, Resolution No. 35 of the 2019 World Radiocommunication Conference requires satellite operators to submit actual deployment parameters of their satellite systems, which, however, facilitates malicious nodes to predict satellite orbits and positions and prepare for potential attacks or eavesdropping. Consequently, DSIN faces severe challenges in ensuring secure information transmission. To ensure the security of information transmission, three mainstream solutions are classic cryptography, physical-layer security, and quantum-domain security, which constitute multi-level security mechanisms and provide comprehensive security at the physical layer, network layer, transport layer, and even application layer, as shown in Figure 19.

Cryptography uses public/private keys to encrypt/decrypt confidential information, rendering it undecipherable without the key [360]. The security essence of key-based cryptosystems lies in that it is difficult for devices with limited computational power to carry out mathematical operations to decipher the transmitted data. The cryptography-based security solutions operate at the network or upper layers and can be easily transplanted to satellite communication systems [361], where the limited on-board storage and processing capabilities need to be prudently considered. With the growth in device computational power and advancements in quantum computing technology, the capability to decipher information will be significantly improved. A possible way for the security enhancement is to increase the key length, which can temporarily alleviate the pressure of cracking brought by the enhancement of classical computation and

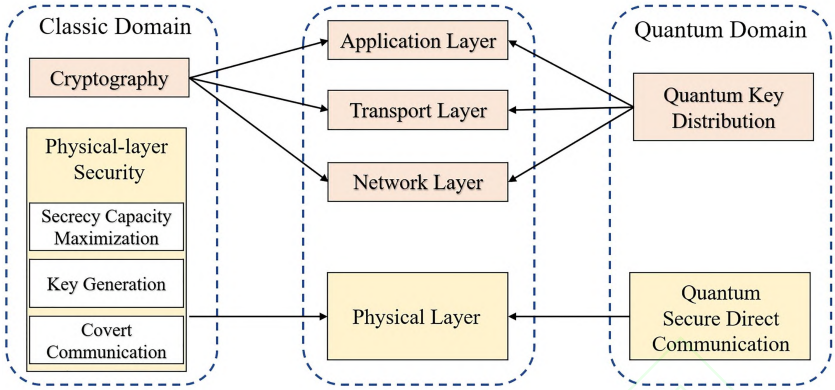

Figure 19 (Color online) Different information security solutions at different open system interconnection layers.

cryptanalysis capabilities. However, the longer the key, the lower the efficiency of the encryption algorithm becomes, resulting in slower encryption, decryption and key distribution speeds. This motivates us to apply more lightweight and efficient algorithms to traditional cryptographic methods and also reveals that cryptography alone is insufficient to guarantee secure information transmission in DSIN.

To cope with this issue, it is promising to move the mechanisms for secure information transmission down to the physical layer directly through physical signal transformation. Based on the concept of information theory security, physical-layer secure communication technologies exploit the differences between legitimate and eavesdropping channels to achieve complete secrecy at the physical signal transmission level, effectively preventing the third parties from intercepting information through eavesdropping channels [362]. Existing methods for secure transmission at the physical layer are mainly divided into three categories. The first is to increase the rate disparity between legitimate and eavesdropping channels by using techniques such as information coding, beamforming, artificial noise and relay cooperative jamming, thereby enhancing the secrecy capacity [363]. The second is to generate inherently random key pairs for physical-layer key encryption of the transmitted data by utilizing the reciprocity, randomness and spatial uniqueness of wireless channels [364]. In this regard, the uniqueness of channel information ensures that the transmitted data cannot be deciphered even if malicious nodes enhance their computational capabilities. Although these methods protect the information contents from being deciphered, malicious nodes can still detect communication behavior, locate the sources of radiation and launch physical attacks, thereby posing a security threat to legitimate nodes. To achieve a higher level of secure physical-layer transmission, the third is the emerging covert communication technology. It typically hides the behavior of information transmission by mimicking environmental noise or other natural signals. Therefore, the transmitted signal appears to be part of the electromagnetic environment and the information is transmitted without drawing the attention of unauthorised entities [365]. The key to information covertness lies in the various uncertainties in the end-to-end physical transmission, such as the signalling uncertainty, noise uncertainty, fading uncertainty and interference uncertainty. In [366], a covert satellite communication scheme over an overt channel by randomizing the variance feature of the transmitted Gaussian signalling to realise a positive covert rate, i.e., increasing the uncertainty at the transmitted signal. The interference uncertainty inherent in large-scale LEO satellite networks is utilised to enhance the covertness in [367], where a two-stage Stackelberg game is used to model the conflict dynamics between the adversarial BS and the satellite network. A Stackelberg equilibrium is achieved to reveal the trade-off between transmission reliability and covertness in leveraging co-channel interference. Since physical-layer secure communications depend on how to transform the physical signal, the information security can be further enhanced via effectively combining physical layer transmission with advanced channel coding, massive multiple input multiple output (mMIMO), and terahertz technologies.

Although cryptography-based and physical-layer solutions have the capability to ensure communication security, there is always a certain possibility of information leakage. Instead, due to the fundamental principles of quantum mechanics, quantum communications provide a promising approach to realize unconditionally secure communications even when malicious nodes have unlimited computational power [368]. Specifically, the no-cloning theorem states that an unknown quantum state cannot be perfectly copied and thus prevents the eavesdroppers from copying the transmitted quantum states. Quantum superposition means that information is encoded in quantum states that can be in a combination of '0'

and '1' simultaneously, and the detection of these states collapses the superposition, making it impossible to extract the full information without knowing the correct measurement basis. Furthermore, quantum entanglement shows that any detection on one entangled particle instantaneously affects the state of the other, thereby making the entanglement suitable for secure communications. To this end, numerous quantum cryptographic or quantum communication protocols have been proposed, which can be classified into four main branches: quantum key distribution (QKD), quantum teleportation, quantum secret sharing and quantum secure direct communication (QSDC). Different from conventional cryptography, QKD utilizes the laws of quantum physics to distribute unconditional secret keys between a pair of legitimate parties rather than the classic wireless exchange, and the messages are encrypted by the agreed secret keys and then transmitted over a classic wireless channel. The QKD systems have been extended to ranges of hundreds of kilometres through optical fibers [369] and thousands of kilometres in satellite systems of the Micius satellite for quantum science experiments [370]. QSDC encodes the secret message directly into quantum states and then allows the direct transmission of secret messages through a quantum channel without establishing a shared secret key first. It should be noted that QKD systems have become commercially available, while QSDC is still an active area of research and development, and its practical implementation is still in its early stages.

To sum up, the three mainstream security solutions constitute the built-in fundamental security capabilities in the future DSIN. The requirements for embedded security capabilities in DSIN can be addressed by integrating these fundamental security capabilities in 6G and making the corresponding adaptive improvements. Hence, the functions of the collection, control, and isolation are provided. On this basis, DSIN can also leverage distributed FL technology to realize decentralized secure and trustworthy mechanisms, constructing a secure and reliable intelligent distributed satellite network. Consequently, the constructed network satisfies the differentiated security requirements of various service scenarios, enhances the autonomous security capabilities of network communications, and establishes a measurable and evolvable intrinsic security protection system.

4.4 Testbeds of DSIN

One of the key challenges faced by DSIN is ensuring reliable and efficient communication between satellites and ground stations, as well as inter-satellite communication. As satellite systems become more complex, the testbeds of DSIN are a critical infrastructure that facilitates the development and validation of advanced space communication technologies. As the space industry continues to evolve, several platforms have been developed to incorporate orbital simulations, inter-satellite communication, ground-station connectivity, and dynamic topology changes for SatCom systems. These platforms are essential for assessing the performance of communication protocols under varying environmental conditions. Historically, a variety of satellite network simulators in the market, such as iTrinegy's network emulators and the DataSoft satellite network simulator, or in OpenSource projects like the satellite network simulator 3 (SNS3), OPSAND and real-time satellite network emulator, have been widely used for network simulations [227]. However, the advent of DSIN requires testbeds that can handle the complexity of multi-satellite networking and their associated communication protocols. The development of these testbeds has evolved from simple, single-satellite simulations to sophisticated systems that can simulate entire satellite constellations.

The international application status of these platforms is marked by several key developments and trends. For instance, the integration of SDN in satellite networks has been a significant step forward. The SDN allows for the separation of the data plane and control plane, enabling centralized management of network resources and dynamic reconfiguration in response to varying traffic demands and network conditions [371]. This approach not only reduces the computational burden on individual satellites but also optimizes the use of network resources, catering to the high-dynamic and heterogeneous nature of satellite networks. As a cost-effective SatCom solution for 5G, the virtualisation of SatCom network functions is validated in the SAT5g project [372], ensuring compatibility with the 5G SDN and NFV architecture. To address the challenge of analyzing and evaluating integrated space, aerial, and ground networks, a SAGIN simulation platform, which integrates multiple network protocols, node mobility, and control algorithms, is developed in [373], optimizing the network functions such as access control and resource orchestration in a combined centralized and decentralized manner. While the flexibility and scalability brought by SDN/NFV technologies are beneficial for heterogeneous network architecture, the issues of deploying the SDN controllers in DSIN and coordinating the actions among the controllers require

careful investigation. As a viable option, cloud/edge computing and network slicing can be incorporated into resource isolation, cloud services, computation offloading, and edge caching, fundamentally meeting the varying requirements of ever-increasing services and applications. In addition, the development of testbeds like the DVB-RCS2/S2 has been instrumental in the design and validation of next-generation satellite systems [374]. These testbeds provide a distributed environment that mimics the operational characteristics of real satellite networks, allowing researchers and engineers to experiment with various protocols, routing algorithms, and network configurations in a controlled setting.

A critical component of testbeds for DSIN is the constellation network simulation. The constellation network simulation includes orbital simulation, topology modeling, and link parameter analysis [375]. Orbital simulation is essential for analyzing satellite link performance and designing communication topologies. By predicting satellite positions and simulating their operational behavior, these platforms enable accurate modeling of communication links and the dynamic nature of DSIN. Using high-precision orbital models such as the Keplerian and SGP4 models is important for the DSCNs to ensure accurate simulations. The Keplerian model is used for scenarios with low precision requirements [376], while the SGP4 model is a more advanced numerical integration model that accounts for factors such as Earth's nonspherical shape and the gravitational influence of the Sun and Moon [377]. The use of these models allows researchers to simulate various scenarios, including LEO constellations and MEO satellites. Furthermore, network topology construction based on simulation tools like NS3 allows the creation of different types of constellations, such as polar, Walker, and hybrid constellations, which are then visualized in 3D. This provides a robust framework for evaluating how satellite movement affects communication links and network performance. Specifically, the platform can simulate changes in the routing topology over time and evaluate metrics such as packet delay, packet loss, and bandwidth utilization.

Distributed computing resources for managing complex simulations are necessary for DSIN. Modern testbeds are envisioned to employ a distributed architecture where the simulation and computation modules are separated. This architecture allows for the efficient allocation of resources across distributed nodes, enabling to simulate large-scale satellite constellations. The platform utilizes a central control center that manages task allocation and monitors resource usage, ensuring efficient execution of simulation tasks. By leveraging distributed computing resources, the platform can concurrently calculate satellite parameters such as orbit, attitude, link budgets, and topology changes, resulting in higher computational efficiency. Moreover, simulation modules must support multiple communication protocols, including satellite-to-ground, inter-satellite, and intra-satellite communication protocols. These modules access real-time computation results from the distributed nodes and display them to the users, enabling the real-time monitoring of network performance. The inclusion of collaborative satellite scenarios further enhances the platform's utility by supporting tasks such as cooperative satellite computation, data transmission coordination, and interference management. The simulation module is designed to cover four key multi-satellite collaboration scenarios: computational collaboration, transmission collaboration, interference coordination, and data injection coordination. For instance, when individual satellites face computational limitations or high latency, a multi-satellite computational collaboration strategy is used. This strategy leverages a grid model to track changes in the satellite network topology, selecting a central node as the computational coordinator. This node processes collaboration requests and returns a list of suitable satellites based on criteria such as maximum computational power and shortest transmission distance, enabling efficient inter-satellite collaboration.

Additionally, the testbed supports the configuration of distributed computational collaboration simulations, providing performance evaluations for these scenarios. For handling large volumes of remote sensing data, the system employs multi-link data transmission strategies, where data are split and transmitted in parallel through multiple channels, significantly reducing latency. The testbed also includes distributed multi-path data transmission simulations, supporting large data downlink tasks and enabling efficient evaluations of remote sensing images delivery. Dynamic beamforming capabilities are used to allocate and manage frequency spectrum resources between satellites, optimizing the use of limited spectrum and avoiding interference. The interference management can achieve large-scale network construction and flexible spectrum allocation. A testbed for MU-MIMO precoding in multi-beam satellite systems is designed by [378] to demonstrate the feasibility of spatial multiplexing with full frequency reuse for video transmission across two co-located GEO satellites. To handle large-scale computational model data injection and synchronization, a combination of SDN-enabled centralized and distributed routing mechanisms can be used to ensure rapid synchronization of ground network data. The platform supports the configuration of injection tasks, enabling collaborative simulation and performance evaluations of data

injection across multiple satellites. In the beyond 5G NTN-terrestrial networks, the SDN-based testbed has been proposed to monitor the substrate network with traffic statistics and apply routing decisions, which allows for managing the procedure proactively to minimize traffic losses [379].

The need for integrated ground and on-orbit testing of various communication protocols should also be highlighted in the testbed of DSIN. Ground testing systems are designed to emulate the satellite environment by software and utilize four key components: space simulation systems, ground simulation systems, testing systems, and auxiliary equipment. This allows for functional verification of network devices and interoperability testing, ensuring that the technology meets specified standards. On-orbit testing, on the other hand, involves deploying the satellite in a real-world environment, performing computing or inference tasks [380], embedding a virtual network [381], and evaluating its performance under actual conditions. This process verifies whether the communication protocols, technical standards, and system performance align with expectations. Furthermore, data obtained from on-orbit tests are used to calibrate the ground-based testing systems, enabling iterative refinement and optimization of DSIN.

5 Emerging directions

In this section, we focus on the evolving trends in DSIN and the associated technical challenges, examining key aspects such as network resource virtualization, distributed AI-enabled on-orbit information processing, semantic communications, direct satellite-to-device communications, and goal-oriented integration of sensing, communication, and computation. These elements are crucial for advancing DSIN to support a wide range of applications.

5.1 Network resource virtualization in DSIN

Resource constraints such as limited computational power, bandwidth, and storage present significant hurdles in DSIN. Unlike traditional terrestrial networks, satellite networks operate in constrained environments where heterogeneous resources must be managed with high precision [3]. Moreover, the fragmented and unbalanced resources across multiple satellites significantly hinder resource utilization in DSIN, particularly when accommodating diverse intelligent applications or tasks. For instance, urgent tasks, such as disaster monitoring, demand higher prioritization and faster resource deployment compared to routine data collection missions. However, the current satellite network lacks the adaptability to prioritize and scale resources based on these differences, leading to either under-allocation or over-allocation of resources and ultimately reducing energy efficiency. As a promising technology to deal with the issue of resource management in DSIN, the network resource virtualization (NRV) abstracts space-air-ground physical resources to create virtualized resource entities and corresponding heterogeneous resource set for diverse services [382], enabling flexible and efficient resource orchestration, as shown in Figure 20.

In NRV-enabled DSIN, heterogeneous resources can be dynamically scheduled and prioritized to address the areas of greatest need, thereby optimizing key performance indicators across multiple satellites. Specifically, based on the distributed satellite computing network architecture presented in Subsection 2.2, the DSIN can decentralize the computation, storage, and networking resources of satellites across different spatial environments, and connects them via ISL to form a collaborative working framework. This architecture allows for the flexible allocation of satellite, ground, and edge network resources, enabling efficient information collection, processing, computation, storage, and distribution. It is particularly well-suited to meet the massive and intelligent information service demands of the future, which will be driven by the vision of the Internet of Everything.

Despite the potential of NRV, several critical challenges hinder its full integration into DSIN, the foremost being the lag effect in resource allocation caused by long-distance communication. The physical distances between satellites introduce latency issues, potentially undermining the effectiveness of NRV, particularly in time-sensitive applications. Research efforts have explored multi-layer architectures and edge computing solutions to reduce latency [383]. By enabling data processing closer to the source, edge computing helps mitigate latency. In satellite networks, LEO satellites can function as edge devices, processing data locally before transmission. However, while edge computing alleviates some latency issues, it only provides a partial solution and demands substantial onboard processing capabilities, posing a challenge for smaller satellites with limited resources. Moreover, achieving resource virtualization

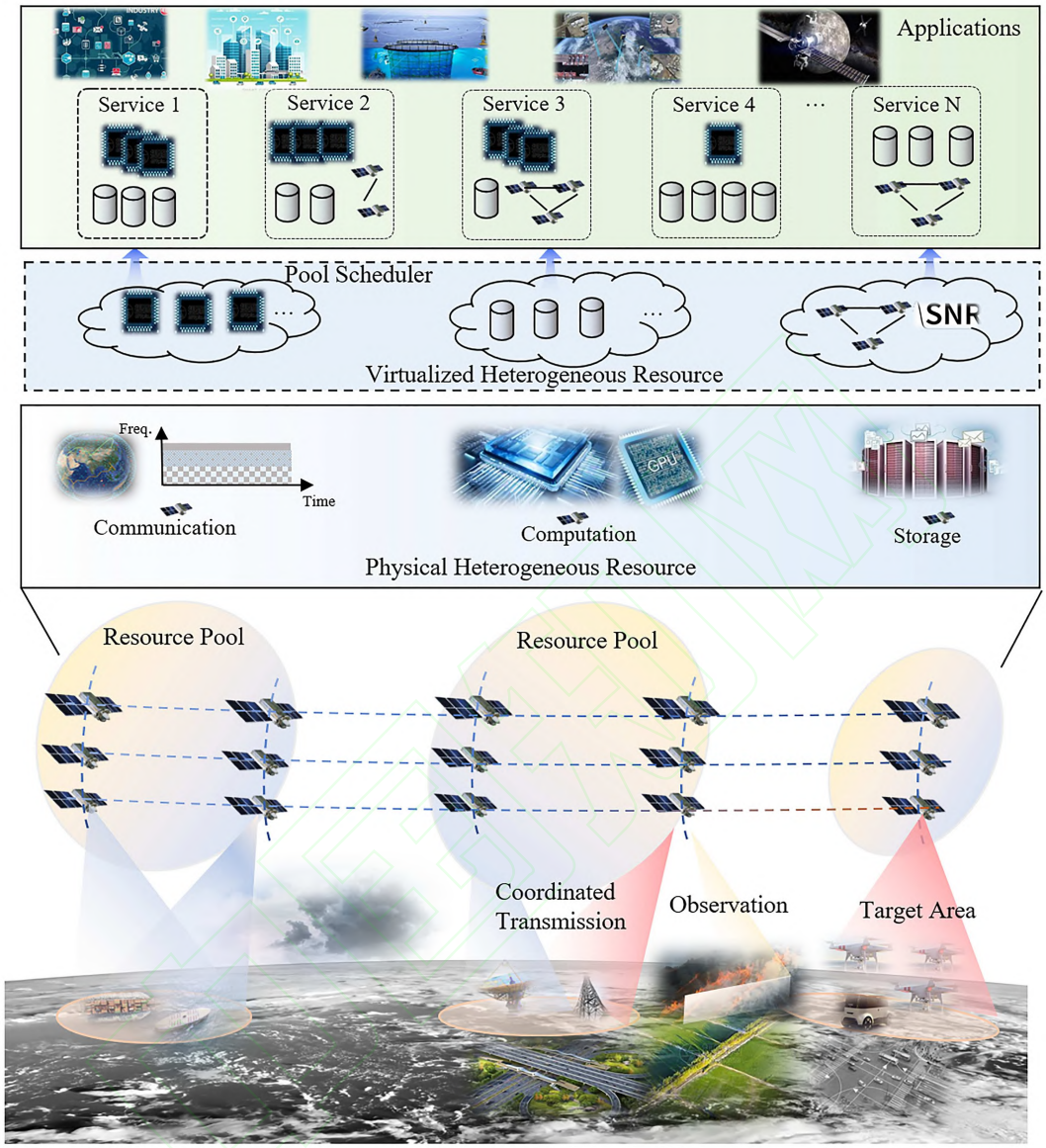

Figure 20 (Color online) Illustration of resource pooling as a foundation for enhanced NRV.

across heterogeneous satellite platforms presents challenges in forming a unified, interoperable resource scheduler.

As a foundation technology for enhanced NRV, the resource pooling concept is proposed by [384] to achieve uniform resource management. The resource pooling technology aggregates heterogeneous satellite network resources, i.e., processing power, bandwidth, and storage, into a centralized, shared pool accessible by any satellite. The multi-dimensional optimization problem is typically NP-hard, but recent advances in meta-heuristic methods, such as genetic algorithms (GA), ant colony optimization (ACO), and simulated annealing (SA), have made it possible to solve many of these problems [385]. Benefiting from this advance, we can utilize the NRV technology to meet the unique requirements of different tasks, allowing for task-oriented resource prioritization. For example, a real-time disaster monitoring mission could be allocated a higher share of computational and bandwidth resources than routine environmental observation tasks, thus ensuring faster data processing and transmission for time-sensitive applications. To achieve resource pooling, physical pooling must be implemented to eliminate barriers between homogeneous and heterogeneous satellites, enabling a shared communication channel that facilitates resource sharing across users, instances, and time scales. In addition, logical pooling can be utilized to aggregate fragmented resources into a cohesive virtual pool in a software-defined manner. Specifically, this approach facilitates fine-grained and flexible resource management, enabling resources to be dynamically

partitioned and allocated based on task requirements and swiftly released and reclaimed upon task completion. Such rapid allocation and deallocation mechanisms significantly improve resource flexibility and availability. It is worth noting that the edge computing combined resource pool architecture has shown promise in optimizing resource sharing for satellite-ground cooperative tasks [386]. However, a significant challenge in establishing a resource pool of clustered satellites lies in the dynamic topology changes caused by orbital movements, which result in instability. The core issue revolves around managing the capacity of the resource pool and dynamically selecting satellites in the CCS system to ensure efficient resource sharing [387].

Another critical issue lies in the limitations of multidimensional heterogeneous resources in DSIN, which are confined to different network layers, lacking interoperability and unified global coordination. To address this, the virtualized hybrid satellite-terrestrial systems (VITAL) project introduces the network functions virtualization (NFV) and SDN technologies into satellite networks, enabling flexible and resilient resource management in satellite-ground networks [388]. The SDN facilitates dynamic resource management in heterogeneous satellite networks. By dynamically deploying SDN controllers in LEO satellite constellations, it becomes possible to effectively address network traffic fluctuations caused by variations in user geographic locations and time zones [389]. Moreover, the SDN/NFV technologies in SAGIN have been proven to effectively transform the network from a connection-based model to a service-based model, offering on-demand resource allocation and service customisation [390]. However, the inherent resource dynamics and service uncertainties in DSIN render traditional scheduling methods unsuitable for decision-making tasks requiring high efficiency and rapid response, such as service function chain deployment and mapping. Reinforcement learning (RL), as a self-learning and adaptive decision-making approach, has been explored in SDN/NFV networks to some extent [391]. Nevertheless, how to apply RL to integrate lightweight SDN/NFV for enabling adaptive and scalable NRV in DSIN remains relatively scarce, highlighting a gap in the current research.

5.2 Distributed AI inference and on-orbit information processing

Given the constrained transmission capacity of satellite-to-ground links, traditional architectures reliant on GS struggle to address the growing demands for massive data return [392, 393]. Moreover, the limited resources of a single satellite render it impractical to process all received data independently. Consequently, leveraging on-orbit computing and storage resources across DSIN becomes essential for the real-time processing of vast sensor data collections. As outlined in Subsection 2.2, the distributed satellite computing network architecture integrates the available computing, communication, sensing, and storage resources of distributed satellites into a virtual resource pool for unified management. Onboard intelligent fusion of multi-source satellites is achieved by managing multi-dimensional and heterogeneous resources, enabling collaborative distributed computing power. Distributed AI technologies, such as federated learning [394], transfer learning and split learning [395], can be effectively employed within the CCS system in DSIN to facilitate on-orbit information processing. Moreover, with the rapid development of AI, large language models (LLMs) have become a star technology in the field of natural language processing. However, the inference process of LLMs often faces significant challenges in terms of computational resources and time costs, especially for on-orbit intelligent applications in satellite networks. To address this issue, distributed AI inference technology has emerged. Distributed AI inference involves breaking down the large language model inference task, which would typically be executed on a single machine, into multiple sub-tasks that can be processed in parallel across multiple nodes with limited computational power. This approach enables the efficient utilisation of computational resources from distributed and edge nodes, significantly improving on-orbit information processing and inference speed, while also reducing inference costs. During the distributed inference offloading process, managing a large number of computational resources, including CPUs, GPUs, and others, is crucial. To accelerate the model inference speed, NRV and distributed satellite computing network architecture, as mentioned in Subsections 2.2 and 5.1, can be employed to achieve compact and efficient computational power collaboration and task priority scheduling. Furthermore, in response to the dynamic network characteristics of DSIN, such as frequent topology changes and narrow communication time windows between the satellites and GSs, distributed incremental inference emerges as a promising technique. By learning the implicit relationships between various tasks or model information, it abstracts a distributed inference model with basic inference capabilities. This approach avoids full-scale learning and model updates, thereby accelerating model inference speed while reducing communication overhead.

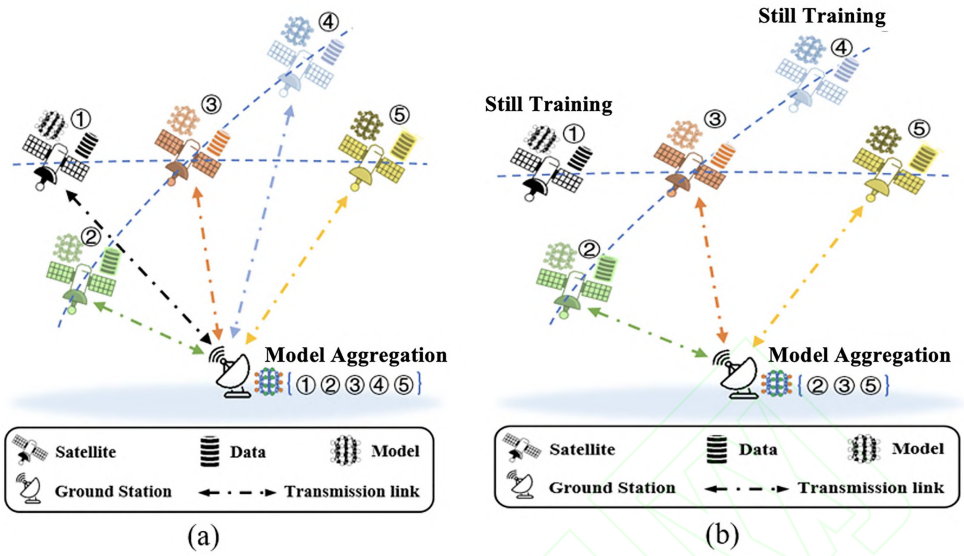

Figure 21 (Color online) Illustration of (a) synchronous and (b) asynchronous architectures of distributed AI for on-orbit information processing.

The research on distributed AI in DSIN involves two architectures: synchronous architecture and asynchronous architecture. Synchronous architecture requires all satellites to update simultaneously to ensure state consistency, making them suitable for tasks that require real-time coordination. However, they may suffer from reduced efficiency due to factors such as network latency. In contrast, the asynchronous architecture allows satellites to update independently, providing flexibility and effectively accommodating delays and node failures, but the inconsistency in update pacing may increase coordination complexity.

In a synchronous architecture, each satellite is equipped with edge computing capabilities and acts as a typical client group of traditional on-orbit distributed architectures. Each satellite can train on its locally relevant data and, after several rounds of local training, obtain a high-performance local model. This model is then sent to the GS for global model aggregation at the same time, as illustrated in Figure 21(a). Similarly to the naive idea of FedAvg [394], after multiple global rounds of model training and aggregation, the CCS system ultimately generated a unified model suitable for global inference. FedProx [396] addresses each satellite's significant differences in data distribution by introducing a loss regularization term that uses the current local model and the previous global aggregated model, ensuring that any onboard model does not deviate too far from the global model parameters. FedNova [397] treats the adjustment of learning rates and weight decay parameters as a joint optimization problem. In each global round, after the server receives all local gradients, it calculates the global gradient and sends it back to each satellite, allowing them to adaptively adjust their learning rates and weight decay parameters to optimize local model performance. To reduce the risks of single-point failure and communication congestion, the CCS system can perform local model updates through communication between co-orbital satellites and cross-orbital exchanges, enabling satellites to collaboratively train models without a central server [398].

The advantages of deploying distributed AI in synchronous architecture lie in the simplicity and intuitiveness of the algorithms. This makes them easy to implement and manage while ensuring that all satellites are in the same state during the training process. However, a significant drawback is the low global efficiency, as faster nodes must wait for slower ones, leading to resource wastage in the CCS system. Additionally, delays or failures from any single node can impact overall performance. As the number of satellites in the CCS system increases, the waiting time for model aggregation will also grow. Therefore, synchronous architecture has poor adaptability to constellations with varying onboard resources.

In an asynchronous architecture, as shown in Figure 21(b), the clustered satellites in the CCS system cannot update or aggregate at the same time. Therefore, there is a greater need to focus on the impact of single satellite model lag on global aggregation. Previous research has extensively explored optimizing local training and global model aggregation. By proposing an improved asynchronous federated learning method to enhance robustness in heterogeneous environments, this approach focuses on utilizing the predictable availability of satellites and introduces a new communication protocol and algorithm frame-

work to improve the efficiency of the training process [399]. Considering the deployment of distributed AI orchestrated by an external constellation parameter server, a novel communication scheme has been proposed that leverages inter-satellite links within the orbit, the predictability of satellite movements, and partial aggregation methods to significantly reduce training time and communication costs [400]. FedSpace [401] dynamically schedules model aggregation, leveraging the deterministic and time-varying connectivity of satellite orbits and Earth's rotation. An asynchronous satellite system AsyncFLEO [402] that utilizes high-altitude platforms as parameter servers, addressing both idle waiting in synchronous FL and the model staleness issues caused by lagging satellites. An asynchronous distributed algorithm called FedGSM [403], introduces a compensation mechanism that leverages the deterministic and time-varying characteristics of satellite orbits to mitigate the adverse effects of gradient staleness. The proposed asynchronous distributed algorithm called FedBuff [404], utilizes buffered asynchronous aggregation to address scalability and privacy issues in cross-device FL.

Compared to synchronous architecture, asynchronous architecture has the advantage of improving computational efficiency in CCS system, as each satellite can independently update its model, reducing global wait time and allowing all satellites to adapt to varying computational capabilities and network delays. When some satellites fail, other satellites can continue to update, which may accelerate the convergence of the global model in certain scenarios. However, the drawbacks of asynchronous architecture include instability in the global state. The independence of the satellites can lead to inconsistencies in the global updates, resulting in fluctuations in the performance of the global model. Additionally, frequent asynchronous updates may increase network burden and communication overhead. The algorithms involved in asynchronous updates are also more complex, requiring greater technical investment to design effective asynchronous mechanisms.

For distributed on-orbit information processing, both synchronous and asynchronous architectures face significant technical challenges and research opportunities. Due to the relative motion between satellites, the constellation topology changes rapidly, making it difficult to maintain stable and efficient inter-satellite links. Optimizing resource allocation among satellites is crucial. Each satellite has limited and heterogeneous computational, energy, and storage resources, necessitating the deployment of effective algorithms for real-time task and resource allocation [405].

5.3 Semantic communications in DSIN

The 6G-enabled DSIN is expected to provide ubiquitous intelligent services with stringent QoS requirements in dynamic network environments, which prompts a shift from conventional architectures that focus on high transmission rates between transceivers to new architectures centered on intelligent connectivity of everything. The resulting extensive scales of data, which require a massive amount of bits to represent, impose a huge burden and bottleneck on communication systems that handle only bit-level reconstruction, especially on DSIN with limited transmission power, storage space and computing resources [406]. In response, a novel paradigm known as semantic communication is inspired as a potential technology to break the bottleneck by utilizing the semantics and content of data to optimize the utilization of communication resources [407, 408]. For example, when the CCS system collaboratively executes a remote sensing task, each satellite extracts semantic information about the target in orbit and transmits it to the ground station or device. This effectively reduces the volume of data transmitted over the link and improves transmission timeliness. In addition, in poor channel conditions, semantic communication proves more robust to fading than traditional bit-level reliable transmission methods, offering greater resilience for DSIN [409, 410]. In this sense, semantic communication can be viewed as a promising development direction for the future of DSIN.

Specifically, motivated by the three communication levels identified by Shannon and Weaver [411].

• Level A addresses the technical problem, which answers "How accurately can the symbols of communication be transmitted?"

• Level B tackles the semantic problem, and asks "How precisely can the transmitted symbols convey the desired meaning?"

• Level C solves the effectiveness problem, where "How effectively do the received symbols affect conduct in the desired way?" is the core.

Semantic communications in DSIN enable all participants to achieve Level B and Level C communication with minimal overhead by leveraging advanced AI technology [412]. This approach ensures that only the most valuable and contextually relevant information is transmitted to applications, efficiently distilling

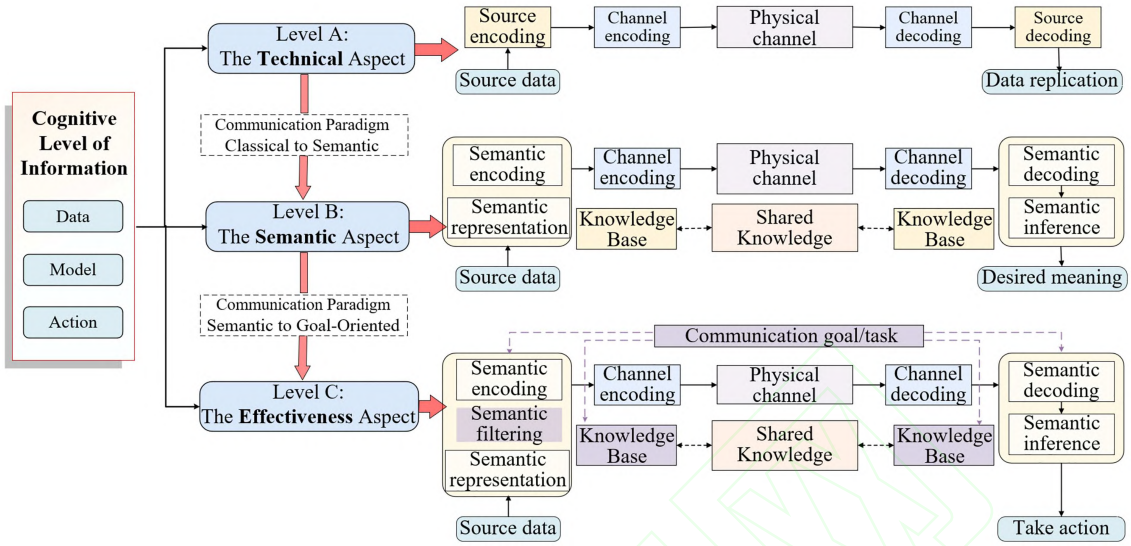

Figure 22 (Color online) Illustration of general communication system models for three communication levels.

insights from massive datasets and delivering them to their destinations at the right time, as illustrated in Figure 22. As a result, it serves as a key technology for ensuring high QoS service in DSIN via extracting, transmitting, and evaluating critical semantic information. Moreover, the advancement of semantic communication and the construction of DSIN are mutually reinforcing. On the one hand, the distributed computing and AI networks facilitated by space-terrestrial collaboration support the large-scale deployment of semantic communication systems [413]. On the other hand, semantic communication significantly enhances network performance, particularly in hardware-constrained satellite networks, which effectively broadens the applications and versions of DSIN. Currently, the work on semantic/effectiveness-empowered communications primarily follows two approaches: the application-centric engineering implementations and the universal representation of information importance.

(1) Application-centric engineering implementations. This approach stems from attempts to leverage DL at the physical layer for representing underlying data meanings in the absence of a rigorous theoretical framework for semantic communications [414]. In this case, the work in this vein focuses on the architecture design and engineering implementation of semantic compression and extraction. An early prototype of end-to-end semantic communication is deep joint source-channel coding (Deep JSCC), in which textual sentences are encoded into fixed-length bit streams within a simple channel environment [415]. Following this, several researchers have developed semantic communication systems tailored for image [416] and speech [417] transmission based on this framework, with the goal of achieving efficient reconstruction of image and speech data at the receiver, respectively. In addition, for task-specific applications, semantic communication systems must be capable of extracting task-relevant information at the transmitter and performing correct decision-making relying on the limited received information at the receiver. For instance, a speech recognition-oriented semantic communication system has been developed in [418], where the receiver can reduce the volume of data from the transmitter by converting the received speech signals into text, and then applying speech synthesis to reconstruct the original signal. Similarly, to achieve accurate image recognition with minimal overhead, the author in [419] proposed a classification-oriented semantic communication system, which can accurately identify various objects based on the transmitted image features, rather than the full image, under limited bandwidth and power constraints.

Moreover, given the practical necessity of gathering multi-modal data from diverse users/devices and fusing them at the receiver, a Transformer-based framework with high representational capacity can be employed to unify the transmitter structure across various tasks [420, 421]. This approach enables the extension from single-user, single-modal systems to multi-user, multi-modal scenarios, such as integrating text and image information for complex tasks like visual question answering (VQA). Furthermore, the complexity of multi-modal multi-tasks and the variability of transmission environments raise the need for stronger generalization in semantic communication systems, which drives the advancements in knowledge base (KB)-based approaches [422]. In typical satellite networks, such as onboard remote sensing processing, KB-sharing mechanisms are widely integrated into semantic encoding modules to enhance

the utilization of ground-based remote sensing knowledge and expert insights, which can sustain long-term semantic alignment between transceivers. Therefore, the KB-assisted semantic extraction further strengthens the ability of satellites for semantic representation across multi-source data and multi-modal tasks in complex space-terrestrial channels [423].

(2) Universal representation of information importance. This approach focuses on the etymological meaning of semantics, i.e., the importance or priority of information [424]. From this perspective, the priority is not determined by classical information entropy but rather relies on the semantics conveyed by the information, where the semantic metrics are no longer confined to specific applications outlined in the first approach, such as the semantic distance in text or structural similarity index (SSIM) in images [425]; instead, they serve as a more comprehensive multidimensional metric that reflects the overall performance of the network. Within such a paradigm, the semantic communication systems are capable of recognizing high-priority information and adjusting resource allocation accordingly. For example, a packet is given more priority when its destination has not been updated for a while in the AoI-based framework [426]. However, the assessment of information importance in AoI relies on the assumption that fresher messages always contain more valuable information, which makes it overlook the differing temporal requirements in various applications and fails to capture how the content of information impacts the task execution. In response, the value of information (VoI) is born to address the above problems, typically denoted by a non-linear penalty function of AoI, i.e., $f(\mathrm{AoI}(t))$ [427]. Thus, on the one hand, VoI can capture the varying sensitivity to freshness across different applications by mapping age penalty to linear, exponential, or logarithmic functions [428]. Nevertheless, the selection of penalty functions completely relies on empirical setups, and thus some efforts aim to derive the function stringently, such as the mutual information-based construction [429]. On the other hand, it can reflect the differences in information content related to the transceiver state $(X_t, \hat{X}_t)$ through a content-aware error penalty function $g(X_t, \hat{X}_t)$, such as mean squared error (MSE) or threshold error [412]. It is worth noting that the error function can also be regarded as a kind of nonlinear AoI.

Moreover, since the above metrics only focus on one attribute of information at a given time, several semantic metrics that combine multiple attributes have been proposed. One of the most notable examples is age of incorrect information (AoII) [430], which quantifies the utility reduction from asynchronous duration between transceivers by simply integrating $f(\mathrm{AoI}(t))$ with $g(X_t, \hat{X}_t)$ [414]. On this basis, by further incorporating a cost function C dependent on real-world constraints such as bandwidth limits and energy consumption, the authors in [431] proposed a three-dimensional utility of information (UoI) evaluation framework that can simultaneously encompass timeliness, accuracy, and energy efficiency of information. Further, UoI expands the error function $g(X_t, \hat{X}_t)$ to capture not only packet-level physical process mismatches but also the application-layer semantic states required for extraction and recovery, such as KBs [432]. In this regard, the improvement both broadens the application scenarios of UoI and initiates a preliminary integration with the first approach.

Overall, current semantic-empowered communication, particularly in the context of DSIN, still faces several challenges for future investigations. (1) Research in semantic information theory: the shortcomings in foundational theoretical research have led to issues such as the lack of interpretability and explainability in semantic extraction, as well as an unclear relationship between semantic rate and Shannon bit rate, which makes the real implementation of semantic communication systems in DSIN challenging. (2) Flexible access and handover: the inherent flexibility of satellites, especially those in CCS systems, requires constant adaptation of communication modes in response to real-time environmental changes, which imposes more demands on the generalization capabilities of semantic communication systems, including the time and resources needed for model retraining and the continual expansion of KBs in dynamic environments. (3) Inconsistent KB and sharing costs: information acquisition from the broad coverage of DSIN will create inconsistencies in KBs across different nodes, leading to mismatched semantic networks. However, frequent updates to maintain semantic network alignment amid constantly expanding KBs are extremely time- and resource-consuming.

5.4 Goal-oriented integrated sensing, communication, and computation

As DSIN continues to advance, a diverse range of communication-assisted applications is emerging, including remote sensing, disaster management, and environmental monitoring, each requiring integrated sensing and computational capabilities along with distinct communication performance metrics [433]. For instance, emerging applications such as on-orbit object detection and tracking necessitate not only

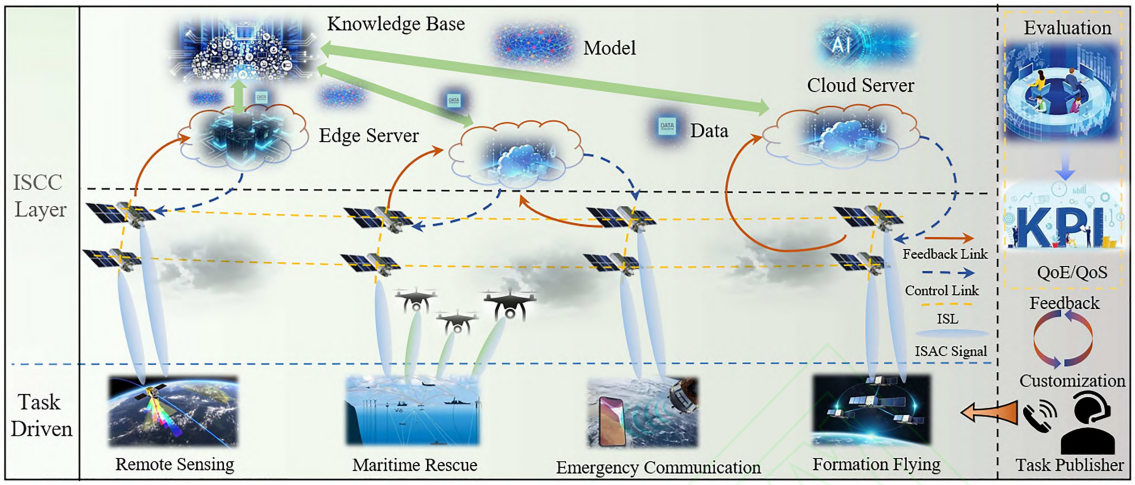

Figure 23 (Color online) Illustration of goal-oriented ISCC for DSIN.

reliable communication capabilities but also sophisticated computing and sensing functionalities, underpinned by the satellite network infrastructure. As a transformative concept within DSIN, the integrated sensing, communication, and computation (ISCC) has begun to attract increasing attention from both the academic and industrial sectors [434, 435], and has recently been adopted among the key usage scenarios for IMT-2030/6G by the Radio Communication Division of the International Telecommunication Union (ITU-R) [436]. While recent studies have proposed several approaches to address these challenges of ISCC in satellite networks, i.e., ISCC in air-ground integrated networks [437] and S-IoT [433], goal-oriented ISCC in DSIN remains under-explored in several key areas, with much of the research in preliminary phases. As shown in Figure 23, the goal-oriented ISCC mechanisms hold the potential to redefine how we utilize heterogeneous space-air-ground resources to provide on-demand services for diverse applications or tasks, bridging the gaps between sensing, communication, and computation to create autonomous and resilient information network. Nevertheless, integrating goal-oriented ISCC into DSIN faces several significant challenges, primarily stemming from the limitations of heterogeneous resources and the inherent difficulties of distributed collaborative scheduling within DSIN.

From a computing perspective, the DSIN is restricted by the computational limitations of on-orbit hardware, which impacts the real-time processing capabilities. Traditional satellites rely on GSs for data processing, causing large delays that undermine mission-critical applications. Current solutions, such as deploying high-performance processors, are limited due to power constraints and radiating availability. One promising direction is the use of edge computing [438], where data is processed closer to the sensing source (e.g., on the satellite itself or nearby satellites). This reduces latency and conserves bandwidth by minimizing the amount of data transmitted to GS. Another promising approach involves hybrid edge-cloud solutions, where less time-sensitive tasks are processed in terrestrial cloud servers, while critical tasks are handled at the edge. However, this model necessitates robust coordination across network layers. Building on this, some research advocates for the use of distributed computing within the CCS system to manage resources more effectively. Despite its potential, this approach faces significant challenges, particularly due to the instability of inter-satellite communications [437].

From a communication perspective, the spectrum resources are actually limited in the CCS system, leading to potential conflicts and interference. Besides, frequent handovers in the CCS system further exacerbate the accuracy of information perception [434]. Thus, goal-oriented dynamic spectrum management, where satellites adaptively select frequency bands based on the availability of resources and the need by task, has been proposed as a solution. However, goal-oriented dynamic spectrum management requires advanced low-complexity algorithms and policies to manage spectrum resources efficiently, particularly when coordinating with terrestrial networks. To deal with this issue, some existing approaches utilize AI-driven or multi-agent cooperative scheduling approaches to maximize resource utilization [439]. The AI and ML techniques aim to allocate resources dynamically based on real-time satellite and mission status, allowing for enhancing the adaptability of ISCC in DSIN, but they are still in the early stages and remain susceptible to inaccuracies in rapidly changing environments, such as unexpected environmental conditions or equipment malfunctions. In a multi-agent CCS system, each satellite is equipped with

multiple onboard sensors, such as radars and cameras, enabling the capture of multi-modal, task-specific environmental data and its on-orbit processing. By integrating the dual functions of radio sensing and communication within a unified infrastructure, the goal-oriented ISCC in the CCS system can facilitate signal-level sensing information fusion, complementing information-level exchange among agents [439].

However, the mutual interactions among distributed satellites make the design of online ISAC strategies highly complex. While DRL has been explored for ISAC strategy development in multi-agent CCS systems, the dynamic nature of target determination and ISAC strategy design introduces substantial computational and operational complexity, limiting the applicability of traditional DRL algorithms. In addition, the ISCC operations consume significant power, posing a challenge to maintaining energy efficiency. Energy-aware sensing, computation, and communication protocols have been developed to address this issue [440, 441], yet they often involve trade-offs in performance, which can undermine the effectiveness of ISCC in real-time applications. Research into power-harvesting technologies in space, such as advancements in solar panel efficiency, shows promise but remains limited in its impact, particularly for low-power satellites.

5.5 Direct satellite-to-device communications

With the breakthrough development in satellite communication techniques and the continuous reduction in manufacturing and launch costs of satellite platforms, the DSIN is expected to enable ubiquitous services and diverse applications to user terminals, i.e., direct satellite-to-device communications. This application, as implied by its name, seeks to provide high-performance direct connectivity between satellites and common devices, and can further broaden its services to include mobile phones, and IoT terminals deployed in vehicles, aircraft, and maritime platforms [442]. Therefore, different from the conventional mobile satellite communication systems (MSCS) that focus solely on industrial users, direct satellite-to-device communications can offer diverse services for consumers in regions lacking coverage from terrestrial mobile networks with voice, messaging, and broadband Internet connectivity, which significantly expands the market size of SatCom [442]. Its commercial potential has also fostered deep collaboration across the entire industry chain, including satellite manufacturers and operators, cellular equipment manufacturers, mobile phone manufacturers, and mobile network operators, emerging as a focal point of interest among academia, industry, and research communities over the past three years [443]. Currently, various international organizations and telecommunication regulatory authorities from multiple countries have expressed their intention to promote the development of direct satellite-to-phone communication with substantial policy support. Furthermore, industry leaders across related sectors have continued to increase their strategic investments in this area, making competition increasingly fierce.

5.5.1 *Technical routes*

As of now, a unified route for direct satellite-to-phone communication has yet to be materialized, and it can be roughly categorized into three types based on its commercial progress.

• Dual-mode mobile device with satellite-terrestrial independent systems. This route involves embedding dedicated SatCom chips or specific waveforms into a common mobile device and utilizing existing MSCS to provide services for these intelligent devices. This technical route is relatively mature, with products already available, notably Huawei's Mate60 Pro [444] and Apple's iPhone 14 [445], which utilize the TianTong-1 and GlobalStar satellites to provide direct connectivity services, respectively. However, due to the non-standardized technologies of MSCS and their outdated air interface protocols, the communication capabilities of dual-mode devices are primarily restricted to voice and messaging, intended to provide emergency information transmission services for users in remote areas. Thus, it is mostly viewed as an interim solution for direct satellite-to-phone communication.

• Mobile device with terrestrial system directly connected to satellites. This route aims to provide direct connectivity services without altering the configurations of current ground-based mobile phones by manufacturing and launching satellite constellations specifically designed for this application. The primary challenge of this route lies in overcoming the significant attenuation associated with long-distance satellite-terrestrial transmission, particularly given the extremely limited performance of mobile device antenna. In response to this challenge, ASTS utilizes the BlueWalker-3 satellite equipped with a 64 m^2 phased array antenna to conduct 5G connectivity tests with existing mobile phones [446]. Meanwhile, Lynk Global plans to deploy a complete LTE network in space and has currently completed the field experiments of the bidirectional voice communication between the test satellite and existing mobile phones

[447]. Furthermore, SpaceX introduces the Starlink V2 satellite, which features an additional 25 m^2 array antenna compared to V1, and achieves a downlink communication speed of 16.9 Mbit/s with unmodified mobile phones during tests [448]. Nevertheless, while this route can provide broadband internet services, it places all modifications on the satellites, leading to higher engineering implementation costs.

• Integration of satellite-terrestrial protocols based on 3GPP NTN. This route requires the adoption of a unified protocol on both the satellite and terrestrial sides, enabling the next generation of mobile phones to have the capability to access both satellite and terrestrial networks simultaneously. To realize this vision, the NTN working group in 3GPP has completed the 5G New Radio (NR) standard for NTN in Release-17 [449] and its enhancement technologies to support direct satellite-to-device service in Release-18 [450]. Moreover, 3GPP plans to conduct research on supporting the standardization of spaceborne processing mode and spaceborne BS in Release-19 and Release-20 [451]. In terms of standard implementation, satellite operators like Omnispace and EchoStar, equipment manufacturers such as ZTE, Unisoc and MediaTek, and mobile network operators like China Telecom, and China Mobile, have initiated validation work for spaceborne BSs based on 3GPP NTN standards and their performance for direct NTN satellite-to-device communications. However, the time required for standards to move from development to implementation is considerable, which hinders the rapid commercialization of this route.

5.5.2 *Challenges and solutions*

In practice, direct satellite-to-device communication encounters numerous technical challenges. Firstly, due to the size and shape limitations, the typically low-gain omnidirectional antennas used in existing mobile phones struggle to establish communication with conventional satellites over long distances with significant signal loss. As a result, improvements in aspects such as antenna design, power consumption, and weight constrain are needed for the next generation of satellites to enhance the transmission and reception capabilities of spaceborne antennas [452].

Another challenge lies in the spaceborne payload design. Note that spaceborne BS deployment is a key trend in SatCom, as it significantly reduces network latency and leverages inter-satellite links to enable large-scale satellite networking, which can provide more efficient service for mass mobile users. Following this, the weight and power limitations of satellite platforms impose stringent requirements on spaceborne BS payloads, necessitating advancements in lightweight design, high integration, and thermal management [453]. Although the partial BS deployment based on DU-CU separation is a potential scheme enabling effective tradeoffs between payload and cost, consensus on how to segment functional modules for optimal performance within acceptable cost remains elusive.

Additionally, to address the challenges posed by high Doppler shifts from satellite movement, time-varying network topologies, and uneven spatial and temporal distribution of traffic, it is essential to further enhance the adaptability of existing SatCom protocols to the characteristics of satellite networks. More precisely, the technology advancements in adaptive coding and modulation, time-frequency synchronization among DSIN, large-scale random access, intra-/inter-satellite beam switching, frequency-sharing between satellite and terrestrial network, and so forth [79], are necessary to meet the demand for high QoS to mobile users with direct connectivity.

Motivated by the above analysis, DSIN can be seen as an efficient transmission architecture to address challenges shown in Figure 24.

• Distributed satellite-based array antennas. The satellites equipped with standard array antennas can tune with each other to create a larger equivalent aperture, providing substantial gain and narrow beam coverage [454]. Alternatively, they can also operate as a virtual MIMO system through collaboration, allowing individual mobile users to receive signals from multiple satellites [455]. The above configuration enables coherent signal transmission and reception, i.e., enhances the desired gain while suppressing interference, ultimately improving the SNR and spatial resolution. Moreover, the deployment of sub-arrays or units of a distributed array antenna across various satellite platforms cleverly circumvents the challenges posed by the conventional approach of loading an ultra-large multi-beam array antenna on a single satellite, which includes the technical demands of large deployable array antennas and the stringent design requirements for rocket fairings. As a result, the distributed satellite-based array antenna significantly reduces the costs of satellite production and launch while offering similar gains.

• Degree of freedom and fault tolerance. The distributed nature of DSIN introduces greater flexibility regarding inter-satellite spacing, number of satellites, and payload configurations [456]. Considering the design of spaceborne payload, a BS can be categorized into different functional modules, and then be

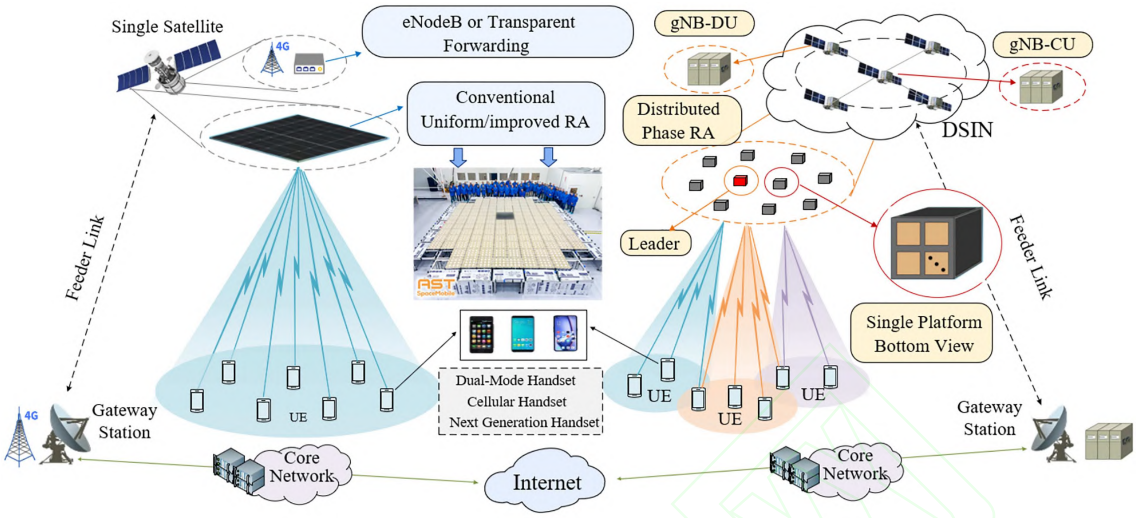

Figure 24 (Color online) Comparison of direct satellite-to-device communication in DSIN and existing satellite communication system.

allocated to various satellite platforms within a CCS system formation to achieve the full functionality of a complete BS, while satisfying constraints such as weight and power consumption of a satellite through collaboration. On the other hand, leveraging the CU-DU separation architecture, we can position the CU on the leader satellite within a formation, which enables the CU satellite to manage and coordinate multiple follower DU satellites for access and forwarding, thus collectively functioning as a powerful spaceborne BS. Moreover, more refined network function slicing policies can be implemented, such as further dividing the CU into CU-CP and CU-UP modules [14], with the CU-UP placed on select follower satellites to adapt to various communication goals or tasks. Furthermore, this distributed functional slicing or redundancy also enhances the overall fault tolerance of the CCS system, since the failure of one or more satellite platforms within the DSIN may result in performance degradation, but it does not lead to service interruption.

• Scalability with adjustable topology and platform. In DSIN, the number of satellites, the platform capabilities, and the topological configuration are all adjustable, which allows for low-cost maintenance and flexible application changes by simply replacing certain satellites within a CCS formation, showcasing its strong scalability [7]. On this basis, by integrating the previously discussed air interface technologies and collaborative cross-layer optimization into individual satellite transmission mechanisms and multi-satellite coordinated management, respectively, the system can effectively tackle the challenges arising from the complex transmission environments, uneven network traffic, and diverse user demands in the satellite network. Consequently, the DSIN can provide continuous, reliable, and timely services to ground mobile users under limited payloads.

It is worth noting that the practical deployment of these technologies still faces substantial hurdles. For example, the distributed satellite-based array antennas and spaceborne BS architectures remain largely at the theoretical and research stages, facing challenges like maintaining precise time-frequency synchronization, ensuring stable formation flying under limited power, and adapting beam management and resource optimization to complex scenarios. Further development and rigorous validation are essential for advancing these technologies toward application readiness.

6 Conclusion

Benefiting from rapid progress in commercial space and satellite technologies, satellite networks are transitioning into a new era, evolving towards DSIN. This survey provides an in-depth exploration of the critical network architectures underpinning DSIN, focusing on distributed regenerative satellite network architecture, distributed satellite computing network architecture, and reconfigurable formation control. A comprehensive review for enabling technologies of DSIN followed, covering essential air interface and collaborative transmission technologies (e.g., waveform design, GFRA, NOMA/RSMA multicast, cloud-native distributed MIMO cooperation) and cross-layer optimization techniques (e.g., mobility manage-

ment, resource allocation, secure communication), alongside the testbeds of DSIN. Finally, the survey identified several open research challenges and promising directions for future investigation, providing a valuable roadmap for advancing DSIN research. By providing a thorough, multi-faceted overview, this work serves as a critical resource for researchers and developers, offering valuable insights to guide future innovations and propel the evolution of DSIN technologies at the forefront of satellite network research.

Acknowledgements This work was supported in part by Major Key Project of PCL (Grant No. PCL2024A01), National Natural Sciences Foundation of China (NSFC) (Grant Nos. 62027802, 62071141), Shenzhen Science and Technology Program (Grant Nos. JCYJ20241202123904007, GXWD20231127123203001, JSGG20220831110801003), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2025A1515010281), Fundamental Research Funds for the Central Universities (Grant No. HIT.OCEF.2024046), National Key R&D Program of China (Grant No. 2020YFB1806900), and Beijing Natural Science Foundation (Grant No. L242012).

References

1 You X, Wang C X, Huang J, et al. Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts. Sci China Inf Sci, 2021, 64: 110301
2 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
3 Peng C, He Y, Zhao S, et al. Integration of data center into the distributed satellite cluster networks: challenges, techniques, and trends. IEEE Netw, 2023, 37: 52–58
4 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
5 He Y, Wang C, Qi C, et al. Spatial ultra-sparse distributed antenna satellite-ground cooperative transmission architecture: challenges, key technologies, and trends. IEEE Commun Mag, 2024, 62: 136–143
6 Zhang N, Li Z Y, Li R Y, et al. 5G-Advanced: network-controlled repeater (in Chinese). Telecommun Sci, 2022, 38: 169–176
7 Tuzi D, Delamotte T, Knopp A. Satellite swarm-based antenna arrays for 6G direct-to-cell connectivity. IEEE Access, 2023, 11: 36907–36928
8 Tuzi D, Delamotte T, Knopp A. Performance assessment of sparse satellite swarms for 6G direct-to-cell connectivity. In: Proceedings of IEEE International Workshop on Signal Processing Advances in Wireless Communications, Lucca, 2024. 616–620
9 Generation Partnership Project (3GPP). Study on new radio (NR) to support nonterrestrial networks (NTN) (Release 16). TR 38.811. 2020. https://www.3gpp.org/ftp/Specs/archive/38_series/38.811
10 Bhandari S, Vu T X, Chatzinotas S. User-centric flexible resource management framework for LEO satellites with fully regenerative payload. IEEE J Sel Areas Commun, 2024, 42: 1246–1261
11 Yahia O B, Zineb G, Olivier B, et al. Evolution of high-throughput satellite systems: a vision of programmable regenerative payload. IEEE Commun Surv Tut, 2025, 27: 1565–1597
12 Choi J, Li B, Al Homssi B, et al. Spectrum sharing through marketplaces for O-RAN based non-terrestrial and terrestrial networks. IEEE Int Things M, 2024, 7: 128–134
13 Yoshida Y, Mobile Xhaul evolution: enabling tools for a flexible 5G Xhaul network. In: Proceedings of Optical Fiber Communications Conference and Exposition, San Diego, 2018. 1–85
14 Generation Partnership Project (3GPP). NG-RAN; architecture description (Release 17). TS 38.401. 2025. https://www.3gpp.org/ftp/Specs/archive/38_series/38.401
15 Generation Partnership Project (3GPP). Study on new radio access technology; radio access architecture and interfaces (Release 14). TR 38.801. 2017. https://www.3gpp.org/ftp/Specs/archive/38_series/38.801
16 Joda R, Pamuklu T, Iturria-Rivera P E, et al. Deep reinforcement learning-based joint user association and CU-DU placement in O-RAN. IEEE Trans Netw Serv Manage, 2022, 19: 4097–4110
17 Wang S, Li Q. Satellite computing: vision and challenges. IEEE Int Things J, 2023, 10: 22514–22529
18 Li S, Maddah-Ali M A, Yu Q, et al. A fundamental tradeoff between computation and communication in distributed computing. IEEE Trans Inform Theor, 2018, 64: 109–128
19 Ng J S, Lim W Y B, Luong N C, et al. A comprehensive survey on coded distributed computing: fundamentals, challenges, and networking applications. IEEE Commun Surv Tut, 2021, 23: 1800–1837
20 Duan S, Wang D, Ren J, et al. Distributed artificial intelligence empowered by end-edge-cloud computing: a survey. IEEE Commun Surv Tut, 2023, 25: 591–624
21 Yang Y, He X, Lee J, et al. Collaborative deep reinforcement learning in 6G integrated satellite-terrestrial networks: paradigm, solutions, and trends. IEEE Commun Mag, 2025, 63: 188–195
22 Deng S, Zhao H, Huang B, et al. Cloud-native computing: a survey from the perspective of services. Proc IEEE, 2024, 112: 12–46
23 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
24 Pan X H, Lu L, Deng P K, et al. Architecture, technologies and experiment of satellite-terrestrial network based on space-based simplified core network (in Chinese). Telecommun Sci, 2024, 40: 49–59
25 Wang Y, Yang C, Lan S, et al. End-edge-cloud collaborative computing for deep learning: a comprehensive survey. IEEE Commun Surv Tut, 2024, 26: 2647–2683
26 Gao G, Yao L, Li W, et al. Onboard information fusion for multisatellite collaborative observation: summary, challenges, and perspectives. IEEE Geosci Remote Sens Mag, 2023, 11: 40–59
27 Li D X, Xie W, Li Y, et al. FedFusion: manifold-driven federated learning for multi-satellite and multi-modality fusion. IEEE Trans Geosci Remote Sens, 2024, 62: 1–13
28 Xi S Y, Shang B D, Zhang H X, et al. Multi-satellite-enabled edge computing: an offloading and computation integration approach. In: Proceedings of International Conference on Intelligent Communications and Computing, Nanchang, 2023. 151–156

29 Cui G, Duan P, Xu L, et al. Latency optimization for hybrid GEO-LEO satellite-assisted IoT networks. IEEE Int Things J, 2023, 10: 6286–6297
30 Gong Y, Yao H, Nallanathan A. Intelligent sensing, communication, computation, and caching for satellite-ground integrated networks. IEEE Netw, 2024, 38: 9–16
31 Zhang P, Xu X, Dong C, et al. Model division multiple access for semantic communications. Front Inform Technol Electron Eng, 2023, 24: 801–812
32 Liu G P, Zhang S. A survey on formation control of small satellites. Proc IEEE, 2018, 106: 440–457
33 Yin J F, Zhang Q J, Liu J, et al. A review on development of formation flying interferometric SAR satellite system. Spacecraft Eng, 2018, 27: 116–122
34 Clohessy W H, Wiltshire R S. Terminal guidance system for satellite rendezvous. J Aerospace Sci, 1960, 27: 653–658
35 Tschauner J, Hempel P. Optimale beschleunigungsprogramme fur das rendezvous-manover. Astron Acta, 1964, 10: 296
36 Lei C, Feng W, Wei P, et al. Edge information hub: orchestrating satellites, UAVs, MEC, sensing and communications for 6G closed-loop controls. IEEE J Sel Areas Commun, 2025, 43: 5–20
37 Wang D W, Wu B L, Poh E K. Satellite Formation Flying: Relative Dynamics, Formation Design, Fuel Optimal Maneuvers and Formation Maintenance. Singapore: Springer Press, 2016
38 Cai H, Huang J. Leader-following adaptive consensus of multiple uncertain rigid spacecraft systems. Sci China Inf Sci, 2016, 59: 010201
39 Zou A M, Kumar K D. Neural network-based distributed attitude coordination control for spacecraft formation flying with input saturation. IEEE Trans Neural Netw Learn Syst, 2012, 23: 1155–1162
40 Balch T, Arkin R C. Behavior-based formation control for multirobot teams. IEEE Trans Robot Automat, 1998, 14: 926–939
41 Schlanbusch R, Kristiansen R, Nicklasson P J. Spacecraft formation reconfiguration with collision avoidance. Automatica, 2011, 47: 1443–1449
42 Abbasi Y, Moosavian S A A, Novinzadeh A B. Formation control of aerial robots using virtual structure and new fuzzy-based self-tuning synchronization. Trans Inst Measurement Control, 2017, 39: 1906–1919
43 Song Y Y, Zhou Q R, Chen Q W, et al. Optimal reconfiguration control of electromagnetic satellite formation. In: Proceedings of the 41st Chinese Control Conference, Hefei, 2022. 1809–1814
44 Wang P K C, Hadaegh F Y. Minimum-fuel formation reconfiguration of multiple free-flying spacecraft. J Astronaut Sci, 1999, 47: 77–102
45 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
46 Blackmore L, Hadaegh F. Necessary and sufficient conditions for attitude estimation in fractionated spacecraft systems. In: Proceedings of AIAA Guidance, Navigation, and Control Conference, Chicago, 2009
47 Li H, Liao X, Huang T, et al. Event-triggering sampling based leader-following consensus in second-order multi-agent systems. IEEE Trans Automat Contr, 2014, 60: 1998–2003
48 Lü Q, Li H, Xia D. Distributed optimization of first-order discrete-time multi-agent systems with event-triggered communication. Neurocomputing, 2017, 235: 255–263
49 Garcia E, Cao Y C, Casbeer D W. Cooperative control with general linear dynamics and limited communication: centralized and decentralized event-triggered control strategies. In: Proceedings of American Control Conference, Portland, 2014. 159–164
50 Zhu W, Jiang Z P, Feng G. Event-based consensus of multi-agent systems with general linear models. Automatica, 2014, 50: 552–558
51 Fan M C, Chen Z, Zhang H T. Semi-global consensus of nonlinear second-order multi-agent systems with measurement output feedback. IEEE Trans Automat Contr, 2014, 59: 2222–2227
52 Seuret A, Prieur C, Tarbouriech S, et al. LQ-based event-triggered controller co-design for saturated linear systems. Automatica, 2016, 74: 47–54
53 Sconzo P. The use of Lambert's theorem in orbit determination. Astron J, 1962, 67: 19–21
54 Nelson S L, Zarchan P. Alternative approach to the solution of Lambert's problem. J Guidance Control Dyn, 1992, 15: 1003–1009
55 Avanzini G. A simple Lambert algorithm. J Guidance Control Dyn, 2008, 31: 1587–1594
56 Zhang G. Terminal-velocity-based lambert algorithm. J Guidance Control Dyn, 2020, 43: 1529–1539
57 Yang Z, Luo Y Z, Zhang J, et al. Homotopic perturbed Lambert algorithm for long-duration rendezvous optimization. J Guidance Control Dyn, 2015, 38: 2215–2223
58 Zhang H, Zhang G. Reachable domain of ground track with a single impulse. IEEE Trans Aerosp Electron Syst, 2020, 57: 1105–1122
59 Thorne J D. Convergence behavior of series solutions of the Lambert problem. J Guidance Control Dyn, 2015, 38: 1821–1826
60 Wang C X, You X, Gao X, et al. On the road to 6G: visions, requirements, key technologies, and testbeds. IEEE Commun Surv Tut, 2023, 25: 905–974
61 Xiang Z, Gao X, Li K X, et al. Massive MIMO downlink transmission for multiple LEO satellite communication. IEEE Trans Commun, 2024, 72: 3352–3364
62 Milojevic M, Haardt M, Eberlein E, et al. Channel modeling for multiple satellite broadcasting systems. IEEE Trans Broadcast, 2009, 55: 705–718
63 Tawfik M M, Sree M F A, Abaza M, et al. Performance analysis and evaluation of inter-satellite optical wireless communication system (IsOWC) from GEO to LEO at range 45000 km. IEEE Photon J, 2021, 13: 1–6
64 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
65 Homssi B A, Chan C C, Wang K, et al. Deep learning forecasting and statistical modeling for Q/V-band LEO satellite channels. Trans Mach Learn Comm Netw, 2023, 1: 78–89
66 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
67 Zuo Y, Yue M, Zhang M, et al. OFDM-based massive connectivity for LEO satellite Internet of Things. IEEE Trans Wireless Commun, 2023, 22: 8244–8258
68 You L, Li K X, Wang J, et al. Massive MIMO transmission for LEO satellite communications. IEEE J Sel Areas Commun, 2020, 38: 1851–1865
69 Li K X, Gao X, Xia X G. Channel estimation for LEO satellite massive MIMO OFDM communications. IEEE Trans Wireless Commun, 2023, 22: 7537–7550
70 Wang Y, Li Q, Jiao J, et al. ARM: adaptive random-selected multi-beamforming estimation scheme for satellite-based Internet of Things. IEEE Access, 2019, 7: 63264–63276
71 Zhang Y, Liu A, Li P, et al. Deep learning (DL)-based channel prediction and hybrid beamforming for LEO satellite massive MIMO system. IEEE Int Things J, 2022, 9: 23705–23715
72 Lu S, Yu B G, Tang C K, et al. MIMO-OFDM channel detection algorithm in multi-station and multi-satellite uplink system based on deep learning. In: Proceedings of IEEE International Conference on Signal Processing, Communications

and Computing, Xi'an, 2021. 1–4
73 Yuan J, Ngo H Q, Matthaiou M. Machine learning-based channel prediction in massive MIMO with channel aging. IEEE Trans Wireless Commun, 2020, 19: 2960–2973
74 Zhang Y, Wu Y, Liu A, et al. Deep learning-based channel prediction for LEO satellite massive MIMO communication system. IEEE Wireless Commun Lett, 2021, 10: 1835–1839
75 Ma W, Qi C, Zhang Z, et al. Sparse channel estimation and hybrid precoding using deep learning for millimeter wave massive MIMO. IEEE Trans Commun, 2020, 68: 2838–2849
76 Ying M, Chen X, Qi Q, et al. Deep learning-based joint channel prediction and multibeam precoding for LEO satellite Internet of Things. IEEE Trans Wireless Commun, 2024, 23: 13946–13960
77 Gizzini A K, Medjahdi Y, Ghandour A J, et al. Towards explainable AI for channel estimation in wireless communications. IEEE Trans Veh Technol, 2024, 73: 7389–7394
78 Xiang Z, Sun R, Gong X, et al. Massive MIMO uplink transmission for multiple LEO satellite communication. IEEE Trans Aerosp Electron Syst, 2025, 61: 4852–4865
79 Bakhsh Z M, Omid Y, Chen G, et al. Multi-satellite MIMO systems for direct satellite-to-device communications: a survey. IEEE Commun Surv Tut, 2024. doi: 10.1109/COMST.2024.3449430
80 Bjornson E, Sanguinetti L. Making cell-free massive MIMO competitive with MMSE processing and centralized implementation. IEEE Trans Wireless Commun, 2020, 19: 77–90
81 Abdelsadek M Y, Kurt G K, Yanikomeroglu H. Distributed massive MIMO for LEO satellite networks. IEEE Open J Commun Soc, 2022, 3: 2162–2177
82 Wang D M, You X H, Huang Y M, et al. Full-spectrum cell-free RAN for 6G systems: system design and experimental results. Sci China Inf Sci, 2023, 66: 130305
83 Dalgitsis M, Cadenelli N, Serrano M A, et al. Cloud-native orchestration framework for network slice federation across administrative domains in 5G/6G mobile networks. IEEE Trans Veh Technol, 2024, 73: 9306–9319
84 Cao Y, Wang P, Zheng K, et al. Experimental performance evaluation of cell-free massive MIMO systems using COTS RRU with OTA reciprocity calibration and phase synchronization. IEEE J Sel Areas Commun, 2023, 41: 1620–1634
85 Rogalin R, Bursalioglu O Y, Papadopoulos H, et al. Scalable synchronization and reciprocity calibration for distributed multiuser MIMO. IEEE Trans Wireless Commun, 2014, 13: 1815–1831
86 Hamed E, Rahul H, Abdelghany M A, et al. Real-time distributed MIMO systems. In: Proceedings of the ACM SIGCOMM Conference, 2016. 412–425
87 Magounaki T, Kaltenberger F, Knopp R. Modeling the distributed MU-MIMO OAI 5G testbed and group-based OTA calibration performance evaluation. In: Proceedings of IEEE International Workshop on Signal Processing Advances in Wireless Communications, 2020. 1–5
88 Gharanjik A, Shankar M R B, Arapoglou P D, et al. Robust precoding design for multibeam downlink satellite channel with phase uncertainty. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2015. 3083–3087
89 Hong Z, Xu S, Li T, et al. Robust cascaded team MMSE precoding for cell-free distributed downlink under hierarchical fronthaul. IEEE Trans Wireless Commun, 2024, 23: 14515–14529
90 Wang C, Zhang Y R, Li Q, et al. Satellite computing: a case study of cloud-native satellites. In: Proceedings of IEEE International Conference on Edge Computing and Communications, 2023. 262–270
91 Wang D, Zhang C, Du Y, et al. Implementation of a cloud-based cell-free distributed massive MIMO system. IEEE Commun Mag, 2020, 58: 61–67
92 Wang D M, Zhang C, Ji Z H, et al. Live demonstration: a cloud-based cell-free distributed massive MIMO system. In: Proceedings of IEEE International Symposium on Circuits and Systems, Daegu, 2021
93 Hwang T, Yang C, Wu G, et al. OFDM and its wireless applications: a survey. IEEE Trans Veh Technol, 2009, 58: 1673–1694
94 Falconer D, Ariyavisitakul S L, Benyamin-Seeyar A, et al. Frequency domain equalization for single-carrier broadband wireless systems. IEEE Commun Mag, 2002, 40: 58–66
95 Vakilian V, Wild T, Schaich F, et al. Universal-filtered multi-carrier technique for wireless systems beyond LTE. In: Proceedings of IEEE Global Communications Conference Workshops, Atlanta, 2013. 223–228
96 Zhang X, Jia M, Chen L, et al. Filtered-OFDM-enabler for flexible waveform in the 5th generation cellular networks. In: Proceedings of IEEE Global Communications Conference, San Diego, 2015. 1–6
97 Farhang-Boroujeny B. OFDM versus filter bank multicarrier. IEEE Signal Process Mag, 2011, 28: 92–112
98 Michailow N, Matthe M, Gaspar I S, et al. Generalized frequency division multiplexing for 5th generation cellular networks. IEEE Trans Commun, 2014, 62: 3045–3061
99 Sahin A, Arslan H. Edge windowing for OFDM based systems. IEEE Commun Lett, 2011, 15: 1208–1211
100 Huang M, Chen J, Feng S. Synchronization for OFDM-based satellite communication system. IEEE Trans Veh Technol, 2021, 70: 5693–5702
101 Generation Partnership Project (3GPP). Study on narrow-band Internet of Things (NB-IoT)/enhanced machine type communication (eMTC) support for non-terrestrial networks (NTN) (Release 17). TR 36.763. 2021. https://www.3gpp.org/ftp/Specs/archive/36_series/36.763
102 Hadani R, Rakib S, Tsatsanis M, et al. Orthogonal time frequency space modulation. In: Proceedings of IEEE Wireless Communications and Networking Conference, San Francisco, 2017. 1–6
103 Chu T M C, Zepernick H-J, Hüük A, et al. OTFS modulation for non-terrestrial networks: concepts, applications, benefits, and challenges. In: Proceedings of International Conference on Signal Processing and Communication System, Bydgoszcz, 2023. 1–10
104 Ma Y, Ma G, Ai B, et al. Characteristics of channel spreading function and performance of OTFS in high-speed railway. IEEE Trans Wireless Commun, 2023, 22: 7038–7054
105 Generation Partnership Project (3GPP). OTFS modulation waveform and reference signals for new RAT (Release 17). TSG RAN WG1, R1-162930. 2016. https://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_84/Docs
106 Farhang A, RezazadehReyhani A, Doyle L E, et al. Low complexity modem structure for OFDM-based orthogonal time frequency space modulation. IEEE Wireless Commun Lett, 2018, 7: 344–347
107 Shi J, Li Z, Hu J, et al. OTFS enabled LEO satellite communications: a promising solution to severe Doppler effects. IEEE Netw, 2024, 38: 203–209
108 Hedhly W, Musavian L, Thomos N. OTFS-NOMA system for MIMO communication networks with spatial diversity. In: Proceedings of IEEE International Conference on Communications, Denver, 2024. 569–574
109 Li S, Yuan W, Wei Z, et al. Cross domain iterative detection for orthogonal time frequency space modulation. IEEE Trans Wireless Commun, 2022, 21: 2227–2242
110 Wei Z, Yuan W, Li S, et al. Transmitter and receiver window designs for orthogonal time-frequency space modulation. IEEE Trans Commun, 2021, 69: 2207–2223
111 Qu H, Liu G, Zhang L, et al. Low-complexity symbol detection and interference cancellation for OTFS system. IEEE Trans Commun, 2021, 69: 1524–1537
112 Huang K H, Qiu M, Tong J, et al. Performance of ODDM with imperfect channel estimation. In: Proceedings of IEEE

International Workshop on Signal Processing Advances in Wireless Communications, Shanghai, 2023. 561–565
113 Li Q, Yuan J, Qiu M, et al. Low complexity turbo SIC-MMSE detection for orthogonal time frequency space modulation. IEEE Trans Commun, 2024, 72: 3169–3183
114 Tusha A, Dogan-Tusha S, Yilmaz F, et al. Performance analysis of OTFS under in-phase and quadrature imbalance at transmitter. IEEE Trans Veh Technol, 2021, 70: 11761–11771
115 Tusha A, Dŏgan-Tusha S, Yilmaz F, et al. Physical effect of in-phase and quadrature imbalance in delay-Doppler domain. In: Proceedings of IEEE Vehicular Technology Conference, Norman, 2021. 1–6
116 Neelam S G, Sahu P R. Analysis, estimation and compensation of hardware impairments for CP-OTFS systems. IEEE Wireless Commun Lett, 2022, 11: 952–956
117 Murali K, Chockalingam A. On OTFS modulation for high-Doppler fading channels. In: Proceedings of Information Theory and Applications Workshop, 2018. 1–10
118 Raviteja P, Phan K T, Hong Y. Embedded pilot-aided channel estimation for OTFS in delay-doppler channels. IEEE Trans Veh Technol, 2019, 68: 4906–4917
119 Guo Q, Jiang H, Xiang J, et al. A CS-BEM OTFS channel estimation approach for sparse continuous Doppler-spread channels. IEEE Wireless Commun Lett, 2024, 13: 2985–2989
120 Zhao L, Gao W J, Guo W. Sparse Bayesian learning of delay-Doppler channel for OTFS system. IEEE Commun Lett, 2020, 24: 2766–2769
121 Yuan W, Li S, Wei Z, et al. Data-aided channel estimation for OTFS systems with a superimposed pilot and data transmission scheme. IEEE Wireless Commun Lett, 2021, 10: 1954–1958
122 Wang X, Shen W, Xing C, et al. Joint Bayesian channel estimation and data detection for OTFS systems in LEO satellite communications. IEEE Trans Commun, 2022, 70: 4386–4399
123 Buzzi S, Caire G, Colavolpe G, et al. LEO satellite diversity in 6G non-terrestrial networks: OFDM vs. OTFS. IEEE Commun Lett, 2023, 27: 3013–3017
124 Caus C, Shaat M, Pèrez-Neira A I, et al. Cooperative dual LEO satellite transmission in multi-user OTFS systems. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops, Rhodes Island, 2023. 1–5
125 Zhang Z, Wu Y, Ma Z, et al. Coordinated multi-satellite transmission for OTFS-based 6G LEO satellite communication systems. IEEE J Sel Areas Commun, 2025, 43: 156–170
126 Surabhi G D, Augustine R M, Chockalingam A. On the diversity of uncoded OTFS modulation in doubly-dispersive channels. IEEE Trans Wireless Commun, 2019, 18: 3049–3063
127 Ramachandran M K, Chockalingam A. MIMO-OTFS in high-Doppler fading channels: signal detection and channel estimation. In Proceedings of IEEE Global Communications Conference, Abu Dhabi, 2018. 206–212
128 Loyka S L. Channel capacity of MIMO architecture using the exponential correlation matrix. IEEE Commun Lett, 2001, 5: 369–371
129 Thaj T, Viterbo E. Low-complexity linear diversity-combining detector for MIMO-OTFS. IEEE Wireless Commun Lett, 2022, 11: 288–292
130 Bora A S, Phan K T, Hong Y. Spatially correlated MIMO-OTFS for LEO satellite communication systems. In: Proceedings of IEEE International Conference on Communications Workshops, Seoul, 2022. 723–728
131 Qu H, Liu G, Imran M A, et al. Efficient channel equalization and symbol detection for MIMO OTFS systems. IEEE Trans Wireless Commun, 2022, 21: 6672–6686
132 Bora A S, Phan K T, Hong Y. Mitigating spatial correlation in MIMO-OTFS. IEEE Trans Veh Technol, 2024, 73: 3608–3622
133 Shen B, Wu Y, An J, et al. Random access with massive MIMO-OTFS in LEO satellite communications. IEEE J Sel Areas Commun, 2022, 40: 2865–2881
134 Shen B, Wu Y, Gong S, et al. Massive MIMO-OTFS-based random access for cooperative LEO satellite constellations. IEEE J Sel Areas Commun, 2025, 43: 90–106
135 Gunjan G, Shrivastava S, Kashyap S. Modeling and analysis of physical layer security of OTFS systems under transmit antenna selection and passive eavesdropping. IEEE Commun Lett, 2024, 28: 483–487
136 Hu J, Shi J, Ma S, et al. Secrecy analysis for orthogonal time frequency space scheme based uplink LEO satellite communication. IEEE Wireless Commun Lett, 2021, 10: 1623–1627
137 Niu H C, Hu J F, Shi J, et al. Secrecy performance analysis for OTFS modulation based downlink LEO satellite communication. In: Proceedings of IEEE Global Communications Conference Workshops, Kuala Lumpur, 2023. 2135–2139
138 Mudumbai R, Brown D R, Madhow U, et al. Distributed transmit beamforming: challenges and recent progress. IEEE Commun Mag, 2009, 47: 102–110
139 Nanzer J A, Schmid R L, Comberiate T M, et al. Open-loop coherent distributed arrays. IEEE Trans Microwave Theor Techn, 2017, 65: 1662–1672
140 Nanzer J A, Mghabghab S R, Ellison S M, et al. Distributed phased arrays: challenges and recent advances. IEEE Trans Microwave Theor Techn, 2021, 69: 4893–4907
141 Rashid M, Nanzer J A. High accuracy distributed Kalman filtering for synchronizing frequency and phase in distributed phased arrays. IEEE Signal Process Lett, 2023, 30: 688–692
142 Zhao D, Gu P, Zhong J, et al. Millimeter-wave integrated phased arrays. IEEE Trans Circ Syst I, 2021, 68: 3977–3990
143 Zhao D X, Yu P G, Jiang S, et al. W-band CMOS beamforming ICs and integrated phased-array antennas with 20+ Gb/s data rates. Sci China Inf Sci, 2024, 67: 212301
144 Liu H Q, Zhao D X, Yi Y R, et al. A 24.25–27.5 GHz 128-element dual-polarized 5G integrated phased array with 5.6%-EVM 400-MHz 64-QAM and 50-dBm EIRP. Sci China Inf Sci, 2022, 65: 214301
145 Zhao D X, Chen Z H, You X H. Design and implementation of CMOS millimeter-wave ICs and 4096 TX/4096 RX very-large-scale integrated phased-array antenna (in Chinese). Sci Sin Inform, 2021, 51: 505–519
146 Luo X, Ouyang J, Chen Z H, et al. A scalable Ka-band 1024-element transmit dual-circularly-polarized planar phased array for SATCOM application. IEEE Access, 2020, 8: 156084
147 Fu X, You D, Wang Y, et al. A low-power radiation-hardened Ka-band CMOS phased-array receiver for small satellite constellation. IEEE J Solid-State Circ, 2024, 59: 349–363
148 Zhao D, Gu P, Yi Y, et al. A K-band hybrid-packaged temperature-compensated phased-array receiver and integrated antenna array. IEEE Trans Microwave Theor Techn, 2022, 71: 409–423
149 Mudumbai R, Barriac G, Madhow U. On the feasibility of distributed beamforming in wireless networks. IEEE Trans Wireless Commun, 2007, 6: 1754–1763
150 Coleri S, Ergen M, Puri A, et al. Channel estimation techniques based on pilot arrangement in OFDM systems. IEEE Trans Broadcast, 2002, 48: 223–229
151 Yang M, Zhao D, Xu C, et al. K/Ka-band hybrid-packaged four-element four-beam phased-array transmitter and receiver front-ends with optimized beamforming passive networks. IEEE J Solid-State Circ, 2024, 59: 3142–3155
152 Yeh Y S, Floyd B A. Multibeam phased-arrays using dual-vector distributed beamforming: architecture overview and 28 GHz transceiver prototypes. IEEE Trans Circ Syst I, 2020, 67: 5496–5509
153 Hu Y, Zhan J, Jiang Z H, et al. An orthogonal hybrid analog-digital multibeam antenna array for millimeter-wave massive

MIMO systems. IEEE Trans Antennas Propagat, 2020, 69: 1393–1403
154 He G, Gao X, Sun L, et al. A review of multibeam phased array antennas as LEO satellite constellation ground station. IEEE Access, 2021, 9: 147142
155 Hu Y, Hong W, Yu C, et al. A digital multibeam array with wide scanning angle and enhanced beam gain for millimeter-wave massive MIMO applications. IEEE Trans Antennas Propagat, 2018, 66: 5827–5837
156 Yu Y, Hong W, Jiang Z H, et al. Multibeam generation and measurement of a DDS-based digital beamforming array transmitter at Ka-band. IEEE Trans Antennas Propagat, 2019, 67: 3030–3039
157 Jang S, Jeong J, Lu R, et al. A 16-element 4-beam 1 GHz IF 100 MHz bandwidth interleaved bit stream digital beamformer in 40 nm CMOS. IEEE J Solid-State Circ, 2018, 53: 1302–1312
158 Peng N, Gu P, You X, et al. A Ka-band CMOS 4-beam phased-array receiver with symmetrical beam-distribution network. IEEE Solid-State Circ Lett, 2020, 3: 410–413
159 Talisa S H, O'Haver K W, Comberiate T M, et al. Benefits of digital phased array radars. Proc IEEE, 2016, 104: 530–543
160 Sohrabi F, Yu W. Hybrid digital and analog beamforming design for large-scale antenna arrays. IEEE J Sel Top Signal Process, 2016, 10: 501–513
161 Wan S, Zhu H, Kang K, et al. On the performance of fully-connected and sub-connected hybrid beamforming system. IEEE Trans Veh Technol, 2021, 70: 11078–11082
162 Shannon C E. A mathematical theory of communication. Bell Syst Technical J, 1948, 27: 379–423
163 Yang G H, He G N, Chen R R, et al. Progress and prospect of 6G wireless air-interface transmission technology research (in Chinese). Sci Sin Inform, 2024, 54: 1078–1113
164 Blahut R E. Algebraic Codes for Data Transmission. New York: Cambridge University Press, 2003
165 Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inform Theor, 1967, 13: 260–269
166 Arikan E. Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. IEEE Trans Inform Theor, 2009, 55: 3051–3073
167 Eslami A, Pishro-Nik H. A practical approach to polar codes. In: Proceedings of IEEE International Symposium on Information Theory Proceedings, St. Petersburg, 2011. 16–20
168 Wang R, Liu R. A novel puncturing scheme for polar codes. IEEE Commun Lett, 2014, 18: 2081–2084
169 Generation Partnership Project (3GPP). 5G; NR; Multiplexing and channel coding (Release 15). TS 38.212. 2024. https://www.3gpp.org/ftp/Specs/archive/38_series/38.212
170 Gallager R. Low-density parity-check codes. IEEE Trans Inform Theor, 1962, 8: 21–28
171 Berrou C, Glavieux A, Thitimajshima P. Near Shannon limit error-correcting coding and decoding: turbo-codes. In: Proceedings of IEEE International Conference on Communications, Geneva, 1993. 1064–1070
172 Xu H, Feng D, Luo R, et al. Construction of quasi-cyclic LDPC codes via masking with successive cycle elimination. IEEE Commun Lett, 2016, 20: 2370–2373
173 CCSDS. Synchronization and channel coding. 131.0-B-2-S, 2011. https://public.ccsds.org/Pubs/131x0b2ec1s.pdf
174 Costello D, Dolecek L, Fuja T, et al. Spatially coupled sparse codes on graphs: theory and practice. IEEE Commun Mag, 2014, 52: 168–176
175 Kudekar S, Richardson T J, Urbanke R L. Threshold saturation via spatial coupling: why convolutional LDPC ensembles perform so well over the BEC. IEEE Trans Inform Theor, 2011, 57: 803–834
176 Wijekoon V B, Viterbo E, Hong Y. LDPC-staircase codes for soft decision decoding. In: Proceedings of IEEE Wireless Communications and Networking Conference, Seoul, 2020. 1–6
177 Jimenez Felstrom A, Zigangirov K S. Time-varying periodic convolutional codes with low-density parity-check matrix. IEEE Trans Inform Theor, 1999, 45: 2181–2191
178 Mitchell D G M, Lentmaier M, Costello D J. Spatially coupled LDPC codes constructed from protographs. IEEE Trans Inform Theor, 2015, 61: 4866–4889
179 Du J. A partially coupled LDPC coded scheme for the Gaussian wiretap channel. IEEE Commun Lett, 2020, 24: 7–10
180 Zhang Y S, Jiao J, Wang Y, et al. Soft OSD-sliding window decoding for staircase LDPC codes in deep space communications. In: Proceedings of IEEE/CIC International Conference on Communications in China, Dalian, 2023. 1–6
181 Ordentlich O, Polyanskiy Y. Low complexity schemes for the random access Gaussian channel. In: Proceedings of IEEE International Symposium on Information Theory, Aachen, 2017. 2528–2532
182 Yang T, Yang L, Guo Y J, et al. A non-orthogonal multiple-access scheme using reliable physical-layer network coding and cascade-computation decoding. IEEE Trans Wireless Commun, 2017, 16: 1633–1645
183 Kowshik S S, Andreev K, Frolov A, et al. Energy efficient coded random access for the wireless uplink. IEEE Trans Commun, 2020, 68: 4694–4708
184 Zamir R, Shamai S, Erez U. Nested linear/lattice codes for structured multiterminal binning. IEEE Trans Inform Theor, 2002, 48: 1250–1276
185 Yang T, Yu F T, Chen Q Z, et al. On the design of efficient lattice-code based multiple access. In: Proceedings of IEEE Global Communications Conference, Kuala Lumpur, 2023. 1–7
186 Tal I, Vardy A. List decoding of polar codes. IEEE Trans Inform Theor, 2015, 61: 2213–2226
187 Niu K, Chen K. CRC-aided decoding of polar codes. IEEE Commun Lett, 2012, 16: 1668–1671
188 Miloslavskaya V, Vucetic B. Design of short polar codes for SCL decoding. IEEE Trans Commun, 2020, 68: 6657–6668
189 Fossorier M P C, Mihaljevic M, Imai H. Reduced complexity iterative decoding of low-density parity check codes based on belief propagation. IEEE Trans Commun, 1999, 47: 673–680
190 Cui H, Ghaffari F, Le K, et al. Design of high-performance and area-efficient decoder for 5G LDPC codes. IEEE Trans Circ Syst I, 2021, 68: 879–891
191 Duffy K R, Li J, Médard M. Guessing noise, not code-words. In: Proceedings of IEEE International Symposium on Information Theory, Vail, 2018. 671–675
192 Fossorier M P C, Lin S. Soft-decision decoding of linear block codes based on ordered statistics. IEEE Trans Inform Theor, 1995, 41: 1379–1396
193 Yue C T, Shirvanimoghaddam M, Vucetic B, et al. Ordered-statistics decoding with adaptive Gaussian elimination reduction for short codes. In: Proceedings of IEEE Global Communication Conference Workshops, Rio de Janeiro, 2022. 492–497
194 Liang J, Wang Y, Cai S, et al. A low-complexity ordered statistic decoding of short block codes. IEEE Commun Lett, 2023, 27: 400–403
195 Yue C T, Shirvanimoghaddam M, Li Y, et al. Segmentation discarding ordered-statistic decoding for linear block codes. In: Proceedings of IEEE Global Communication Conference, Waikoloa, 2019. 1–6
196 Yue C, Shirvanimoghaddam M, Park G, et al. Probability-based ordered-statistics decoding for short block codes. IEEE Commun Lett, 2021, 25: 1791–1795
197 Urman Y, Mogilevsky G, Burshtein D. Improving belief propagation list decoding of polar codes by post-processing. In: Proceedings of IEEE International Symposium on Information Theory, Espoo, 2022. 2571–2576
198 Nikopourand H, Baligh H. Sparse code multiple access. In: Proceedings of IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications, London, 2013. 332–336

199 Yang T. Beyond integer-forcing receiver for lattice-code based multi-user MIMO system. IEEE Commun Lett, 2023, 27: 2553–2557
200 Rao Z, Jiao J, Wang Y, et al. Code-domain collision resolution grant-free random access for massive access in IoT. IEEE Trans Wireless Commun, 2023, 22: 4611–4624
201 Gao Z, Ke M, Mei Y, et al. Compressive-sensing-based grant-free massive access for 6G massive communication. IEEE Int Things J, 2024, 11: 7411–7435
202 Wang W, Chen T, Ding R, et al. Location-based timing advance estimation for 5G integrated LEO satellite communications. IEEE Trans Veh Technol, 2021, 70: 6002–6017
203 Yu N Y. Binary Golay spreading sequences and Reed-Muller codes for uplink grant-free NOMA. IEEE Trans Commun, 2021, 69: 276–290
204 Yu N Y, Yu W. Joint activity and data detection for massive grant-free access using deterministic non-orthogonal signatures. IEEE Trans Wireless Commun, 2024, 23: 9474–9487
205 Wang Y, Xu W, Juntti M, et al. Composite preambles based on differential phase rotations for grant-free random access systems. IEEE Int Things J, 2023, 10: 17035–17046
206 Qi T, Lyu B, Hoang D T. Pilot sequences with low coherence and PAPR for grant-free massive access. IEEE Wireless Commun Lett, 2023, 12: 1254–1258
207 Xu L, Jiao J, Wang Y, et al. Low-correlation superimposed pilot grant-free massive access for satellite Internet of Things. IEEE Trans Commun, 2023, 71: 7087–7101
208 Fengler A, Musa O, Jung P, et al. Pilot-based unsourced random access with a massive MIMO receiver, interference cancellation, and power control. IEEE J Sel Areas Commun, 2022, 40: 1522–1534
209 Zhang Z, Li Y, Huang C, et al. User activity detection and channel estimation for grant-free random access in LEO satellite-enabled Internet of Things. IEEE Int Things J, 2020, 7: 8811–8825
210 Zhou X, Ying K, Gao Z, et al. Active terminal identification, channel estimation, and signal detection for grant-free NOMA-OTFS in LEO satellite Internet-of-Things. IEEE Trans Wireless Commun, 2023, 22: 2847–2866
211 Shen B, Wu Y, Zhang W, et al. LEO satellite-enabled random access with large differential delay and Doppler shift. IEEE Trans Wireless Commun, 2025, 24: 2876–2893
212 Zhang C, Liu Y, Hu J, et al. Joint user identification, channel estimation, and data detection for grant-free NOMA in LEO satellite communications. IEEE J Sel Areas Commun, 2025, 43: 107–121
213 Luo Q, Xiao P, Liu Z, et al. AFDM-SCMA: a promising waveform for massive connectivity over high mobility channels. IEEE Trans Wireless Commun, 2024, 23: 14421–14436
214 Le T T T, Hassan N U, Chen X, et al. A survey on random access protocols in direct-access LEO satellite-based IoT communication. IEEE Commun Surv Tut, 2025, 27: 426–462
215 Kaul S, Yates R, Gruteser M. Real-time status: how often should one update? In: Proceedings of International Conference on Computer Communications, Orlando, 2012, 2731–2735
216 Yang T, Jiao J, Wu S, et al. Grant free age-optimal random access protocol for satellite-based Internet of Things. IEEE Trans Commun, 2022, 70: 3947–3961
217 Su S, Jiao J, Yang T, et al. Unequal timeliness protection massive access for mission critical communications in S-IoT. IEEE Trans Commun, 2024, 72: 3211–3226
218 Zhao B, Ren G, Zhang H. Multisatellite cooperative random access scheme in low Earth orbit satellite networks. IEEE Syst J, 2019, 13: 2617–2628
219 Ying K, Gao Z, Chen S, et al. Quasi-synchronous random access for massive MIMO-based LEO satellite constellations. IEEE J Sel Areas Commun, 2023, 41: 1702–1722
220 Ding Z, Lei X, Karagiannidis G K, et al. A survey on non-orthogonal multiple access for 5G networks: research challenges and future trends. IEEE J Sel Areas Commun, 2017, 35: 2181–2195
221 Jiao J, Sun Y, Wu S, et al. Network utility maximization resource allocation for NOMA in satellite-based Internet of Things. IEEE Int Things J, 2020, 7: 3230–3242
222 Dai L, Wang B, Ding Z, et al. A survey of non-orthogonal multiple access for 5G. IEEE Commun Surv Tut, 2018, 20: 2294–2323
223 Ding Z, Zhao Z, Peng M, et al. On the spectral efficiency and security enhancements of NOMA assisted multicast-unicast streaming. IEEE Trans Commun, 2017, 65: 3151–3163
224 Zhang H, Zhang H, Liu W, et al. Energy efficient user clustering, hybrid precoding and power optimization in terahertz MIMO-NOMA systems. IEEE J Sel Areas Commun, 2020, 38: 2074–2085
225 Di B, Song L, Li Y. Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks. IEEE Trans Wireless Commun, 2016, 15: 7686–7698
226 Hong H, Jiao J, Yang T, et al. Age of incorrect information minimization for semantic-empowered NOMA system in S-IoT. IEEE Trans Wireless Commun, 2024, 23: 6639–6652
227 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
228 Perez-Neira A I, Caus M, Vazquez M A. Non-orthogonal transmission techniques for multibeam satellite systems. IEEE Commun Mag, 2019, 57: 58–63
229 Chu J, Chen X. Robust design for integrated satellite-terrestrial Internet of Things. IEEE Int Things J, 2021, 8: 9072–9083
230 Lin Z, Lin M, Wang J B, et al. Joint beamforming and power allocation for satellite-terrestrial integrated networks with non-orthogonal multiple access. IEEE J Sel Top Signal Process, 2019, 13: 657–670
231 Liu X, Lam K Y, Li F, et al. Spectrum sharing for 6G integrated satellite-terrestrial communication networks based on NOMA and CR. IEEE Netw, 2021, 35: 28–34
232 Jiao J, Hong H, Wang Y, et al. Age-optimal downlink NOMA resource allocation for satellite-based IoT network. IEEE Trans Veh Technol, 2023, 72: 11575–11589
233 Zhao M, Yu H, Pan J, et al. Dynamic resource allocation for multi-satellite cooperation networks: a decentralized scheme under statistical CSI. IEEE Access, 2024, 12: 15419–15437
234 Zhao M, Ye N, Ouyang Q, et al. Multi-satellite cooperative communication: exploiting time asynchrony in non-orthogonal transmissions. IEEE Trans Veh Technol, 2023, 72: 6868–6873
235 Huang H, Yang Y, Ding Z, et al. Deep learning-based sum data rate and energy efficiency optimization for MIMO-NOMA systems. IEEE Trans Wireless Commun, 2020, 19: 5373–5388
236 Ding Z. A study on the optimality of downlink hybrid NOMA. IEEE Signal Process Lett, 2025, 32: 511–515
237 Katwe M, Singh K, Li C P, et al. Spectral-efficient downlink systems under imperfect SIC and CSI: MC-NOMA or partial NOMA? IEEE Wireless Commun Lett, 2024, 13: 133–137
238 Shi Z, Lu H, Xie X, et al. Active RIS-aided EH-NOMA networks: a deep reinforcement learning approach. IEEE Trans Commun, 2023, 71: 5846–5861
239 Deshpande R, Katwe M V, Singh K, et al. Resource allocation design for spectral-efficient URLLC using RIS-aided FD-NOMA system. IEEE Wireless Commun Lett, 2023, 12: 1209–1213
240 Mao Y, Dizdar O, Clerckx B, et al. Rate-splitting multiple access: fundamentals, survey, and future research trends. IEEE

Commun Surv Tut, 2022, 24: 2073–2126
241 Mishra A, Mao Y, Dizdar O, et al. Rate-splitting multiple access for downlink multiuser MIMO: precoder optimization and PHY-layer design. IEEE Trans Commun, 2021, 70: 874–890
242 Dizdar O, Mao Y J, Han W, et al. Rate-splitting multiple access for downlink multi-antenna communications: physical layer design and link-level simulations. In: Proceedings of IEEE Annual International Symposium on Personal, Indoor and Mobile Radio Communications, London, 2020. 1–6
243 de Sena A S, Nardelli P H J, da Costa D B, et al. Rate-splitting multiple access and its interplay with intelligent reflecting surfaces. IEEE Commun Mag, 2022, 60: 52–57
244 Yin L F, Clerckx B, Mao Y J. Rate-splitting multiple access for multi-antenna broadcast channels with statistical CSIT. In: Proceedings of IEEE Wireless Communications and Networking Conference Workshops, Nanjing, 2021. 1–6
245 Zhuo B, Gu J, Duan W, et al. RIS-IoE for data-driven networks: new mentalities, trends and preliminary solutions. IEEE Int Things M, 2023, 6: 102–107
246 Huang J, Yang Y, Lee J, et al. Deep reinforcement learning-based resource allocation for RSMA in LEO satellite-terrestrial networks. IEEE Trans Commun, 2024, 72: 1341–1354
247 Xu Y, Yin L, Mao Y, et al. Distributed rate-splitting multiple access for multilayer satellite communications. IEEE Trans Commun, 2024, 72: 6131–6144
248 Vazquez M, Caus M, Perez-Neira A. Rate splitting for MIMO multibeam satellite systems. In: Proceedings of International ITG Workshop on Smart Antennas, Bochum, 2018. 1–6
249 Si Z W, Yin L, Clerckx B. Rate-splitting multiple access for multigateway multibeam satellite systems with feeder link interference. IEEE Trans Commun, 2022, 70: 2147–2162
250 Liu J, Shi Y, Fadlullah Z M, et al. Space-air-ground integrated network: a survey. IEEE Commun Surv Tut, 2018, 20: 2714–2741
251 Leung K C, Li V. Transmission control protocol (TCP) in wireless networks: issues, approaches, and challenges. IEEE Commun Surv Tut, 2006, 8: 64–79
252 CCSDS. Overview of space communications protocols, report concerning space data system standards. 130.0-G-2, 2007. https://public.ccsds.org/Pubs/130x0g2s.pdf
253 Fall K. A delay-tolerant network architecture for challenged Internets. In: Proceedings of ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Karlsruhe, 2003. 27–34
254 Burleigh S, Ramadas M, Farrell S. Licklider transmission protocol-specification. IETF RFC 5325, 2008. https://datatracker.ietf.org/doc/html/rfc5325
255 Shi L, Jiao J, Sabbagh A, et al. Integration of Reed-Solomon codes to licklider transmission protocol (LTP) for space DTN. IEEE Aerosp Electron Syst Mag, 2017, 32: 48–55
256 Hasegawa Y, Ito T, Ono Y, et al. A throughput model of TCP-FSO/ADFR for free-space optical satellite communications. In: Proceedings of IEEE Global Communications Conference, Waikoloa, 2019. 1–6
257 Le H D, Mai V V, Nguyen C T, et al. Design and analysis of sliding window ARQ protocols with rate adaptation for burst transmission over FSO turbulence channels. J Opt Commun Netw, 2019, 11: 151–163
258 Rosas F, Souza R D, Pellenz M E, et al. Optimizing the code rate of energy-constrained wireless communications with HARQ. IEEE Trans Wireless Commun, 2016, 15: 191–205
259 Lee W, Simeone O, Kang J, et al. HARQ buffer management: an information-theoretic view. IEEE Trans Commun, 2015, 63: 4539–4550
260 Ahmed A, Al-Dweik A, Iraqi Y, et al. Hybrid automatic repeat request (HARQ) in wireless communications systems and standards: a contemporary survey. IEEE Commun Surv Tut, 2021, 23: 2711–2752
261 Makki B, Svensson T, Caire G, et al. Fast HARQ over finite blocklength codes: a technique for low-latency reliable communication. IEEE Trans Wireless Commun, 2019, 18: 194–209
262 Shirvanimoghaddam M, Khayami H, Li Y. Dynamic HARQ with guaranteed delay. In: Proceedings of IEEE Wireless Communications and Networking Conference, Seoul, 2020. 1–6
263 Berardinelli G, Khosravirad S R, Pedersen K I, et al. Enabling early HARQ feedback in 5G networks. In: Proceedings of IEEE Vehicular Technology Conference, Nanjing, 2016. 1–5
264 Strodthoff N, Goktepe B, Schierl T, et al. Enhanced machine learning techniques for early HARQ feedback prediction in 5G. IEEE J Sel Areas Commun, 2019, 37: 2573–2587
265 CCSDS. Erasure correcting codes for use in near-earth and deep-space communications. 131.5-O-1, 2014. https://ccsds.org/Pubs/131x5o1c1.pdf
266 Shirvanimoghaddam M, Mohammadi M S, Abbas R, et al. Short block-length codes for ultra-reliable low latency communications. IEEE Commun Mag, 2019, 57: 130–137
267 Luby M. LT codes. In: Proceedings of IEEE Symposium on Foundations of Computer Science, Vancouver, 2002. 271–280
268 Shokrollahi A. Raptor codes. IEEE Trans Inform Theor, 2006, 52: 2551–2567
269 Generation Partnership Project (3GPP). Technical specification group services and system aspects; Multimedia broadcast/multicast service (MBMS); Protocols and codecs (Release 6). TS 26.346. 2024. https://www.3gpp.org/ftp/Specs/archive/26_series/26.346
270 Sørensen J H, Koike-Akino T, Orlik P. Rateless feedback codes. In: Proceedings of IEEE International Symposium on Information Theory, Cambridge, 2012. 1767–1771
271 Jiao J, Nie S X, Yang Y, et al. Distributed systematic Raptor coding scheme in deep space communications (in Chinese). J Astronaut, 2016, 37: 1232–1238
272 Ho T, Medard M, Koetter R, et al. A random linear network coding approach to multicast. IEEE Trans Inform Theor, 2006, 52: 4413–4430
273 Tsimbalo E, Tassi A, Piechocki R J. Reliability of multicast under random linear network coding. IEEE Trans Commun, 2018, 66: 2547–2559
274 Yang S, Yeung R W. Batched sparse codes. IEEE Trans Inform Theor, 2014, 60: 5322–5346
275 Jiao J, Ni Z, Wu S, et al. Energy efficient network coding HARQ transmission scheme for S-IoT. IEEE Trans Green Commun Netw, 2021, 5: 308–321
276 Jiao J, Liu S, Ding J, et al. Age-optimal network coding HARQ transmission scheme for dual-hop satellite-integrated Internet. IEEE Trans Veh Technol, 2022, 71: 10666–10682
277 Ding J, Jiao J, Huang J, et al. Age-optimal network coding HARQ scheme for satellite-based Internet of Things. IEEE Int Things J, 2022, 9: 21984–21998
278 Huang J, Jiao J, Wang Y, et al. Age-critical long erasure coding-CCSDS file delivery protocol for dual-hop S-IoT. IEEE Int Things J, 2023, 10: 17070–17084
279 Li D, Wu S, Jiao J, et al. Towards age-optimal transmission in satellite-integrated IoT: a two-layer coding approach. IEEE Trans Veh Technol, 2023, 72: 1137–1148
280 Le H D, Trinh P V, Pham T V, et al. Throughput analysis for TCP over the FSO-based satellite-assisted Internet of Vehicles. IEEE Trans Veh Technol, 2022, 71: 1875–1890
281 Hu C B, Yang H J, Li B, et al. A high-throughput cooperative network coding HARQ transmission scheme for integrated

satellite-terrestrial networks. In: Proceedings of IEEE Vehicular Technology Conference, Hong Kong, 2023. 1–5
282 Cai D, Ding Z, Fan P, et al. On the performance of NOMA with hybrid ARQ. IEEE Trans Veh Technol, 2018, 67: 10033–10038
283 Shi Z, Zhang C, Fu Y, et al. Achievable diversity order of HARQ-aided downlink NOMA systems. IEEE Trans Veh Technol, 2020, 69: 471–487
284 Ghanami F, Hodtani G A, Vucetic B, et al. Performance analysis and optimization of NOMA with HARQ for short packet communications in massive IoT. IEEE Int Things J, 2021, 8: 4736–4748
285 Marasinghe D, Rajatheva N, Latva-Aho M. Block error performance of NOMA with HARQ-CC in finite blocklength. In: Proceedings of IEEE International Conference Communications Workshops, Dublin, 2020, 1-6
286 Wu S, Deng Z, Li A, et al. Minimizing age-of-information in HARQ-CC aided NOMA systems. IEEE Trans Wireless Commun, 2023, 22: 1072–1086
287 Liu K P, Li A M, Wu S H. Deep reinforcement learning-assisted age-optimal transmission policy for HARQ-aided NOMA networks. In: Proceedings of IEEE Conference on Computer Communications Workshops, Hoboken, 2023. 1–6
288 Samy R, Yang H C, Rakia T, et al. Space-air-ground FSO networks for high-throughput satellite communications. IEEE Commun Mag, 2022, 60: 82–87
289 Al-Hraishawi H, Chougrani H, Kisseleff S, et al. A survey on nongeostationary satellite systems: the communication perspective. IEEE Commun Surv Tut, 2023, 25: 101–132
290 Le H D, Pham A T. Link-layer retransmission-based error-control protocols in FSO communications: a survey. IEEE Commun Surv Tut, 2022, 24: 1602–1633
291 Feng W, Wang Y, Chen Y, et al. Structured satellite-UAV-terrestrial networks for 6G Internet of Things. IEEE Netw, 2024, 38: 48–54
292 Arum S C, Grace D, Mitchell P D. A review of wireless communication using high-altitude platforms for extended coverage and capacity. Comput Commun, 2020, 157: 232–256
293 Kaymak Y, Rojas-Cessa R, Feng J, et al. A survey on acquisition, tracking, and pointing mechanisms for mobile free-space optical communications. IEEE Commun Surv Tut, 2018, 20: 1104–1123
294 Moon H J, Chae C B, Wong K K, et al. Pointing-and-acquisition for optical wireless in 6G: from algorithms to performance evaluation. IEEE Commun Mag, 2024, 62: 32–38
295 Velazco J E, Wernicke D, Griffin J, et al. Inter-spacecraft omnidirectional optical communicator for swarms. In: Proceedings of Annual AIAA/USU Conference, 2019
296 Sundararaman B, Buy U, Kshemkalyani A D. Clock synchronization for wireless sensor networks: a survey. Ad Hoc Netws, 2005, 3: 281–323
297 Jeong S, Farhang A, Flanagan M. Collaborative vs. non-collaborative CFO estimation for distributed large-scale MIMO systems. In: Proceedings of IEEE Vehicular Technology Conference, Victoria, 2020. 1–6
298 Wu H, Sun Z, Zhou X. Deep learning-based frame and timing synchronization for end-to-end communications. J Phys-Conf Ser, 2019, 1169: 012060
299 Wang Y Y, Zhang C, Peng Q, et al. Learning to detect frame synchronization. In: Proceedings of International Conference on Neural Information Processing, Daegu, 2013. 570–578
300 Qi X, Zhang B, Qiu Z, et al. Using inter-mesh links to reduce end-to-end delay in walker delta constellations. IEEE Commun Lett, 2021, 25: 3070–3074
301 Dong Y Y, Xu X F, Zhang Y Y, et al. A novel virtual node-based multi-controller management architecture for LEO mega-constellation satellite networks. In: Proceedings of IEEE International Conference on Wireless Communications and Signal Processing, Hangzhou, 2023. 701–706
302 Hu M, Li J, Cai C, et al. Software defined multicast for large-scale multi-layer LEO satellite networks. IEEE Trans Netw Serv Manage, 2022, 19: 2119–2130
303 Jiang W. Software defined satellite networks: a survey. Digital Commun Netws, 2023, 9: 1243–1264
304 Chen Q, Guo J, Yang L, et al. Topology virtualization and dynamics shielding method for LEO satellite networks. IEEE Commun Lett, 2019, 24: 433–437
305 Wang R, Kishk M A, Alouini M S. Reliability analysis of multi-hop routing in multi-tier LEO satellite networks. IEEE Trans Wireless Commun, 2024, 23: 1959–1973
306 Hu M, Yang R, Hu Y, et al. QoS-aware software-defined multicast in LEO satellite networks. IEEE Trans Aerosp Electron Syst, 2022, 58: 5307–5317
307 Roth M M, Brandt H, Bischl H. Distributed SDN-based load-balanced routing for low Earth orbit satellite constellation networks. In: Proceedings of IEEE Advanced Satellite Multimedia Systems Conference and Signal Processing for Space Communications Workshop, Graz, 2022. 1–8
308 Deng X, Chang L, Zeng S, et al. Distance-based back-pressure routing for load-balancing LEO satellite networks. IEEE Trans Veh Technol, 2022, 72: 1240–1253
309 Kumar P, Bhushan S, Halder D, et al. fybrrLink: efficient QoS-aware routing in SDN enabled future satellite networks. IEEE Trans Netw Serv Manage, 2021, 19: 2107–2118
310 Soret B, Leyva-Mayorga I, Lozano-Cuadra F, et al. Q-learning for distributed routing in LEO satellite constellations. In: Proceedings of IEEE International Conference on Machine Learning for Communication and Networking, Stockholm, 2024. 208–213
311 Pachler N, Crawley E F, Cameron B G. Robust beam-to-satellite routing strategies for megaconstellations. IEEE Wireless Commun Lett, 2024, 13: 3040–3043
312 Ji S, Zhou D, Sheng M, et al. Mega satellite constellation system optimization: from a network control structure perspective. IEEE Trans Wireless Commun, 2021, 21: 913–927
313 Wei D X, Jin C, Low S H, et al. FAST TCP: motivation, architecture, algorithms, performance. IEEE ACM Trans Netwing, 2006, 14: 1246–1259
314 Ahmad S, Arshad M J. Enhancing fast TCP's performance using single TCP connection for parallel traffic flows to prevent head-of-line blocking. IEEE Access, 2019, 7: 148152
315 Deutschmann J, Hielscher K-S, German R. CUBIC local loss recovery vs. BBR on (satellite) Internet paths. In: Proceedings of IEEE International Symposium on Local and Metropolitan Area Networks, London, 2023. 1–3
316 Claypool S, Chung J, Claypool M. Comparison of TCP congestion control performance over a satellite network. In: Proceedings of International Conference on Passive and Active Network Measurement, 2021. 499–512
317 Liu M X, Liu Y C, Ma Z F, et al. The effects of a performance enhancing proxy on TCP congestion control over a satellite network. In: Proceedings of IEEE International Performance, Computing, and Communications Conference, Austin, 2022. 325–331
318 Shreedhar T, Kaul S K, Yates R D. An empirical study of ageing in the cloud. In: Proceedings of IEEE Conference on Computer Communications Workshops, Vancouver, 2021. 1–6
319 Guloglu U, Baghaee S, Uysal E. Evaluation of age control protocol (ACP) and ACP+ on ESP32. In: Proceedings of IEEE International Symposium on Wireless Communication Systems, Berlin, 2021. 1–6
320 Chiu D M, Jain R. Analysis of the increase and decrease algorithms for congestion avoidance in computer networks. Comput Netws ISDN Syst, 1989, 17: 1–14

321 Cao X Y, Zhang X Y. SaTCP: link-layer informed TCP adaptation for highly dynamic LEO satellite networks. In: Proceedings of IEEE Conference on Computer Communications, New York, 2023. 1–10
322 Ma T, Qian B, Qin X, et al. Satellite-terrestrial integrated 6G: an ultra-dense LEO networking management architecture. IEEE Wireless Commun, 2024, 31: 62–69
323 Sun Y, Peng M, Zhang S, et al. Integrated satellite-terrestrial networks: architectures, key techniques, and experimental progress. IEEE Netw, 2022, 36: 191–198
324 Generation Partnership Project (3GPP). Solutions for NR to support non-terrestrial networks (NTN) (Release 16). TR 38.821. https://www.3gpp.org/ftp/Specs/archive/38_series/38.821
325 Portillo I, Cameron B, Crawley C. Ground segment architectures for large LEO constellations with feeder links in EHF-bands. In: Proceedings of IEEE Aerospace Conference, Big Sky, 2018. 1-14
326 Papa A, de Cola T, Vizarreta P, et al. Design and evaluation of reconfigurable SDN LEO constellations. IEEE Trans Netw Serv Manage, 2020, 17: 1432–1445
327 Du P, Bai W, Li J, et al. Dynamic hierarchical VAP-based location management for mega satellite networks. IEEE Int Things J, 2024, 11: 19749–19761
328 Generation Partnership Project (3GPP). 5G; System architecture for the 5G system (5GS) (Release 16). TS 23.501. 2024. https://www.3gpp.org/ftp/Specs/archive/23_series/23.501
329 Ji S, Zhou D, Sheng M, et al. Dynamic space-ground integrated mobility management strategy for mega LEO satellite constellations. IEEE Trans Wireless Commun, 2024, 23: 11043–11060
330 Ji S, Sheng M, Zhou D, et al. Flexible and distributed mobility management for integrated terrestrial-satellite networks: challenges, architectures, and approaches. IEEE Netw, 2021, 35: 73–81
331 Chowdhury P, Atiquzzaman M, Ivancic W. Handover schemes in satellite networks: state-of-the-art and future research directions. IEEE Commun Surv Tut, 2006, 8: 2–14
332 Del Re E, Fantacci R, Giambene G. Efficient dynamic channel allocation techniques with handover queuing for mobile satellite networks. IEEE J Sel Areas Commun, 1995, 13: 397–405
333 Maral G, Restrepo J, del Re E, et al. Performance analysis for a guaranteed handover service in an LEO constellation with a "satellite-fixed cell" system. IEEE Trans Veh Technol, 1998, 47: 1200–1214
334 Wu Z, Jin F, Luo J, et al. A graph-based satellite handover framework for LEO satellite communication networks. IEEE Commun Lett, 2016, 20: 1547–1550
335 Feng L, Liu Y, Wu L, et al. A satellite handover strategy based on MIMO technology in LEO satellite networks. IEEE Commun Lett, 2020, 24: 1505–1509
336 Zhang S, Liu A, Han C, et al. A network-flows-based satellite handover strategy for LEO satellite networks. IEEE Wireless Commun Lett, 2021, 10: 2669–2673
337 Xu H, Li D, Liu M, et al. QoE-driven intelligent handover for user-centric mobile satellite networks. IEEE Trans Veh Technol, 2020, 69: 10127–10139
338 He S, Wang T, Wang S. Load-aware satellite handover strategy based on multi-agent reinforcement learning. In: Proceedings of IEEE Global Communications Conference, Taipei, 2020. 1–6
339 Cao Y, Lien S Y, Liang Y C. Deep reinforcement learning for multi-user access control in non-terrestrial networks. IEEE Trans Commun, 2021, 69: 1605–1619
340 Yang F, Wu W, Gao Y, et al. Multi-agent fingerprints-enhanced distributed intelligent handover algorithm in LEO satellite networks. IEEE Trans Veh Technol, 2024, 73: 15255–15269
341 Liu H, Wang Y, Li P, et al. A multi-agent deep reinforcement learning-based handover scheme for mega-constellation under dynamic propagation conditions. IEEE Trans Wireless Commun, 2024, 23: 13579–13596
342 Pacheco-Paramo D, Akyildiz I F, Casares-Giner V. Local anchor based location management schemes for small cells in HetNets. IEEE Trans Mobile Comput, 2016, 15: 883–894
343 Deng T, Wang X, Fan P, et al. Modeling and performance analysis of a tracking-area-list-based location management scheme in LTE networks. IEEE Trans Veh Technol, 2016, 65: 6417–6431
344 Johnson D, Perkins C, Arkko J. Mobility support in IPv6. IETF, RFC 3775, 2004. https://www.rfc-editor.org/pdfrfc/rfc6275.txt.pdf
345 Darwish T, Kurt G K, Yanikomeroglu H, et al. Location management in internet protocol-based future LEO satellite networks: a review. IEEE Open J Commun Soc, 2022, 3: 1035–1062
346 Shahriar A Z M, Atiquzzaman M, Rahman S. Mobility management protocols for next-generation all-IP satellite networks. IEEE Wireless Commun, 2008, 15: 46–54
347 Du P, Li J, Bai W, et al. Dual location area based distributed location management for hybrid LEO/MEO mega satellite networks. IEEE Trans Veh Technol, 2023, 72: 2307–2321
348 Li D G, Li H Y, Zhang S, et al. Virtual agent clustering based mobility management over the satellite networks. In: Proceedings of IEEE International Conference on Wireless Communications and Signal Processing, Hangzhou, 2018. 1–5
349 Jeon S, Figueiredo S, Aguiar R L, et al. Distributed mobility management for the future mobile networks: a comprehensive analysis of key design options. IEEE Access, 2017, 5: 11423–11436
350 Zhang X, Shi K, Zhang S, et al. Virtual agent clustering based mobility management over the satellite networks. IEEE Access, 2019, 7: 89544–89555
351 Xu X B, Zhao H, Liu C, et al. On the aggregated resource management for satellite edge computing. In: Proceedings of IEEE International Conference on Communications, Montreal, 2021. 1–6
352 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
353 Deng D, Wang C, Pang M, et al. Dynamic resource allocation with deep reinforcement learning in multibeam satellite communication. IEEE Wireless Commun Lett, 2022, 12: 75–79
354 Cao X, Yang B, Shen Y, et al. Edge-assisted multi-layer offloading optimization of LEO satellite-terrestrial integrated networks. IEEE J Sel Areas Commun, 2022, 41: 381–398
355 He H, Zhou D, Sheng M, et al. Hierarchical cross-domain satellite resource management: an intelligent collaboration perspective. IEEE Trans Commun, 2023, 71: 2201–2215
356 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
357 Guo B, Chang Z, Han Z, et al. Network slicing strategy for real-time applications in large-scale satellite networks with heterogeneous transceivers. IEEE Wireless Commun Lett, 2024, 13: 2195–2199
358 He M C, Wu H Q, Zhou C H, et al. Digital twin-assisted robust and adaptive resource slicing in LEO satellite networks. 2024. arXiv:2411.03635
359 Feng W, Lin Y, Wang Y, et al. Radio map-based cognitive satellite-UAV networks towards 6G on-demand coverage. IEEE Trans Cogn Commun Netw, 2024, 10: 1075–1089
360 Diffie W, Hellman M. New directions in cryptography. IEEE Trans Inform Theor, 1976, 22: 644–654
361 Ingemarsson I, Wong C. Encryption and authentication in on-board processing satellite communication systems. IEEE Trans Commun, 1981, 29: 1684–1687

362 Liu Y, Chen H H, Wang L. Physical layer security for next generation wireless networks: theories, technologies, and challenges. IEEE Commun Surv Tut, 2017, 19: 347–376
363 Han S, Li J, Meng W, et al. Challenges of physical layer security in a satellite-terrestrial network. IEEE Netw, 2022, 36: 98–104
364 Zeng K. Physical layer key generation in wireless networks: challenges and opportunities. IEEE Commun Mag, 2015, 53: 33–39
365 Chen X, An J, Xiong Z, et al. Covert communications: a comprehensive survey. IEEE Commun Surv Tut, 2023, 25: 1173–1198
366 Yu H, Yu J, Liu J, et al. Covert satellite communication over overt channel: a randomized Gaussian signalling approach. IEEE Trans Aerosp Electron Syst, 2025, 61: 2355–2368
367 Feng S, Lu X, Sun S, et al. Covert communication in large-scale multi-tier LEO satellite networks. IEEE Trans Mobile Comput, 2024, 23: 11576–11587
368 Pan D, Long G L, Yin L, et al. The evolution of quantum secure direct communication: on the road to the Qinternet. IEEE Commun Surv Tut, 2024, 26: 1898–1949
369 Yin H L, Chen T Y, Yu Z W, et al. Measurement-device-independent quantum key distribution over a 404 km optical fiber. Phys Rev Lett, 2016, 117: 190501
370 Liao S K, Cai W Q, Liu W Y, et al. Satellite-to-ground quantum key distribution. Nature, 2017, 549: 43–47
371 Quan W, Liu Y, Zhang H, et al. Enhancing crowd collaborations for software defined vehicular networks. IEEE Commun Mag, 2017, 55: 80–86
372 European Commission. SAT-5G Project. CORDIS, 2020. https://www.sat5g-project.eu/
373 Cheng N, Quan W, Shi W, et al. A comprehensive simulation platform for space-air-ground integrated network. IEEE Wireless Commun, 2020, 27: 178–185
374 Erl S, de Cola T. Deep learning for joint source-channel coding of text. In: Proceedings of Advanced Satellite Multimedia Systems Conference, Livorno, 2014. 382–389
375 Afhamisis M, Barillaro S, Palattella M R. A testbed for LoRaWAN satellite backhaul: design principles and validation. In: Proceedings of IEEE International Conference on Communications Workshops, Seoul, 2022. 1171–1176
376 Spiridonov A A, Saetchnik V A, Ushakov D V, et al. Small satellite orbit determination using single pass Doppler measurements. IEEE J Miniat Air Space Syst, 2022, 3: 162–170
377 Tiwari A K, Mehto R, Tiwari S, et al. 3D satellite visualization using SGP4. In: Proceedings of IEEE International Conference on Advances in Electronics, Computers and Communications, Bengaluru, 2023. 1–5
378 Storek K-U, Schwarz R T, Knopp A. Multi-satellite multi-user MIMO precoding: testbed and field trial. In: Proceedings of IEEE International Conference on Communications, Dublin, 2020. 1–7
379 Minardi M, Drif Y, Vu T X, et al. SDN-based testbed for emerging use cases in beyond 5G NTN-terrestrial networks. In: Proceedings of IEEE Network Operations and Management Symposium, Miami, 2023. 1–6
380 Qiao Y, Teng S, Luo J, et al. On-orbit DNN distributed inference for remote sensing images in satellite Internet of Things. IEEE Int Things J, 2025, 12: 5687–5703
381 Minardi M, Vu T X, Lei L, et al. Virtual network embedding for NGSO systems: algorithmic solution and SDN-testbed validation. IEEE Trans Netw Serv Manage, 2022, 20: 3523–3535
382 Wang C, An W Y, Li X. On-board hybrid heterogeneous distributed computing resource virtualization. In: Proceedings of International Conference on Optical Communications and Networks, Qufu, 2023. 1–3
383 Syamala M, Kaliappan S, M P H, et al. Performance analysis of lightweight virtualization for environments with edge computing based on NFV. In: Proceedings of International Conference on Smart Electronics and Communication, Trichy, 2023. 687–691
384 Xu X D, Tao X F, Wu C L, et al. Interference analysis of OFDMA based distributed network architecture and resource pooling (in Chinese). J Beijing Univ Post Telecommun, 2007, 30: 19–22
385 Xu X D, Wang D, Tao X F, et al. Resource pooling for frameless network architecture with adaptive resource allocation. Sci China Inf Sci, 2013, 56: 022314
386 Zhu X, Jiang C, Kuang L, et al. Two-layer game based resource allocation in cloud based integrated terrestrial-satellite networks. IEEE Trans Cogn Commun Netw, 2020, 6: 509–522
387 Qin J, Guo X, Ma X, et al. Application and performance evaluation of resource pool architecture in satellite edge computing. Aerospace, 2022, 9: 451
388 Shen X M, Cheng N, Zhou H B, et al. Space-air-ground integrated networks: review and prospect (in Chinese). Chinese J Int Things, 2020, 4: 3–19
389 Papa A, de Cola T, Vizarreta P, et al. Dynamic SDN controller placement in a LEO constellation satellite network. In: Proceedings of IEEE Global Communications Conference, Montreal, 2018. 206–212
390 Ahmed T, Alleg A, Ferrus R, er al. On-demand network slicing using SDN/NFV-enabled satellite ground segment systems. In: Proceedings of IEEE Conference on Network Softwarization and Workshops, Abu Dhabi, 2018. 242–246
391 Ku H-J, Jung J-H, Kwon G-I. A study on reinforcement learning based SFC path selection in SDN/NFV. Int J Appl Eng Res, 2017, 12: 3439–3443
392 Pachler N, del Portillo I, Crawley E F, et al. An updated comparison of four low Earth orbit satellite constellation systems to provide global broadband. In: Proceedings of IEEE International Conference on Communications Workshops, Montreal, 2021. 1–7
393 Chen Q, Giambene G, Yang L, et al. Analysis of inter-satellite link paths for LEO mega-constellation networks. IEEE Trans Veh Technol, 2021, 70: 2743–2755
394 McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of Artificial Intelligence and Statistics, Ft. Lauderdale, 2017. 1273–1282
395 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
396 Li T, Sahu A K, Zaheer M, et al. Federated optimization in heterogeneous networks. In: Proceedings of Machine Learning and Systems, Austin, 2020. 429–450
397 Wang J Y, Liu Q H, Liang H, et al. Tackling the objective inconsistency problem in heterogeneous federated optimization. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2020. 7611–7623
398 Wu C R, Zhu Y F, Wang F X. DSFL: decentralized satellite federated learning for energy-aware LEO constellation computing. In: Proceedings of IEEE International Conference on Satellite Computing, Shenzhen, 2022. 25–30
399 Razmi N, Matthiesen B, Dekorsy A, et al. Ground-assisted federated learning in LEO satellite constellations. IEEE Wireless Commun Lett, 2022, 11: 717–721
400 Razmi N, Matthiesen B, Dekorsy A, et al. On-board federated learning for dense LEO constellations. In: Proceedings of IEEE International Conference on Communications, Seoul, 2022. 4715–4720
401 So J H, Hsieh K, Arzani B, et al. Fedspace: an efficient federated learning framework at satellites and ground stations. 2022. ArXiv:2202.01267
402 Elmahallawy M, Luo T. AsyncFLEO: asynchronous federated learning for LEO satellite constellations with high-altitude

platforms. In Proceedings of IEEE International Conference on Big Data, Osaka, 2022. 5478–5487
403 Wu L L, Zhang J J. FedGSM: efficient federated learning for LEO constellations with gradient staleness mitigation. In Proceedings of IEEE International Workshop on Signal Processing Advances in Wireless Communications, Shanghai, 2023. 356–360
404 Nguyen J, Malik K, Zhan H Y, et al. Federated learning with buffered asynchronous aggregation. 2021. ArXiv:2106.06639v4
405 Rodrigues T K, Kato N. Hybrid centralized and distributed learning for MEC-equipped satellite 6G networks. IEEE J Sel Areas Commun, 2023, 41: 1201–1211
406 Uysal E, Kaya O, Ephremides A, et al. Semantic communications in networked systems: a data significance perspective. IEEE Netw, 2022, 36: 233–240
407 Gunduz D, Qin Z, Aguerri I E, et al. Beyond transmitting bits: context, semantics, and task-oriented communications. IEEE J Sel Areas Commun, 2023, 41: 5–41
408 Shi G, Xiao Y, Li Y, et al. From semantic communication to semantic-aware networking: model, architecture, and open problems. IEEE Commun Mag, 2021, 59: 44–50
409 Peng H, Zhang Z, Liu Y, et al. Semantic communication in non-terrestrial networks: a future-ready paradigm. IEEE Netw, 2024, 38: 119–127
410 Deng D, Wang C, Xu L, et al. Semantic communication empowered NTN for IoT: benefits and challenges. IEEE Netw, 2024, 38: 32–39
411 Shannon C, Weaver W. Recent contributions to the mathematical theory of communication. ETC, 1953, 10: 261–281
412 Yang W, Du H, Liew Z Q, et al. Semantic communications for future Internet: fundamentals, applications, and challenges. IEEE Commun Surv Tut, 2023, 25: 213–250
413 Strinati E C, Barbarossa S. 6G networks: beyond Shannon towards semantic and goal-oriented communications. Comput Netws, 2021, 190: 107930
414 Qin Z, Ye H, Li G Y, et al. Deep learning in physical layer communications. IEEE Wireless Commun, 2019, 26: 93–99
415 Farsad N, Rao M, Goldsmith A. Deep learning for joint source-channel coding of text. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, 2018. 2326–2330
416 Bourtsoulatze E, Burth Kurka D, Gunduz D. Deep joint source-channel coding for wireless image transmission. IEEE Trans Cogn Commun Netw, 2019, 5: 567–579
417 Weng Z, Qin Z. Semantic communication systems for speech transmission. IEEE J Sel Areas Commun, 2021, 39: 2434–2444
418 Jankowski M, Gunduz D, Mikolajczyk K. Wireless image retrieval at the edge. IEEE J Sel Areas Commun, 2021, 39: 89–100
419 Weng Z, Qin Z, Tao X, et al. Deep learning enabled semantic communications with speech recognition and synthesis. IEEE Trans Wireless Commun, 2023, 22: 6227–6240
420 Xie H, Qin Z, Tao X, et al. Task-oriented multi-user semantic communications. IEEE J Sel Areas Commun, 2022, 40: 2584–2597
421 Xie H, Qin Z, Li G Y. Semantic communication with memory. IEEE J Sel Areas Commun, 2023, 41: 2658–2669
422 Yang Y, Guo C L, Liu F F, et al. Semantic communications with AI tasks. 2021. ArXiv:2109.14170
423 Wang M, Wang H, Qi G, et al. Richpedia: a large-scale, comprehensive multi-modal knowledge graph. Big Data Res, 2020, 22: 100159
424 Kountouris M, Pappas N. Semantics-empowered communication for networked intelligent systems. IEEE Commun Mag, 2021, 59: 96–102
425 Luo X, Chen H H, Guo Q. Semantic communications: overview, open issues, and future research directions. IEEE Wireless Commun, 2022, 29: 210–219
426 Yates R D, Sun Y, Brown D R, et al. Age of information: an introduction and survey. IEEE J Sel Areas Commun, 2021, 39: 1183–1210
427 Kosta A, Pappas N, Ephremides A, et al. The cost of delay in status updates and their value: non-linear ageing. IEEE Trans Commun, 2020, 68: 4905–4918
428 Wang Z, Badiu M A, Coon J P. A framework for characterizing the value of information in hidden Markov models. IEEE Trans Inform Theor, 2022, 68: 5203–5216
429 Holm J, Chiariotti F, Kalør A E, et al. Goal-oriented scheduling in sensor networks with application timing awareness. IEEE Trans Commun, 2023, 71: 4513–4527
430 Maatouk A, Assaad M, Ephremides A. The age of incorrect information: an enabler of semantics-empowered communication. IEEE Trans Wireless Commun, 2023, 22: 2621–2635
431 Lu M, Huang J, Yang T, et al. Utility loss of information optimal for semantic empowered RSMA in satellite-integrated internet. IEEE Int Things J, 2024, 11: 40572–40587
432 Xu L, Jiao J, Yang T, et al. Semantic utility loss of information for energy efficient semantic status update communications. IEEE Trans Cogn Commun Netw, 2025, 11: 59–74
433 Zuo Y, Yue M, Yang H, et al. Integrating communication, sensing and computing in satellite Internet of Things: challenges and opportunities. IEEE Wireless Commun, 2024, 31: 332–338
434 You L, Zhu Y, Qiang X, et al. Ubiquitous integrated sensing and communications for massive MIMO LEO satellite systems. IEEE Int Things M, 2024, 7: 30–35
435 Zhang Y, Wang J, Li Q, et al. Joint communication, sensing, and computing in space-air-ground integrated networks: system architecture and handover procedure. IEEE Veh Technol Mag, 2024, 19: 70–78
436 Kaushik A, Singh R, Dayarathna S, et al. Toward integrated sensing and communications for 6G: key enabling technologies, standardization, and challenges. IEEE Comm Stand Mag, 2024, 8: 52–59
437 Fei Z, Wang X, Wu N, et al. Air-ground integrated sensing and communications: opportunities and challenges. IEEE Commun Mag, 2023, 61: 55–61
438 Xing H, Zhu G, Liu D, et al. Task-oriented integrated sensing, computation and communication for wireless edge AI. IEEE Netw, 2023, 37: 135–144
439 Sun Z, Yu Z, Guo B, et al. Integrated sensing and communication for effective multi-agent cooperation systems. IEEE Commun Mag, 2024, 62: 68–73
440 Khalili A, Rezaei A, Xu D F, et al. Energy-aware resource allocation and trajectory design for UAV-enabled ISAC. In: Proceedings of IEEE Global Communications Conference, Kuala Lumpur, 2023. 4193–4198
441 Liu Y, Liu S, Liu X, et al. Sensing fairness-based energy efficiency optimization for UAV enabled integrated sensing and communication. IEEE Wireless Commun Lett, 2023, 12: 1702–1706
442 Brandon W, Mahle C. Key technology trends-ground terminals. Space Commun, 2000, 16: 125–137
443 Zhang S J, Zhao X T, Zhao Y F, et al. Integrated satellite-terrestrial networks: integrated mode, frequency usage and application prospects (in Chinese). Radio Commun Technol, 2023, 49: 775–787
444 Johnson A. The first phone maker to add satellite texting to its devices is... Huawei. The Verge, 2022. https://www.theverge.com/2022/9/6/23339717/huawei-mate-50-pro-satellite-text-china-beidou
445 Jewett R. Apple to debut iPhone with emergency messaging enabled by Globalstar satellites. Via Satellite, 2022. https://www.satellitetoday.com/mobile-connectivity/2022/09/07/apple-to-debut-iphone-with-emergency-messaging-enabled-by-globalstar-satellites/

446 Bissinger B. AST SpaceMobile announces summer launch date of BlueWalker-3 for direct-to-cell phone connectivity testing AST science. AST SpaceMobile, 2022. https://ast-science.com/2022/06/13/bluewalker-3-launch-date
447 Robert S T, Miller C E. Cellular core network and radio access network infrastructure and management in space. Lynk Global, EP3830980A1. https://patents.google.com/patent/EP3830980A1/en
448 Wall M. SpaceX Starlink satellites to beam service straight to smartphones. Future PLC, 2022. https://www.space.com/spacex-starlink-direct-service-smartphones-t-mobile
449 Generation Partnership Project (3GPP). Study on architecture aspects for using satellite access in 5G (Release 17). TR 23.737. 2021. https://www.3gpp.org/ftp/Specs/archive/23_series/23.737
450 Generation Partnership Project (3GPP). NR NTN (non-terrestrial networks) enhancements (Release 18). TSG RAN, RP-223534. 2023. https://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_98e/Docs
451 Generation Partnership Project (3GPP). New WID: non-terrestrial networks (NTN) for NR phase 3 (Release 19). TSG RAN, RP-234078. 2024. https://www.3gpp.org/ftp/tsg_ran/TSG_RAN/TSGR_102/Docs
452 He Y, Xiao Y, Zhang S, et al. Direct-to-smartphone for 6G NTN: technical routes, challenges, and key technologies. IEEE Netw, 2024, 38: 128–135
453 Sun Y H, Xu H T, Peng M G. Direct-to-mobile low earth orbit satellite communication: architecture, key technologies, and future perspective (in Chinese). Mobile Commun, 2024, 48: 103–110
454 Abdelsadek M Y, Karabulut-Kurt G, Yanikomeroglu H, et al. Broadband connectivity for handheld devices via LEO satellites: is distributed massive MIMO the answer? IEEE Open J Commun Soc, 2023, 4: 713–726
455 Heo J, Sung S, Lee H, et al. MIMO satellite communication systems: a survey from the PHY layer perspective. IEEE Commun Surv Tut, 2023, 25: 1543–1570
456 Tuzi D, Aguilar E F, Delamotte T, et al. Distributed approach to satellite direct-to-cell connectivity in 6G non-terrestrial networks. IEEE Wireless Commun, 2023, 30: 28–34

Appendix A

Table A1 List of abbreviations.

Abbreviation	Definition	Abbreviation	Definition
3GPP	3rd generation partnership project	MEO	Medium earth orbit
5G	Fifth generation	MIMO	Multiple-input and multiple-output
6G	Sixth generation	MIPv6	Mobile IP version 6
APD	Avalanche photodiode	ML	Maximum likelihood
ACI	Adjacent channel interference	MMF	Mobility management function
ACK	Acknowledgment	mMIMO	Massive MIMO
AoA	Angles of arrival	AoD	Angles of departure
AIMD	Additive increase multiplicative decrease	MIMD	Multiplicative increase multiplicative decrease
AMF	Access and mobility management function	MMSE	Minimum mean square error
AoI	Age of information	MPA	Message passing algorithm
AoII	Age of incorrect information	MRB	Most reliable basis
ARQ	Automatic repeat request	MRC	Maximum ratio combining
AWGN	Additive white Gaussian noise	MS	Min-sum
BBSP	Best beam selection policy	MSCS	Mobile satellite communication system
BCH	Bose-Chaudhuri-Hocquenghem	MT	Mobile terminal
BER	Bit error rate	NACK	Non-ACK
BLER	Block error rate	NC	Network coding
BP	Belief propagation	NFV	Network functions virtualization
BPP	Binomial point process	NOMA	Non-orthogonal multiple access
CC	Chase combining	NR	New radio
sCSI	Statistical CSI	iCSI	Instantaneous CSI
CCS	Cohesive clustered satellites	SatCom	Satellite communication
CCI	Co-channel interference	NRV	Network resource virtualization
DD	Delay-Doppler	DSIN	Distributed satellite information networks
ETSI	European Telecommunications Standards Institute	EO	Earth observations
		MDI	Mirror Doppler interference
F6	Fast, flexible, fractionated, free-flying	DARPA	Defense advanced research projects agency
BS	Base station	OFDM	Orthogonal frequency-division multiplexing
CCSDS	Consultative committee for space data systems	NTN	Non-terrestrial network
		NCR	Network controllable repeater
CDMA	Code division multiple access	OSD	Ordered statistics decoding
CFO	Carrier frequency offset	OTFS	Orthogonal time frequency space
CP	Cyclic prefix	PARP	Peak-to-average power ratio
CRC	Cyclic redundancy check	PGD	Pure ground-based deployment
CS	Compressive sensing	PHY	Physical layer
CU	Central unit	DU	Distributed unit
RRC	Radio resource control	PDCP	Packet data convergence protocol
MAC	Medium access control	RLC	Radio link control
RU	Radio unit	NGSO	Non-geostationary satellite orbits

(Continued on next page)

(Continued)

Abbreviation	Definition	Abbreviation	Definition
CSI	Channel state information	PMIPv6	Proxy MIPv6
ISL	Inter-satellite links	O-RAN	Open radio access networks
OTA	Over-the-air	TDD	Time-division duplex
OAI	OpenAirInterfac	UDP	User datagram protocol
D3QN	Dueling double DQN	POMDP	Partially observable Markov decision process
Deep JSCC	Deep joint source-channel coding	QC-LDPC	Quasi-cyclic LDPC
SAT-CU	Satellite centralized unit	SAT-DU	Satellite distributed units
UDM	Unified data management	SMF	Session management function
UPF	User plane function	PCF	Policy control function
AUSF	Authentication server function	SAR	Synthetic aperture radar
MPC	Model predictive control	RSFF	Reconfigurable satellite formation flying
ML	Machine learning	LDS	Low-density signatures
CR	Cognitive radio	MADRL	Multi-agent deep reinforcement learning
DQN	Deep Q-network	RF	Radio frequency
DRL	Deep reinforcement learning	RIS	Reconfigurable intelligent surface
DTN	Delay/interrupt tolerant networking	RL	Reinforcement learning
DVB-S2	Digital video broadcasting-satellite -second generation	RLNC	Random linear NC
		RS	Reed-solomon
eMBB	Enhanced mobile broadband	RSMA	Rate-splitting multiple access
FBMC	Filter bank multi-carrier	SAGIN	Space-air-ground-sea integrated network
FDMMA	Flexible and distributed mobility management architecture	HAPs	High-altitude platforms
FSO	Free space optical	SBL	Sparse Bayesian learning
FEC	Forward error correction	SC	Successive cancellation
FMIPv6	Fast handover for MIPv6	SC-LDPC	Spatially-coupled LDPC
GE	Gaussian elimination	SCL	SC list
HTS	High-throughput satellite	IRS	Intelligent reflecting surfaces
TCP/IP	Transmission control protocol/Internet protocol	DBPR	Distance-based back-pressure routing
		RTT	Round-trip time
GFDM	Generalized frequency division multiplexing	SCMA	Sparse code multiple access
GFRA	Grant-free random access	SCPS	Space communication protocol standard
GRAND	Guessing random additive noise decoding	SDN	Software-defined networking
		LFNs	Long fat networks
GS	Ground station	SE	Spectral efficiency
QKD	Quantum key distribution	QSDC	Quantum secure direct communication
HARQ	Hybrid ARQ	SFF	Satellite formation flying
HMIPv6	Hierarchical MIPv6	SFFT	Symplectic finite Fourier transform
IAI	Inter-antenna interference	SGIMM	Space-ground integrated mobility management
ICI	Inter-carrier interference		
IDI	Inter-Doppler interference	S-IoT	Satellite Internet of Things
IETF	Internet engineering task force	SIC	Successive interference cancellation
IP	Internet protocol	SIGMA	Seamless IP diversity based generalized mobility architecture
IQI	In-phase and quadrature imbalance		
IR	Incremental redundancy	SLA	Satellite LA
ITU-R	Radio Communication Division of the International Telecommunication Union	VITAL	Virtualized hybrid satellite-terrestrial systems
MTs	Mobile terminals	SMFD	Space-based management function deployment
ISFFT	Inverse SFFT		
ISI	Inter-symbol interference	SNR	Signal-to-noise ratio
ISCC	Integrated sensing, communication, and computation	SSIM	Structural similarity index
		TBCC	Tail-biting convolutional code
JD	Joint decoder	TCP	Transmission control protocol
KB	Knowledge base	TEP	Test error pattern
LA	Location area	TF	Time-frequency
LCMA	Lattice-code multiple access	UFMC	Universal filtered multi-carrier
LDPC	Low-density parity-check	ULA	User LA
LEC	Long erasure code	UoI	Utility of information
LEO	Low earth orbit	URLLC	Ultra-reliable and low latency communications
LM	Location management		
LMS	Land mobile satellite	VAC	Virtual agent cluster
LNA	Low noise amplifier	VAD	Virtual agent domain
LoS	Line of sight	VAP	Virtual attachment point

(Continued on next page)

(Continued)

Abbreviation	Definition	Abbreviation	Definition
LSMR	Least squares minimum residual	VBI	Variational Bayesian inference
PAPR	Peak-to-average power ratio	RAR	Random access response
RAPID	Random access preamble identifier	C-RNTI	Cell-radio network temporary identifier
TA	Timing advance	TSC	Terrestrial-satellite communication
ZC	Zadoff-Chu	DFT	Discrete Fourier transform
LSP	Longer side priority	VMIPv6	Virtual MIPv6
RMS	Root mean square	TDD	Time division duplexing
SNS3	Satellite network simulator 3	GA	Genetic algorithms
ACO	Ant colony optimization	SA	Simulated annealing
LTE	Long-term evolution	VQA	Visual question answering
LTP	Licklider transmission protocol	VoI	Value of information
MBMS	Multimedia broadcast/multicast service	WT	Whitening transformation
LOS	Line-of-sight	IP	Internet protocol
UAVs	Unmanned aerial vehicles	LCRD	Laser communications relay demonstration
PAT	Pointing, acquisition, and tracking	FOV	Field-of-view
RRU	Remote radio unit	FDD	Frequency-division duplex
LLMs	Large language models	LPS	Low-correlation-zone periodic sequence