Toward Communication-Efficient Space Data Centers: Bottlenecks, Architectures, and New Paradigms

Minghao Sun, Zehui Chen, Jinbo Hou, Kezhi Wang, *Senior Member, IEEE,* Xiaoli Chu, *Senior Member, IEEE,*

***Abstract*—The rapid growth of foundation model training and large-scale AI services has driven ground data centers toward unprecedented power densities, intensifying challenges in energy supply, cooling, and spatial scalability. Space Data Centers (SDCs) have emerged as a promising paradigm for hosting energy-intensive computing infrastructures in orbit, leveraging continuous solar energy and radiative cooling advantages. However, unlike ground facilities primarily constrained by power and site availability, SDCs are fundamentally limited by communication capability. The gap between petabit-scale internal data exchange in ground data centers and the gigabit-scale capacity of ground–space links forms a critical bottleneck. This article systematically analyzes communication constraints in SDC architectures and explores semantic communication as a key enabling paradigm. By transmitting compact, task-relevant semantic representations instead of raw data, uplink pressure can be substantially reduced. The feasibility of communication-efficient orbital AI infrastructures is demonstrated through the evaluation of a multi-layer heterogeneous SDC framework consisting of relay satellites and orbital computing nodes operating under coupled energy and thermal constraints. The article further outlines open research challenges toward scalable deployment.**

***Index Terms*—Space data centers, semantic communication, orbital AI infrastructure, energy and thermal constraints.**

I. Introduction

THE rapid advancement of Artificial Intelligence (AI) has led to an unprecedented growth in data volume required for model training and continuous adaptation [1]. At present, large hyper-scale data centers already operate at power levels close to 100 MW, and several proposals envision further scaling toward gigawatt-level computing facilities [2]. Such power densities are comparable to those of medium-sized power plants, making grid integration and peak-load management increasingly challenging. In this context, Space Data Centers (SDCs) leveraging satellite platforms have emerged as a promising paradigm for hosting large-scale computing resources outside densely populated regions, offering potential advantages in energy acquisition, cooling mechanisms, and spatial scalability [3]. Recent industrial developments further illustrate this trend. In May 2025, Zhijiang Laboratory in China launched the first batch of computing satellites under the Three-Body Computing Constellation program. In November 2025, the Tongyi Qianwen (Qwen3) Large Language Model (LLM) was uploaded to the above satellites, enabling on-orbit deployment and inference. Also in November 2025, StarCloud (formerly Lumen Orbit) launched a prototype orbital computing facility equipped with an H100 GPU, demonstrating onboard training with nano-GPT and inference testing using Google's Gemma model. Google has announced Project Suncatcher, targeting the launch and hardware validation of prototype satellites to establish a space-based data infrastructure platform by 2027. Meanwhile, SpaceX has outlined plans to deploy computing-enabled satellites integrated with the Starlink network, forming a distributed architecture that dynamically coordinates communication and computational resources in orbit. Such developments position SDCs as a new frontier in the evolution of AI infrastructure. However, this emerging computing paradigm introduces new challenges for communication architectures and system design, particularly under the coupled constraints of energy availability and thermal dissipation in space environments. Existing studies on SDCs have largely focused on energy harvesting and thermal management [4], or provided a high-level overview of software control for SDC-oriented cloud networking [5], while communication mechanisms and communication-oriented architectural design remain relatively underexplored. This gap raises a fundamental question: Can current communication technologies support the massive, heterogeneous data transmission requirements of SDCs under multi-dimensional resource constraints?

In this article, we address this question through a systematic analysis of communication bottlenecks and representative service scenarios in SDC systems. We first identify the fundamental bottlenecks in existing SDC communication architectures and characterize the key service scenarios that determine their operational requirements. Based on this analysis, we identify candidate communication technologies capable of supporting SDC deployment across diverse application contexts. A further contribution of this work is the incorporation of semantic communication as a foundational paradigm for SDC systems. Instead of transmitting full-volume raw datasets, semantic communication conveys compact, task-relevant representations extracted from large-scale data. This shift reduces dependence on raw-data backhaul and alleviates pressure on ground-space link capacity. To examine the feasibility of this approach, we develop a multi-layer heterogeneous satellite network architecture that integrates SDC nodes with communication satellites acting as relays between SDC nodes and ground stations. By incorporating semantic communication mechanisms, we evaluate the system's ability to sustain large-scale data transmission services under joint energy and thermal constraints. The results offer practical insights into the design of communication-efficient SDC infrastructures. Finally, we outline a research roadmap for semantic-enabled SDC frameworks, identifying key development stages and future research directions.

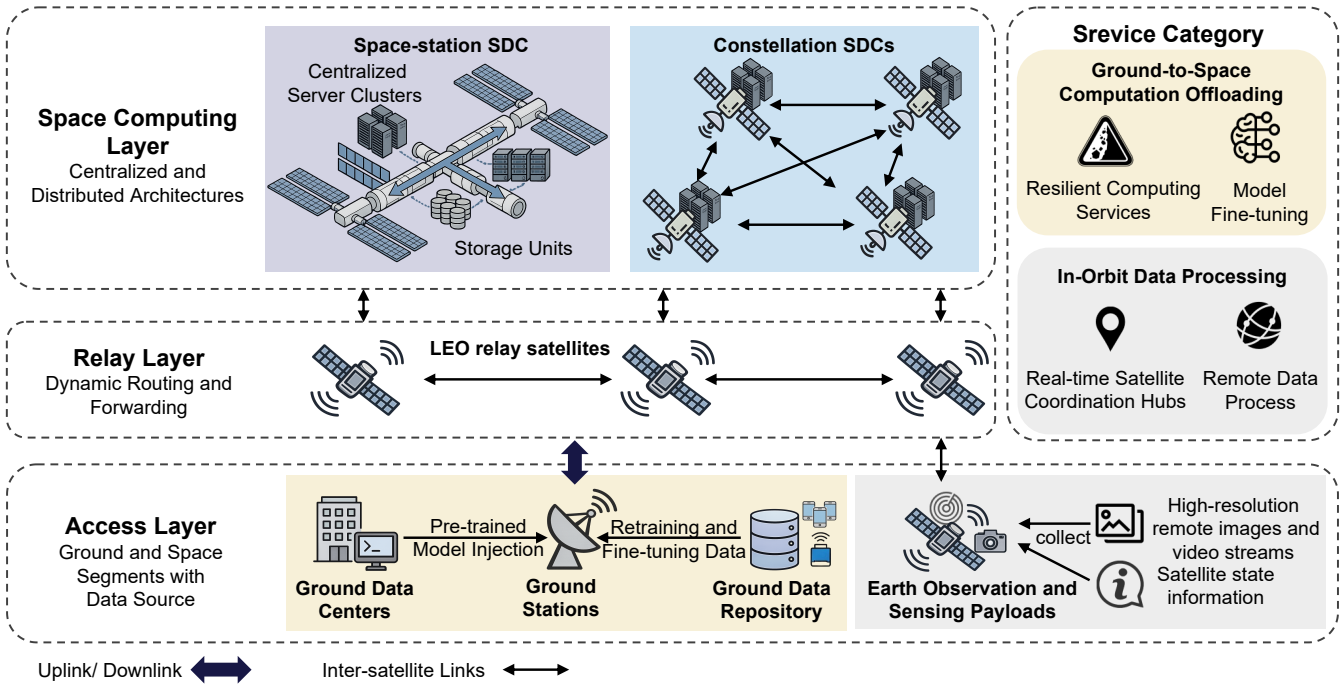


Fig. 1. Space data centers: multi-layer architecture, and heterogeneous service types.

II. Toward Sustainable and Efficient Computing in Space

In this section, we outline the system-level advantages of SDCs over conventional ground-based data centers and present a multi-layer heterogeneous network architecture that consists of a space computing layer, a relay layer, and an access layer, as shown in Fig. 1. Within this framework, GSs communicate directly with SDC platforms when coverage is available; otherwise, outside the SDC service footprint, ground traffic is first offloaded to relay satellites and then forwarded to the SDC. Multi-satellite scheduling mechanisms enable dynamic workload allocation across computing nodes, allowing tasks to be routed, migrated, or replicated based on resource availability and service requirements. This flexibility enables the system to accommodate time-varying demand and heterogeneous workloads.

A. Why Space Data Centers?

Energy: At present, a key bottleneck of ground data centers lies in the growing gap between increasing data processing demand and the limited power supply of regional power grids. In recent years, ultra-large-scale orbital solar arrays have been proposed to support space computing platforms, with the objective of providing long-term and reliable energy supplies for energy-intensive computing workloads. The space environment offers nearly uninterrupted access to solar energy for large-scale computing infrastructures. In contrast to ground-based solar systems, which are constrained by day–night cycles, weather conditions, and geographical locations, solar radiation in orbital space is more stable and predictable. By deploying large-scale solar arrays, SDCs can operate continuously and achieve more stable and efficient power utilization than ground photovoltaic systems [4]. Such sustained energy availability not only supports long-duration computing tasks but also enables more communication-intensive system designs. In particular, it creates the possibility of powering large-scale antenna arrays, high-frequency transceivers, and other energy-demanding communication modules, thereby facilitating higher-capacity space–ground and inter-satellite communications. As a result, the reliance of SDCs on solar energy during operation may reduce their dependence on terrestrial power grids. Compared with ground data centers that draw large amounts of electricity from regional grids, SDCs have the potential to alleviate the energy-system pressure and concentrated carbon emissions associated with the centralized deployment of ultra-large-scale computing facilities.

Cooling: Approximately 58% of the electricity consumed by ground data center facilities is currently used for cooling systems rather than functional computing [6]. The deep-space radiative environment provides fundamentally different conditions for reducing cooling complexity and resource consumption at the system level. Although the lack of physical media in space makes traditional heat dissipation methods relying on convection and water circulation inapplicable, its extremely low background temperature enables the heat generated by onboard computing systems to be dissipated directly into deep space through passive radiative heat transfer, reducing a significant amount of cooling water consumption. Moreover, the heat dissipation capacity of the facilities can be enhanced as the effective heat sink area expands, resulting in a relatively high scalability in heat dissipation capacity.

Space Scalability: The vastness of orbital space allows computing infrastructures unprecedented flexibility in spatial scalability and deployment locations. Unlike ground-based data centers that are constrained by land availability, urban planning, and regional energy policies, SDCs can expand on demand through modular deployment and multi-satellite

collaboration. Such an approach enables distributed space computing networks capable of supporting cross-regional and global computing services. This multi-satellite cooperative computing paradigm opens new possibilities for the elastic expansion and coordinated operation of future ultra-large-scale computing systems.

B. Potential Architectures of SDCs

LEO Satellite Constellation Architecture: Within the constellation, satellites assume complementary roles to form a multi-layer distributed computing architecture in orbit. Computing satellites equipped with onboard servers provide primary processing capability, while dedicated communication satellites extend ground coverage, support task offloading, and act as relay nodes between computing satellites and ground stations. Dynamic interconnections among the computing nodes enhance architectural flexibility, enabling new computing satellites to be added and aging platforms to be replaced without interrupting service continuity.

Space Station–Hosted Data Center Architecture: This approach deploys server clusters directly on a space station, effectively establishing a centralized data center in orbit. A key advantage is that its structural design and maintenance requirements closely resemble those of conventional ground-based data centers, thereby reducing the need for additional maintenance mechanisms. Moreover, co-locating computing resources on a single space platform centralizes processing capability and lowers the overhead associated with multi-satellite trajectory coordination and inter-satellite communication, simplifying overall system management.

C. What Can SDCs Do?

In-Orbit Data Processing: Such services leverage onboard pre-processing and intelligent analysis of remote sensing images or video streams to reduce downlink data volume and improve real-time responsiveness. In addition, SDCs can act as coordination hubs for large-scale space-based systems, using real-time satellite state information to optimize orbital trajectories and service allocation. This enables dynamic, near real-time coordination across satellite clusters. In this scenario, latency-sensitive tasks are prevalent, making efficient multi-satellite routing and communication resource allocation critical.

Ground-to-Space Computation Offloading: Through ground–satellite links, data can be offloaded for in-orbit processing, where continuous solar energy supports sustained large-scale computation. This architecture is particularly suitable for latency-tolerant yet computation-intensive workloads, such as large-model training, retraining, and multi-task fine-tuning. Beyond dedicated space-based services, SDCs can also serve as complementary computing resources to ground data centers. For instance, during peak demand periods or in the event of natural disasters, selected workloads can be shifted to orbital platforms to relieve pressure on ground infrastructure and enhance overall system resilience.

III. Bottlenecks and Solutions

In this section, we examine the fundamental bottlenecks limiting current SDC deployments and outline potential technical solutions along with their applicable scenarios.

A. Challenges Facing SDCs

Communication Capacity: The rapid rise of LLMs has made training and inference for foundation models core workloads in modern data centers. Supporting model pre-training, continual adaptation, and large-scale inference requires massive datasets and extensive computational resources, driving rapid expansion in data center capacity and scale. In ground data centers, infrastructure expansion is primarily constrained by power supply and site availability. In contrast, SDCs face system-level limitations that are more closely tied to communication capability. According to 'Speed, Scale and Reliability: 25 Years of Google Data-Center Networking Evolution' reported by Google Cloud in 2024, the data throughput of a single aggregation block in modern data center networks has reached approximately 200 Tb/s. When dozens of such aggregation blocks are interconnected, the aggregate data exchange capacity can scale to the Pb/s level. However, due to limited transmit power, constrained spectrum resources, severe propagation losses over long distances, and intermittent connectivity, existing ground–satellite access links typically provide capacities of only tens to hundreds of Gb/s [7].

Table I lists the uplink capacity of several representative satellite platforms and communication technologies currently under development or deployment. The table shows that existing commercial access links typically range from Mbps to multi-Gbps, while emerging optical communication systems can achieve rates of 10–100 Gbps or higher. Despite these advances, current satellite links remain far below the data-transfer requirements of large-scale SDCs, making communication capacity a fundamental deployment bottleneck.

Software-Defined System Scheduling: Software-defined scheduling faces multidimensional challenges in SDC environments. Due to the time-varying and intermittent connectivity of satellite links, task feedback, control signals, and model updates cannot always be disseminated across distributed nodes in real time. As a result, during no-contact intervals, different nodes may temporarily operate under inconsistent software versions or configuration states. Task data may accumulate in local buffers and be transmitted in batches once connectivity is restored, leading to bursty traffic patterns and transient congestion. To ensure reliable system operation, scheduling mechanisms must therefore accommodate not only dynamically varying channel capacity but also delayed state synchronization. This requires elastic buffering strategies and priority-aware scheduling policies to maintain stable task execution under asynchronous updates and propagation delays, while preventing adverse impacts on task-related transmission rates.

Data Security: The openness and broadcast nature of SDCs' long-distance ground-to-satellite wireless transmissions increase the risk of data leakage and eavesdropping [8]. Future SDC infrastructures may support multiple organizations,

TABLE I
SINGLE-UPLINK CAPACITIES UNDER DIFFERENT PLATFORMS AND TECHNOLOGIES

Platform / technology	Single uplink capacity	Source
Starlink (typical service)	5–20 Mbps	Starlink, Specifications
Starlink lower-speed tier	20 Mbps	Starlink, Network Update
SES O3b mPOWER (per terminal)	Multi-Gbps	SES, O3b mPOWER
SES O3b mPOWER IP Transit	Up to 10 Gbps	SES, O3b mPOWER IP Transit Executive Guide
Airbus / CNES TELEO optical feeder link	10 Gbps	Airbus, TELEO Announcement
NASA next-generation optical uplink	>100 Gbps	NASA, Next-Generation Optical Communications Relay
NASA optical relay pathfinder	100 Gbps	NASA, Optical Relay Pathfinder

services, and heterogeneous tasks operating on shared orbital resources. In such multi-tenant, diverse-service environments, security mechanisms designed to protect a specific task may introduce additional overhead that degrades the performance of other workloads. Moreover, ensuring data privacy and establishing trust across different stakeholders adds further complexity to system design.

B. SDC Enabling Technologies

Within the conventional communication paradigm, the most critical structural bottleneck in SDC systems is limited communication capacity, which fundamentally constrains the scalability of ground-to-space data transmission. The other challenges discussed above become significant only after this primary bottleneck has been alleviated and the system is capable of supporting the simultaneous transmission and coordination of numerous heterogeneous tasks. Accordingly, this subsection focuses on the key technologies that may help overcome the communication-capacity limitations of current SDC architectures.

Semantic Communication: Semantic communication has emerged as a promising approach to addressing the communication capacity bottleneck in SDC scenarios. Unlike bit-level communication, semantic communication prioritizes semantic fidelity or task utility rather than exact bit reconstruction. By proactively filtering redundant or task-irrelevant content under bandwidth constraints, semantic communication can significantly improve effective throughput and communication efficiency [9]. A typical semantic communication system includes a shared knowledge base or semantic model, with the receiver using this shared knowledge for semantic inference and information recovery. In this regard, SDCs offer a unique advantage: their substantial computational power and large-scale storage capacity allow maintenance and continuous updating of massive semantic models and knowledge repositories. With these resources, a semantic receiver can actively retrieve factual or structured knowledge from SDC-hosted knowledge bases to assist in semantic decoding and reasoning [10]. In SDC-oriented AI workloads, in-orbit fine-tuning of deployed LLMs represents a major task. Such a training objective is typically well defined and does not require preserving the complete data distribution. Samples with marginal contribution to model updates can be discarded, while only high-impact gradient information or task-relevant features are transmitted [11]. This aligns naturally with task-oriented semantic communication, which prioritizes task performance over bit-level reliability. Instead of minimizing bit error rate, the communication process is designed to maximize task success probability or minimize task loss. By transmitting high-value samples or semantically compressed feature representations, link utilization can be significantly reduced. Semantic encoders extract task-relevant features and jointly optimize task success rate, training loss, and downstream performance metrics. Such techniques can reduce communication overhead by up to 60% compared with raw data transmission [12].

Multi-Antenna Beamforming for SDC Networks: Multiple-Input Multiple-Output (MIMO) employs multiple antennas at the transmitter and receiver to improve link performance through array gain, beamforming, and spatial multiplexing. In high-SNR scenarios, these mechanisms can increase the effective capacity of an individual link to the Gb/s level [13]. In SDC systems, relatively stable energy supplies and substantial onboard processing capabilities can support large antenna arrays, high-frequency transceivers, and real-time beam control. Nevertheless, practical deployment requires sufficiently large antenna apertures and introduces additional constraints on onboard power consumption, thermal dissipation, mass, and spacecraft accommodation. Narrow-beam transmission also demands accurate beam steering and tracking to maintain alignment under rapid relative motion, orbital dynamics, and platform vibrations in LEO.

Inter-Satellite Laser Communication: Inter-Satellite Laser (ISL) communication is a high-capacity optical transmission technology that employs narrow laser beams to establish direct space-to-space links between satellites. According to the 2025 *IEEE Spectrum* report [14], ISL links can achieve data rates of up to 400 Gbps. Compared with conventional radio-frequency links, ISL links offer significantly higher bandwidth and lower transmission latency. Such high-throughput connectivity provides a critical networking foundation for distributed computing and real-time coordination within SDC architectures, enabling satellites to operate as an integrated orbital computing system. In practice, such capacities may require advanced optical terminals, wavelength-division multiplexing, and precise ephemeris-aided acquisition, tracking, and pointing under rapid LEO relative motion.

IV. A COMMUNICATION-EFFICIENT SDC FRAMEWORK INCORPORATING SEMANTIC COMMUNICATION UNDER COUPLED ENERGY AND THERMAL CONSTRAINTS

In practice, SDC operation requires co-design across three tightly coupled domains: energy, thermal management, and communication. Computing workloads consume power and generate heat, while thermal constraints limit sustainable

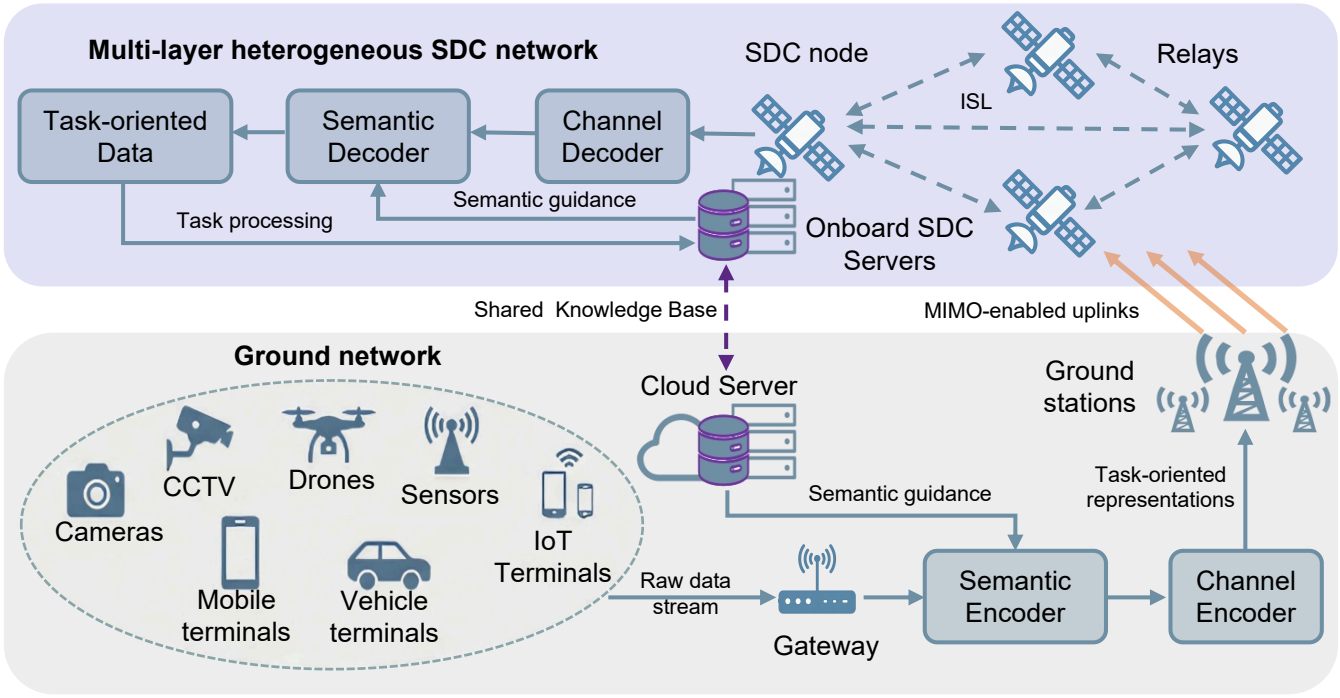


Fig. 2. Architecture of a Semantic Communication-enabled Multi-Layer Heterogeneous SDC Communication Framework.

compute intensity, and intermittent ground-space connectivity governs when data, model updates, and control signals can be exchanged. Consequently, the system must be managed through both long-term window-level scheduling and real-time computing control mechanisms. Accordingly, in this section, we first introduce the fundamental network architecture of SDC systems. Based on this framework, we then compare semantic communication with conventional bit-level communication in terms of uplink utilization and ground station (GS) energy consumption under the joint constraints of power supply, computational demand, and thermal dissipation.

A. A Multi-layer Heterogeneous Semantic Communication-enabled SDC Framework

As illustrated in Fig. 2, a semantic communication (SemCom)-enabled SDC framework is designed to support large-volume ground-space transmission. The proposed framework integrates communication and computing into an end-to-end SDC service chain. Semantic encoders first extract compact task-relevant representations from data generated by ground or space sources. To ensure compatibility, each encoder-decoder pair is associated with a model-version identifier and a knowledge-base snapshot index, allowing the ground and space segments to verify consistency before transmission and decoding. The representations are then delivered through MIMO-enhanced ground-space links and, when direct SDC access is unavailable, forwarded by relay satellites through high-capacity ISLs. Because ground-space contacts are intermittent, semantic models and knowledge bases are updated through contact-window-aware incremental synchronization, where only changed model parameters, task-specific dictionaries, or knowledge entries are transmitted during high-capacity access periods. Frequently used knowledge can be cached or replicated across SDC nodes via ISLs, while less frequently accessed content remains in centralized orbital storage or is retrieved from the ground when latency permits. SDC nodes then perform in-orbit AI processing and return task results or model updates through the same network. In this framework, semantic communication reduces transmission demand, MIMO improves ground-space access capacity, ISLs support both orbital transport and knowledge replication, and version-controlled synchronization prevents semantic mismatch between distributed ground and space entities. Furthermore, the dynamic interconnection among satellites enhances architectural flexibility, allowing new computing nodes to be incrementally integrated and aging platforms to be replaced without disrupting ongoing services. This modular and evolvable structure supports scalable expansion and long-term system upgrades.

TABLE II
SIMULATION PARAMETERS [2]

Parameter	Value	Parameter	Value
Power generation density	449 W/m^2	Ground elevation angle	30°
Heat dissipation density	240 W/m^2	ISL capacity penalty	6 dB
Number of SDC nodes	1	Carrier frequency	30 GHz
RF bandwidth	1.6 GHz	Transmit antenna gain	43 dBi
Receive antenna gain	43 dBi	System noise temperature	600 K

B. Case Study for Ground-Space Computation Offloading

To evaluate whether the proposed framework can sustain high-throughput SDC operation, we compare conventional bit-level communication (BitCom) with SemCom within a multi-layer multi-LEO satellite network. Specifically, we examine the ground-space uplink capacity required by each scheme to achieve a throughput matching the maximum sustainable data processing volume.

In the simulated framework, the total SDC service duration is set to 10s, divided into 10 time slots of 1s each. The ground segment consists of 30 GSs, while the space segment includes 24 communication satellites and 1 SDC operating at an altitude of 500 km. The communication satellites are assumed to have no onboard computing capability; instead, they forward all received semantic information to the SDC, which is equipped with multiple H100 GPUs, through 400 Gbps ISL links for processing. Each GS is assumed to access up to two orthogonal ground-satellite channels simultaneously, where each satellite can maintain up to six simultaneous inter-satellite links. The remaining simulation parameters are summarized in Table II. A load-aware scheduling mechanism is incorporated, such that when a relay satellite experiences high traffic load, data are preferentially routed to alternative satellites with lower load whenever connectivity permits.

The SDC platform is equipped with a square solar array and a rear-mounted radiative panel for heat dissipation to minimize structural footprint. Let the total solar power generation be denoted by P, and the computing power consumption by E_1. Since most of the electrical power consumed by GPUs is converted into heat, the thermal generation rate is also approximated as E_1. The processing energy consumption is assumed to be proportional to the processed data volume X [15]. Since the energy consumption associated with signal reception is significantly smaller than the computing power demand under the considered system configuration, it is neglected in the subsequent analysis for simplicity. The radiative panel provides a net heat dissipation capacity of E_2 after excluding the absorbed solar radiation [2]. Due to the need for active coolant circulation, we assume that the pumping power consumption accounts for 2% of the radiative dissipation power. Therefore, sustainable system operation requires satisfying the following conditions: $E_1 + 0.02E_2 \leq P$ and $E_1 \leq E_2$. The first constraint ensures sufficient solar power for both computation and cooling, while the second guarantees that generated heat does not exceed the radiative dissipation capacity. Under a fixed power budget P, these constraints determine the maximum sustainable data processing volume $X_{\max}$.

In the semantic communication configuration, we employ two representative datasets, CIFAR-10 and AG News, corresponding to image and text modalities, respectively, to evaluate the proposed framework across different data types relevant to future AI workloads. Each GS employs a modality-specific encoder: a modified ResNet18 with a 3×3 input convolution for CIFAR-10, and DistilBERT for each 128-token AG News sequence. The ResNet18 features are compressed into an 80-dimensional latent vector, while the 768-dimensional DistilBERT feature is projected through a 256-dimensional hidden layer into a 4-dimensional latent vector. Using straight-through quantization, the CIFAR-10 and AG News latent vectors are quantized to 3 and 2 bits per dimension, respectively. With an additional 16 bits of overhead, they form 256-bit and 24-bit application payloads.

Fig. 3 shows the equivalent single-channel data rate per GS required by SemCom and Bitcom to achieve $X_{\max}$ on the CIFAR-10 and AG News datasets under different SDC power budgets. According to Table I, a single ground-space link can support data rates of up to 100 Gbps, which is indicated by the horizontal reference line in Fig.3. The results show that, for an SDC with a processing capability exceeding 50 MW, conventional bit-level communication requires per-link capacities beyond the 100 Gbps limit across all evaluated scenarios. This suggests that, under current communication capabilities, BitCom alone cannot sustain full-load SDC operation. In contrast, SemCom dramatically reduces the required ground-space link utilization by more than 98.9% for CIFAR-10 and 98.7% for AG News. We have also evaluated several additional performance metrics, including end-to-end latency, energy consumption at GS, and task accuracy, as summarized in Table III. Across different data volumes, SemCom consistently outperforms BitCom in terms of latency and energy consumption for both datasets, while achieving comparable task accuracy.

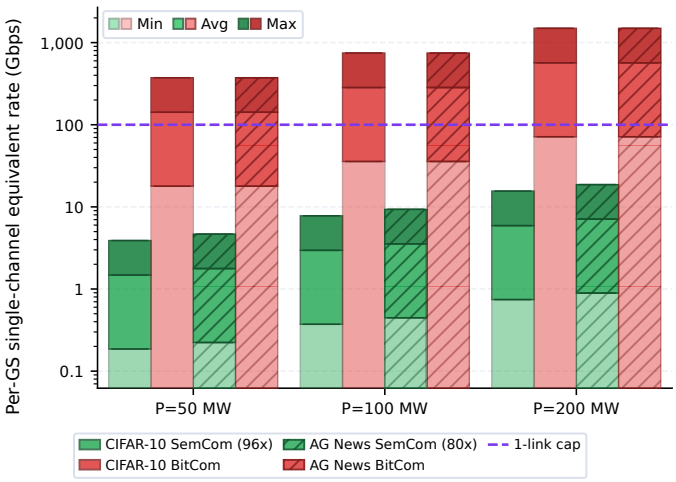


Fig. 3. Per-GS equivalent single-channel data rate required to achieve $X_{\max}$ on the CIFAR-10 and AG News datasets under different SDC power budgets.

TABLE III
PERFORMANCE COMPARISON BETWEEN SEMCOM AND BITCOM

Dataset	Volume	Latency (ms)		Energy (J)		Accuracy (%)	
		SemCom	BitCom	SemCom	BitCom	SemCom	BitCom
CIFAR-10	1 Gbit	23.3	33.2	2.841	6.990	94.43	94.67
	1.5 Gbit	23.7	37.2	2.911	8.972		
	3 Gbit	25.2	52.2	3.159	15.281		
AG News	1 Gbit	23.2	32.2	2.838	6.870	94.68	94.71
	1.5 Gbit	24.1	37.2	2.973	8.972		
	3 Gbit	25.5	52.2	3.221	15.281		

The results show that despite the additional computational overhead introduced by the semantic encoder, the performance of semantic communication still remains substantially better than that of conventional bit-level transmission, demonstrating that the proposed semantic communication framework consistently achieves substantial communication-capacity savings across both image and text modalities, further supporting its potential for future multimodal AI workloads. Notably, since many of the underlying technologies are still at the conceptual stage, these simulations should be regarded as proof-of-concept studies. More detailed technical implementations, engineering considerations, and system-level optimizations will be investigated in our future work.

V. Future Research Directions

Realising the full potential of semantic communication for supporting SDCs still presents several open challenges. Based on the multi-layer heterogeneous SDC framework evaluated in this study, we outline a set of future research directions, identify key challenges, and discuss potential solution pathways.

Token-Level Semantic Reconstruction: Large-scale pre-training of LLMs requires learning general data distributions from fully tokenized sequences. In such settings, defining task importance or selectively filtering tokens without compromising distributional integrity becomes highly challenging. An alternative approach is semantic-oriented communication, which prioritizes reconstruction fidelity at the semantic level rather than strict task-specific objectives. While this strategy may alleviate the limitations of task-oriented filtering, it introduces new challenges, including distribution shift, over-normalization effects, hallucinated reconstruction errors, and reduced reproducibility across training runs. If these challenges can be systematically addressed, semantic communication could potentially extend beyond fine-tuning scenarios to support a broader range of LLM training workloads, thereby substantially reducing ground–space link pressure in large-scale SDC deployments.

Cross-Task Knowledge Generalization: Semantic communication relies on shared knowledge bases or semantic models for encoding and decoding. Existing semantic communication systems typically construct task-specific knowledge repositories tailored to predefined objectives, which limits their flexibility across heterogeneous workloads. We plan to adopt generative foundation models to facilitate knowledge transfer and dynamic knowledge adaptation across tasks, which may improve cross-task generalization and reduce the need for manual knowledge engineering, thereby enhancing the scalability and flexibility of semantic communication in SDC environments.

Long-Term Security Risks: Semantic information (SI) can be cached, aggregated, and reused for training or retrieval. Even without access to the original data, an adversary may exploit large volumes of accumulated SI as a queryable semantic index to infer sensitive attributes or critical patterns. In SDC communication architectures, where semantic data may be stored, shared, and reused across multiple transmissions and tasks over extended periods, the security focus of semantic communication shifts from protecting transient communication channels and bitstreams to safeguarding SI as a long-term data asset, thereby requiring fundamentally new security mechanisms. For example, local differential privacy (DP) or DP-aware perturbation can be applied to latent semantic representations before uplink transmission, while privacy-sensitive dimensions can be selectively suppressed or perturbed to mitigate inversion and inference risks.

VI. Conclusion

In this article, we have presented a systematic analysis of communication constraints in SDC architectures and examined representative service scenarios under joint energy and thermal considerations. Building upon a heterogeneous multi-layer architecture integrating relay satellites and inter-satellite laser links, we have further proposed a potential communication-efficient SDC framework. By incorporating semantic communication as a foundational communication paradigm, we have shown that task-oriented semantic representations can provide a promising approach for alleviating ground-space uplink pressure while preserving task performance.

Despite these promising findings, several open challenges remain, including token-level semantic reconstruction for large-scale pre-training, adaptive cross-task knowledge generalization, and long-term security risks associated with accumulated semantic information. Addressing these issues will be essential for extending semantic communication beyond fine-tuning scenarios and enabling scalable orbital AI infrastructures. Overall, the practical viability of future gigawatt-scale SDC systems will depend not only on computational capability, but also on communication-efficient system design.

Acknowledgment

This work was supported in part by the Horizon Europe Research and Innovation Program under grants 101086219 and 101086228, the UK EPSRC grants EP/X038971/1 and EP/Y028031/1, Innovate UK COMET project (No: 10099265), and Royal Society Industry Fellowship (IF\R2\23200104). Minghao Sun and Zehui Chen are co-first authors. Xiaoli Chu is the corresponding author.

References

[1] B.A. Arcas, et al., "Towards a future space-based, highly scalable AI infrastructure system design," *arXiv preprint arXiv:2511.19468*, vol. 4, 2025.

[2] E. Feilden, et al., "Why we should train AI in space," StarCloud, White Paper, 2024.

[3] L. Kuang, et al., "Towards space-based computing infrastructure network: Development trends, network architecture, challenges analysis, and key technologies," *arXiv preprint arXiv:2503.06521*, 2025.

[4] A. Aili, et al., "The development of carbon-neutral data centres in space," *Nature Electronics*, pp. 1–11, Oct. 2025.

[5] W. Ge, et al., "Space cloud networks: Technologies, architectures, and key enablers," *IEEE Communications Magazine*, Mar. 2026.

[6] A. Shehabi, et al., "2024 united states data center energy usage report," Lawrence Berkeley National Laboratory, Dec. 2024.

[7] O.B. Yahia, et al., "Evolution of high throughput satellite systems: Vision, requirements, and key technologies," *arXiv preprint arXiv:2310.04389*, 2023.

[8] P. Tedeschi, et al., "Satellite-based communications security: A survey of threats, solutions, and research challenges," *Computer Networks*, vol. 216, p. 109246, Oct. 2022.

[9] W. Yang, et al., "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 213–250, Nov. 2022.

[10] B. Wang, et al., "Knowledge enhanced semantic communication receiver," *IEEE Communications Letters*, vol. 27, no. 7, pp. 1794–1798, Jul. 2023.

[11] M. Xia, et al., "Less: Selecting influential data for targeted instruction tuning," *arXiv preprint arXiv:2402.04333*, 2024.

[12] W. Chen, et al., "Free space optical semantic communication for satellite remote sensing image transmission," *IEEE Transactions on Communications*, Oct 2025.

[13] T. M. Nguyen, *Satellite Systems: Design, Modeling, Simulation and Analysis*. IntechOpen, 2021.

[14] A. Jones, "China makes high-speed laser links in orbit," https://spectrum.ieee.org/satellite-internet-china-crosslink, May 2025, IEEE Spectrum, accessed Apr. 8, 2026.

[15] Y. Mao, et al., "A survey on mobile edge computing: The communication perspective," *IEEE communications surveys & tutorials*, vol. 19, no. 4, pp. 2322–2358, Aug. 2017.