CHAPTER 1

INTRODUCTION

1.1 Background of Space Exploration

Modern space exploration systems, encompassing everything from advanced Earth-observation satellites to deep-space interplanetary probes, are currently facing a data processing crisis. These missions are increasingly equipped with highly sophisticated sensors, including high-resolution cameras, spectrometers, synthetic aperture radar (SAR), and lidar systems, all of which generate exponentially increasing volumes of data. It is common for a single Earth-observation satellite or a space telescope to produce terabytes of raw imagery and telemetry data every day.

Historically, space missions have operated under a centralized data paradigm, often referred to as the "bent-pipe" architecture. In this model, the spacecraft's onboard computer primarily formats the raw sensor data, which is then downlinked to ground stations for subsequent processing and analysis. While simple, this approach introduces severe limitations. The primary issues include high latency, which delays the availability of crucial information; constrained and costly bandwidth for transferring massive files; and delays resulting from intermittent or infrequent contact windows with ground stations. For time-critical operations, such as near-Earth object detection, swift disaster response, or autonomous hazard avoidance for a planetary rover, these delays can render the collected data nearly useless by the time it reaches the analyst.

1.2 Need for Edge Intelligence in Space

The limitations of traditional ground-centric processing have driven the search for solutions that move computation closer to the data source. Edge intelligence, the deployment of Artificial Intelligence (AI) and computational capabilities directly onto the spacecraft, offers a powerful paradigm shift. By performing computation and AI inference directly on the satellite, rover, or probe (the "edge" of the network), data can be processed in real time without the lengthy round-trip delays required for ground-based analysis.

This approach dramatically improves mission autonomy and responsiveness. For instance, a remote sensing satellite equipped with edge AI can perform onboard analysis of imagery to immediately detect critical events like wildfires, floods, or vessels at sea, transmitting only the critical detection alerts or processed maps rather than gigabytes of raw data. Similarly, autonomous systems on planetary surfaces, such as rovers operating on Mars or the Moon, rely on onboard object recognition and navigation algorithms to operate independently, as communication delays can exceed ten minutes. In essence, edge intelligence shifts the space data processing paradigm from passive downlinking to active, on-orbit computation, enabling systems to react in minutes or seconds, rather than hours or days.

1.3 Objectives and Contributions

This paper systematically explores how edge AI is being integrated into space exploration systems. It begins by establishing the technical foundations of edge intelligence and then surveys the specific architectures, technologies, and system-level changes necessary for deployment in the harsh, resource-constrained space environment.

The objectives include a detailed analysis of the hardware trade-offs between different accelerators (Field-Programmable Gate Arrays, Graphics Processing Units), a comprehensive review of the algorithmic optimizations required for flight (such as model compression), and an examination of emerging distributed learning paradigms. Crucially, the paper integrates and reviews cutting-edge developments from 2024–2025 research, including real-world case studies demonstrating on-orbit capability, such as the inference of Geospatial Foundation Models (GeoFMs) and high-efficiency Synthetic Aperture Radar (SAR) vessel detection.

1.4 Organization of the Paper

This report follows the standard structure outlined in the Table of Contents. Section 2 establishes the fundamentals of edge intelligence. Section 3 provides an overview of existing space platforms and their constraints. Section 4 examines the system-level integration of edge computing into spacecraft architecture. Section 5 details the specialized hardware and software architectures required for space edge AI. Section 6 covers representative applications, and Section 7 analyzes the persistent operational and engineering challenges. Section 8 explores future trends and research directions, followed by concrete case studies in Section 9, and concluding remarks in Section 10.

CHAPTER 2

FUNDAMENTALS OF EDGE INTELLIGENCE

Edge intelligence represents a powerful merger of two technical domains: **edge computing** and **Artificial Intelligence** (AI). Understanding this merger requires first distinguishing the computational environment.

2.1 Computing Paradigms: Cloud, Fog, and Edge

The traditional environment for sophisticated data analysis is **Cloud Computing**, characterized by massive, centralized data centers that offer virtually unlimited power and high processing throughput. However, cloud environments are geographically distant from the data sources, imposing latency unsuitable for real-time space operations.

Edge Computing stands in direct contrast. It refers to the practice of performing computation close to where the data is generated, at the "network edge." This edge can be physical devices, local servers, or, in the space context, satellites and rovers. A formal definition characterizes edge computing as the "enabling technologies that allow computation to be performed at the network edge so that computing happens near data sources." By minimizing the distance between sensor and processor, edge computing significantly reduces the need to transmit raw data to distant servers, thus lowering latency and conserving bandwidth.

2.2 Definitions and Concepts of Edge Intelligence

Edge Intelligence (Edge AI) specifically incorporates machine learning (ML) and AI algorithms, such as deep learning, computer vision, and sensor fusion, into these local, edge devices. Unlike generic edge computing, which might handle simple data formatting or relay tasks, edge AI implies the presence of onboard, smart capabilities, allowing systems to perform complex functions like pattern recognition, prediction, and complex decision-making with minimal dependence on ground infrastructure.

The operation of edge intelligence is governed by several core principles:

1. **Data Locality:** Processing data immediately where it is generated, allowing the system to act on new information instantly.
2. **Resource Constraints:** Edge devices typically operate with tight limits on power, memory, and computational throughput (CPU/GPU), necessitating optimized AI models.
3. **Distributed Intelligence:** In networked scenarios, such as satellite constellations, edge nodes must share data and cooperate to solve joint tasks.

2.3 Enabling Technologies (IoT, AI, ML)

The implementation of edge AI in space relies on leveraging advanced machine learning algorithms optimized for low-resource environments, supported by specialized hardware.

One critical area is the **Space Internet of Things (IoT)**. This paradigm envisions LEO satellites acting as communication hubs for billions of terrestrial IoT devices, such as environmental sensors, trackers, or agricultural monitoring systems. Edge intelligence is essential here because it enables the satellite to pre-process sensor data—filtering, summarizing, or running local decision logic—before the data is transmitted to the ground. This dramatically reduces network congestion and improves service speed for remote telemetry networks.

Successfully deploying ML/AI models at the edge requires significant algorithmic innovation. Techniques such as model compression (pruning, quantization) and carefully selected hardware (FPGAs, ASICs, low-power GPUs) are employed to balance computational performance with the severe resource usage limits.

2.4 Edge Intelligence Architectures and Components

Edge architectures in space are inherently heterogeneous. They combine highly reliable, often radiation-hardened Central Processing Units (CPUs) for fundamental flight control and telemetry with specialized co-processors or accelerators dedicated to AI workloads. These accelerators, such as Field-Programmable Gate Arrays (FPGAs) or custom Application-Specific Integrated Circuits (ASICs), provide the parallel processing capabilities needed for deep learning inference. The software frameworks used, such as OpenCL or tailored ML runtimes, are crucial for orchestrating tasks and managing the load across these disparate components.

A fundamental challenge in this area is ensuring the AI models fit and function within the tight memory and power envelopes. To achieve this, aggressive methods of model adaptation and compression must be utilized.

The necessity of model compression is not merely a technical optimization; it reflects a fundamental difference in AI development pipelines compared to terrestrial applications. While modern terrestrial AI focuses on maximizing performance scale and utilizing models with billions of parameters, space AI development is primarily driven by minimizing power consumption and mass. This power-aware engineering constraint requires model efficiency to be prioritized over peak theoretical performance. Aggressive algorithmic techniques, such as **Quantization-Aware Pruning (QAP)**, are essential because the computational complexity is dramatically sensitive to the precision (bit-width) of the parameters. For instance, a model can be compressed by factors of up to 25 times in terms of computational operations without significant loss in task accuracy by applying such methods. The enduring pressure to achieve maximal efficiency under minimal resource ceilings means that the space sector is positioned to become a significant driver for innovation in ultra-efficient, highly compressed, and intrinsically reliable AI hardware, potentially accelerating the development of specialized, low-power AI ASICs that can bridge the current performance-reliability gap.

2.5 Benefits of Edge Intelligence (Low Latency, Privacy)

The advantages of deploying intelligence directly at the edge are particularly pronounced in space environments:

- **Low Latency:** Decisions, alerts, or system adjustments can be generated instantly onboard, bypassing long round-trip delays, a benefit critical for autonomous navigation and emergency response.
- **Bandwidth Reduction:** Only processed results (e.g., summaries, detections, compressed maps) need to be downlinked, dramatically cutting the communication demand and cost associated with transmitting raw data. In a disaster scenario, a satellite can transmit only the damage map, saving crucial bandwidth.
- **Autonomy:** Spacecraft gain the resilience to operate and make mission-critical decisions independently, even if communication with Earth is lost or unstable.
- **Privacy and Security:** Sensitive raw data can be analyzed locally, reducing the risk of exposure during transmission to ground stations.

CHAPTER 3

SPACE EXPLORATION SYSTEMS OVERVIEW

3.1 History and Evolution of Space Exploration

Space exploration began with single-function, monolithic spacecraft relying heavily on ground command and control: the bent-pipe model. The evolution has been driven by the increasing sophistication of sensors and the need for greater operational throughput and resilience. The industry is now transitioning toward distributed, networked nodes with increasing computational capability, where autonomy is a core design requirement rather than an optional add-on.

3.2 Key Space Mission Components (Satellites, Rovers)

Space exploration platforms are highly diverse, each presenting unique computational and environmental challenges :

- **Observation Satellites:** Platforms in Low Earth Orbit (LEO) and Medium Earth Orbit (MEO) generate vast, multidimensional datasets (e.g., multispectral images, SAR radar scans) that require rapid filtering and processing.
- **Planetary Rovers/Landers:** Operating in deep space (Mars, Moon), these systems face long communication delays (often >10 minutes one-way) and must rely on high levels of autonomy for navigation, obstacle avoidance, and scientific sample selection.
- **Space Stations and Crewed Vehicles (ISS, Orion):** These platforms have more robust power availability and can host powerful computing experiments, such as Microsoft's Azure edge nodes and the MIT Spaceborne Computer, to test high-performance computing (HPC) concepts in orbit.
- **IoT Satellites:** Small LEO satellites serving as communication relays for billions of ground-based IoT sensors globally, demanding edge AI for efficient stream management.

Across all these systems, common constraints apply: limited onboard power (often <20 W for CubeSats), strict mass and volume budgets, and exposure to a harsh radiation environment.

3.3 Onboard Computing and Autonomy in Spacecraft

For robotic missions, onboard computing directly translates into enhanced safety and efficiency. Planetary vehicles, in particular, cannot rely on continuous human guidance. Edge intelligence enables autonomous navigation (AutoNav) through computer vision-based path planning and obstacle avoidance.

A prime example is the NASA Perseverance Mars Rover mission. The rover utilizes enhanced AutoNav and MLNav (Machine Learning Navigation) systems to perform real-time decision-making. By analyzing terrain images with its onboard computer, Perseverance can identify hazards and navigate around them, enabling the rover to drive autonomously for up to 88% of its total operational distance on Mars. Furthermore, AI systems like the Autonomous Exploration for Gathering Increased Science (AEGIS) autonomously select scientific targets based on analysis of images, deciding where to drill or study geological features, maximizing scientific return without constant Earth intervention.

3.4 Communication and Data Handling Systems

Communication systems are the choke point that edge computing seeks to mitigate. Traditional satellite links are constrained by low data rates, with modern LEO downlinks typically maxing out around 100 Mbps to 150 Mbps. For communications involving large bandwidth-delay products (like the 0.8-second round-trip delay common in satellite links), standard TCP protocols are inefficient. To maximize utilization, network protocols must implement specialized features like TCP-LW (large window extensions), which increases the maximum window size to 2^{32} to allow for efficient data flow across high-latency links.

A major enabling technology for modern space systems is the **Inter-Satellite Link (ISL)**, which is critical for satellite constellations and mesh networks. ISLs allow data to be routed through space from satellite to satellite, reducing reliance on ground stations. These links utilize both radio frequency (RF) and Free Space Optical (FSO), or laser communications. FSO links are becoming increasingly vital due to their high-speed, high-capacity, and secure transmission properties.

Managing ISLs introduces significant network complexity. Because satellites are constantly moving, links in the network (especially inter-plane links in polar regions) are not always available. This necessitates sophisticated, deterministic routing algorithms that can predict link shutdowns, reroute traffic (link handover), and optimize paths based on maximum coverage time and connection statistics to ensure low delay jitter and reliable communication across the network.

The fundamental shift from RF to FSO (lasercomms) for Inter-Satellite Links is a prerequisite for scaling complex, distributed edge AI applications like federated learning. Sharing massive models, like fine-tuned GeoFMs, or communicating the frequent, bandwidth-intensive gradient updates needed for Federated Learning (FL) requires enormous, reliable bandwidth capacity. FL studies show that using high-capacity ISLs for parameter aggregation is essential for mitigating intermittent ground connectivity and achieving faster model convergence. Therefore, FSO provides the massive, secure bandwidth needed for rapid, real-time synchronization, establishing FSO technology adoption not merely as an optional upgrade but as a non-negotiable enabling technology for sophisticated, distributed space AI.

3.5 Challenges of Traditional Space Exploration

The traditional approach, defined by the constraints of limited communication windows and ground-based processing, fundamentally restricts mission scalability and responsiveness. As sensors become more prolific, the volume of data generated quickly overwhelms the capacity of the downlink channel. This reliance on ground intervention for nearly all mission-critical decisions severely limits the responsiveness required for scenarios like orbital debris avoidance or real-time disaster monitoring, driving the industry toward autonomous, intelligent processing at the edge.

CHAPTER 4

INTEGRATION OF EDGE INTELLIGENCE IN SPACE SYSTEMS

Successfully integrating edge intelligence into spacecraft requires a complete reevaluation of the system architecture, moving away from simple data handlers toward heterogeneous, load-balancing compute nodes.

4.1 Rationale and Benefits of Edge Intelligence in Space

The core rationale for edge integration is the need for speed. In many time-critical scenarios, mission response must occur in minutes or seconds, a timeline impossible to meet if commands or processed data must travel from Earth (often involving multiple hours for ground-based processing and human review). Edge intelligence provides the necessary localized processing capability to close these decision loops quickly, maximizing operational efficiency and safety.

4.2 System Architecture for Space Edge Intelligence

The system architecture for space edge intelligence is typically hybrid and modular. It features the established, highly reliable **On-Board Data Handling (OBDH)** computer, which handles core spacecraft control, telemetry, and basic data formatting. To support AI, the OBDH is augmented by, and connected to, one or more powerful **Edge Computers**.

These Edge Computers are the dedicated AI acceleration modules. They host the specialized processors—CPUs, GPUs, FPGAs, and ASICs—that handle the high-throughput, energy-intensive AI workloads. This design creates a two-tiered system where the reliable OBDH manages satellite operations while the specialized edge module focuses on computational augmentation. The Edge Computer may receive data from the satellite’s local sensors or from neighboring satellites via Inter-Satellite Links (ISLs).

The physical modularity of these architectures is largely a reliability strategy that also facilitates rapid technology turnover. Reliable, radiation-hardened (rad-hard) components are typically slow and extremely expensive, whereas high-performance commercial off-the-shelf (COTS) components are fast but highly vulnerable to the harsh radiation environment. By maintaining separation, the mission-critical flight software, run by the certified rad-hard OBDH, is protected from potential failures or data corruption occurring in the experimental, high-performance accelerator modules. This architectural separation allows space agencies to adopt the "AI-First" design paradigm and test cutting-edge hardware components quickly, as only the specialized edge payload needs to be qualified, minimizing the impact of technology refresh cycles on core mission safety standards.

4.3 Data Processing and Workflow Onboard Spacecraft

The onboard workflow must be exquisitely managed to accommodate the constraints of the space environment. The most power-intensive AI tasks, such as running large neural networks, are often batched and executed during power-rich phases of the orbit, such as when the satellite's solar arrays are fully illuminated. Conversely, lower-power AI tasks (like basic health monitoring) may run continuously.

The OBDH's extended software stack must manage task scheduling, load distribution across the heterogeneous hardware, and the communication bus interfaces (e.g., PCIe, SpaceWire) connecting the main CPU to the accelerators. Systems often adopt a client-server model where an operator can uplink a configuration script specifying which AI kernels to execute on the available accelerators.

Software innovation is helping to bridge the gap between complex space hardware and accessible development. Specialized toolkits, such as the GPU@SAT DevKit, implement a General-Purpose GPU (GPGPU)-inspired parallel architecture on FPGA fabric, allowing developers to program the accelerators using familiar high-level parallel computing APIs like OpenCL. This greatly accelerates edge AI deployment by abstracting the underlying hardware complexity for application engineers.

4.4 Communication Protocols and Ground Integration

Integration allows for distributed processing across the entire network. The "computing out" approach enables an onboard satellite to offload parts of a heavy task (like processing a large neural network) to a nearby edge node satellite with spare capacity via an ISL. This strategy distributes the computational load across the network.

Furthermore, the integration trend is leading toward commercial models, exemplified by the Thales Alenia Space and Microsoft collaboration on the IMAGIN-E payload aboard the International Space Station (ISS). This testbed links commercial Azure edge servers to space payloads, exploring the concept of **Edge-as-a-Service**, where satellites effectively offer their compute resources to external users on demand.

4.5 Example Scenarios of Edge Integration in Missions

Several missions have demonstrated the feasibility of integrating high-performance computing (HPC) and AI components:

- **JAXA's GEMINI payload:** This mission flew consumer-grade GPU components to test their viability and resilience in orbit, pushing the boundaries of what is possible with

COTS hardware.

- **IMAGIN-E (ISS):** This platform has successfully hosted the first on-orbit demonstration of advanced foundation models (discussed in Section 9), validating the integration of modern, powerful co-processors into space systems under stringent orbital constraints.

These examples confirm that the technology is rapidly moving from conceptual design to active in-orbit testing, driven by the modularity of modern satellite architectures.

CHAPTER 5

EDGE AI ARCHITECTURES FOR SPACE EXPLORATION

Space edge AI architectures are defined by stringent requirements for radiation tolerance, extreme power efficiency, and physical compactness. They necessitate a combination of reliable, legacy components and high-throughput, customized accelerators.

5.1 Hardware Architectures (Radiation-Hardened Chips, GPUs, FPGAs)

Radiation and Reliability Requirements

All onboard computing hardware must be designed to withstand the harsh space environment. This involves mitigating the effects of Total Ionizing Dose (TID), which causes general component degradation, and Single-Event Upsets (SEUs), where cosmic rays can flip bits or cause momentary data corruption. This leads to a complex choice between highly reliable, but technologically lagging, radiation-hardened (rad-hard) chips, and faster, but less resilient, commercial off-the-shelf (COTS) components.

CPUs and SoCs

Radiation-hardened CPUs, often based on older ARM or POWER architectures (like the Space Micro Rad750), serve as the central control unit for flight software. While exceptionally reliable, these processors lack the parallel throughput required for modern deep learning models, limiting them to managing operations and running only lightweight inference tasks (e.g., decision trees).

FPGAs as High-Efficiency Accelerators

Field-Programmable Gate Arrays (FPGAs) are currently the preferred solution for AI acceleration in power-constrained space systems. They offer a crucial combination of parallel processing capability, reconfigurability, and inherent radiation tolerance (via Triple-Modular Redundancy—TMR). Leading space-grade FPGAs include the AMD RT Kintex UltraScale, the Virtex 5QV, and the Microsemi RTG4.

FPGAs are highly valued for their superior power efficiency. Comparative studies show that FPGAs deliver **1.2 to 22.3 times lower energy-per-frame** compared to state-of-the-art GPUs in embedded vision tasks. In general terms of floating-point operations per watt (GFLOPS/W), FPGAs are typically **three to four times better** than commercial GPUs, making them the most practical choice for systems with power envelopes often below 20 W.

GPUs and VPUs

Traditional commercial GPUs are generally unsuitable for space due to their high power consumption and susceptibility to radiation. However, testing initiatives, such as the ESA-led GPU4S, are exploring the viability of specialized COTS chips like Intel's Movidius Vision Processing Units (VPUs), which offer dedicated neural network acceleration at high efficiency. For customized applications, FPGAs are often programmed to emulate GPU-like parallel processing, as demonstrated by the GPU@SAT DevKit, which utilizes OpenCL programming kernels to run parallel compute tasks on FPGA fabric.

ASICs and Neuromorphic Processors

Application-Specific Integrated Circuits (ASICs) and Neuromorphic Processing Units (NPUs), like Google's TPU or Intel's Loihi research chips, represent the future of maximum efficiency. While not yet standard in space, they promise the highest performance-per-watt ratio due to their highly optimized architectures, provided they can be successfully radiation-hardened.

5.2 Software Frameworks (Operating Systems, Middleware, AI Platforms)

The software stack must manage the heterogeneity of the hardware. The onboard systems typically run specialized operating systems (like Linux or RTOS) that host drivers and runtimes for the various accelerators.

- **Abstraction and Portability:** Key frameworks aim to simplify development. The GPU@SAT DevKit provides an OpenCL interface, allowing developers to utilize familiar parallel computing paradigms on a customizable FPGA accelerator. Similarly, standard ML runtimes (TensorFlow Lite, OpenVINO) are cross-compiled and tailored for the specific space hardware.
- **Distributed Architecture:** For large satellite constellations, systems are moving toward sophisticated orchestration. Proposals include "microservice-empowered" architectures, where complex tasks, such as large model inference, are broken down into lightweight services that can be partitioned and run across different satellite nodes, dynamically managed by the onboard middleware.

5.3 Distributed and Federated Edge Architectures (Satellite Constellations)

Distributed intelligence is a necessity for growing satellite constellations. **Federated Learning (FL)** is emerging as a crucial paradigm. FL allows satellites to train models locally using their collected data and periodically share only the compressed model updates (gradients or parameters) over ISLs, rather than sending massive volumes of raw data back to Earth. This harnesses the collective intelligence of the network while dramatically reducing communication load and mitigating the impacts of intermittent connectivity. Research shows that using ISLs in FL can yield a seven-fold increase in convergence speed and a ten-fold reduction in communication bandwidth requirements compared to traditional methods.

5.4 Security, Reliability, and Fault-Tolerance in Space

The requirement for mission reliability dictates rigorous fault tolerance. This involves using Error-Correcting Code (ECC) memory, implementing watchdog timers to monitor system stability, and employing TMR in reconfigurable logic (FPGAs) to automatically correct radiation-induced bit flips.

Security must also be addressed, as edge devices can be vulnerable to tampering or malicious model uploads. Secure boot processes and strong authentication protocols for software updates are vital. Furthermore, for mission-critical functions, the AI accelerator must be compartmentalized so that in the event of an error or failure, the AI system does not interrupt or jeopardize the essential operations managed by the primary OBDH.

The performance bottleneck in space AI is fundamentally shifting from raw processing speed (FLOPS) to I/O efficiency and thermal management under severe power constraints. FPGAs, with their ability to be customized at the hardware level, provide the current optimal solution. FPGAs are preferred because their superior GFLOPS/W efficiency (3-4 times better than generic GPUs) allows them to meet strict thermal and power limits (e.g., <10 W) while still handling heavy computational loads. This customization allows engineers to precisely tailor the data path, computation units, and memory access, minimizing unnecessary data movement and power wastage—an essential design philosophy when processing gigapixel images, such as those from SAR sensors. This dominance of FPGAs confirms that generic, software-defined computing is suboptimal in space. The industry's reliance on FPGAs signals a commitment to hardware-defined efficiency, suggesting that the logical next step is the dedicated, radiation-hardened ASIC, which is the ultimate form of specialization.

CHAPTER 6

APPLICATIONS OF EDGE INTELLIGENCE IN SPACE

Edge intelligence is enabling new levels of capability across nearly every domain of space exploration by providing timely, localized data processing.

6.1 Autonomous Navigation and Guidance (Rovers, Landers)

For planetary missions, edge AI provides the real-time decision-making necessary to overcome communication delays. Rovers, such as Perseverance, rely on computer vision and deep learning models to identify obstacles, map terrain, and calculate optimal paths in milliseconds. Beyond navigation, AI systems support scientific decision-making, such as identifying a geological anomaly in a rock sample and autonomously commanding the rover to drill or take a high-resolution image, maximizing scientific data collection efficiency.

6.2 Onboard Data Processing (Earth Observation, Scientific Data)

Earth observation (EO) satellites benefit profoundly from onboard AI, transforming the massive raw data streams into actionable intelligence before transmission.

Real-Time SAR Analysis

A critical application is maritime surveillance using Synthetic Aperture Radar (SAR) imagery, which can penetrate clouds and darkness. By implementing custom deep learning models, such as YOLOv8 architectures optimized for FPGAs, satellites can detect vessels (e.g., tracking ships disabling their transponders). This onboard processing allows the system to analyze large 700-megapixel SAR images in under a minute with power budgets below 10 W, achieving results comparable to ground-based GPU processing. This capability dramatically speeds up alerts for illegal fishing or search-and-rescue targets.

Data Filtering and Reduction

Edge AI performs autonomous data analysis to filter out irrelevant information. This includes cloud detection and quality assessment in optical imagery, or flagging anomalous events like fires, floods, or deforestation as soon as the data is captured. This "Area of Interest" (AOI) identification ensures that only mission-relevant data products, not massive raw files, consume limited downlink bandwidth.

6.3 Spacecraft Health Monitoring and Fault Diagnosis

Spacecraft integrity is continuously monitored by onboard sensors tracking temperature, voltage, vibration, and radiation levels. Edge intelligence provides localized, real-time health assessments to predict and preempt subsystem failures.

#### Thermal Anomaly Detection

An important example involves FPGA-based thermal monitoring systems. These systems utilize an infrared sensor and a simple Machine Learning classifier, such as a Support Vector Machine (SVM), running directly on the FPGA to detect thermal anomalies or unusual vibrations that precede failures. By acting immediately, the satellite can autonomously shut down a faulty component or reconfigure itself without waiting for human analysis of delayed telemetry logs.

6.4 Satellite Constellation Coordination and Swarms

For multi-satellite systems or spacecraft swarms, distributed edge intelligence is necessary for coordination. AI supports complex tasks such as formation flying, where satellites share real-time position and sensor data via ISLs to maintain precise spacing. This capability is vital for creating virtual instruments, such as synthetic aperture radar arrays or distributed telescope networks, that require synchronized data acquisition and processing.

6.5 Communication Network Optimization (Inter-Satellite Links)

In LEO mega-constellations, edge AI dynamically optimizes the network itself. By analyzing real-time traffic and link quality, AI can manage network routing, prioritize data traffic, and allocate bandwidth across inter-satellite links, effectively providing "network intelligence" to ensure seamless connectivity and throughput.

The success of specific, narrowly-trained edge models, such as SAR vessel detection or thermal anomaly classification, provides the critical operational credibility needed to deploy larger, more generalized AI frameworks in orbit. Space missions demand absolute reliability, meaning that the verification and explainability of AI systems are paramount. Simple, specialized models are inherently easier to verify than complex, general-purpose models. The successful, low-power operation of these specialized models builds flight heritage and confidence in the underlying computational architecture—validating hardware resilience, power management, and software scheduling. This heritage of successful, low-complexity AI deployments is the necessary foundation for the eventual leap to highly sophisticated foundation models (GeoFMs) , allowing mission planners to confidently deploy generalized intelligence capable of complex semantic reasoning.

CHAPTER 7

CHALLENGES AND LIMITATIONS

Despite the transformative potential of edge intelligence, its application in space is severely constrained by persistent engineering and environmental challenges.

7.1 Harsh Environmental Conditions (Radiation, Temperature Extremes)

Radiation Effects

The space environment exposes electronics to high levels of radiation from cosmic rays and solar particles, leading to data corruption and component damage. This necessitates using radiation-hardened or radiation-tolerant hardware, which creates a significant performance gap: rad-hard chips are often 10 to 100 times slower than their commercial counterparts.

A major engineering trade-off exists regarding performance. High power efficiency requires manufacturing processors using small semiconductor nodes (e.g., 7nm), but these smaller nodes are intrinsically less resistant to radiation damage and more susceptible to stress due to lower operating and breakdown voltages. This forces designers into a constant compromise between maximizing computational performance and ensuring long-term reliability.

This engineering dichotomy has established two distinct markets in space AI: (1) low-cost, high-performance, short-lifespan constellations (which accept the risk associated with COTS components), and (2) high-cost, high-reliability, low-performance missions (like deep-space probes) that must rely on slower, legacy rad-hard technology. This performance gap can only be resolved by substantial investment in specialized, radiation-tolerant AI ASICs that overcome the intrinsic limitations of current small-node COTS manufacturing processes.

Thermal Variability

Spacecraft operate across extreme temperature ranges. Onboard systems must manage heat dissipation efficiently, as the components are often tightly packed and reliant on passive or limited active cooling, further limiting the clock speeds and power budget of processors.

7.2 Resource Constraints (Power, Mass, Size)

Space systems, especially CubeSats, operate under highly restrictive power, mass, and volume budgets. Computation is often limited to a few watts. This strict power envelope dictates that only highly optimized, small AI models can be deployed. To manage this, edge AI systems frequently use **power gating**—turning off accelerators when idle—and operate in short bursts during high-power periods (e.g., in full sunlight). The limitations on memory volume also restrict the ability to store large datasets or multiple large models simultaneously.

7.3 Communication Constraints (Latency, Bandwidth, Reliability)

While edge AI dramatically reduces the volume of *raw* data transferred, initial model deployment or subsequent updates still require significant downlink capacity. For deep learning, even a compressed model can be tens or hundreds of megabytes. Mission planning must account for these occasional large-volume transfers. Furthermore, intermittent communication links—especially when relying on ground stations—complicate the reliable coordination of distributed learning algorithms like Federated Learning. Although laser communications (FSO) improve data rates, the underlying network topology remains highly dynamic, requiring sophisticated, robust protocols.

7.4 Algorithmic and Data Limitations (Training Data, Robustness)

Edge AI models must be aggressively compressed to meet power and memory limits, often leading to a trade-off where some task accuracy is inevitably sacrificed. Ensuring that these lightweight models maintain mission-critical performance remains an ongoing challenge.
The complexity is compounded by the techniques used for optimization:

Algorithmic Compression Techniques for Space Edge AI

Technique	**Primary Mechanism**	**Benefit for Space Edge**	**Operational Example**
Quantization	Reduces parameter precision (e.g., 32-bit float to 8-bit integer)	Lower memory footprint, faster compute, less power	ViT Compression (GeoFM) inference on IMAGIN-e
Pruning	Removes redundant weights or connections	Reduced model size and computation, increased hardware density	Quantization-aware pruning (QAP) for maximum efficiency
Knowledge Distillation	Training a small 'student' model using outputs from a large 'teacher'	Maintains high accuracy with a dramatically smaller model	Key technique for adapting large foundation models
Neural Architecture Search (NAS)	Automated optimization of network structure	Custom model creation highly specific to low-power hardware (e.g., FPGA optimization)	

Furthermore, models trained on terrestrial data distributions frequently face a "domain shift" when deployed in orbit. For example, satellite imagery taken from high altitudes presents unique subtle differences in terrain and atmospheric clutter. Careful domain adaptation strategies are essential to maintain performance under real operational conditions.

7.5 Safety, Security, and Standardization Issues

For autonomous systems making mission-critical decisions, reliability and transparency are paramount. AI systems must be thoroughly verified, robust against adversarial attacks, and auditable. Verifying autonomous AI against all possible on-orbit anomalies is inherently complex. This demands strict adherence to Responsible AI (RAI) principles, such as those mandated by NASA (following EO 13960), which require AI systems to be transparent and accountable.

The security of the platform is also a concern. Edge devices, being remote, require robust protection against cyber threats, including secure boot mechanisms and cryptographically authenticated over-the-air updates to prevent the upload of malicious models.

CHAPTER 8

FUTURE TRENDS AND RESEARCH DIRECTIONS

The field of edge intelligence in space is characterized by rapid technological evolution, pointing toward highly autonomous and collaboratively intelligent systems.

8.1 Advances in Space-Grade AI Hardware (Neural Chips, Quantum)

The current reliance on FPGAs is transitional. The next major leap will be the development of dedicated, radiation-hardened **Neural Processing Units (NPUs)**. These ASICs will be optimized for the specific convolutional and matrix operations common in space algorithms, offering maximal performance-per-watt efficiency.

A transformative direction involves **Neuromorphic Computing**. This approach utilizes brain-inspired systems, such as spiking neural networks (SNNs) and neuromorphic chips (like Intel’s Loihi), which offer orders-of-magnitude greater energy efficiency than conventional digital architectures. Neuromorphic devices could enable "always-on" sensing capabilities with minimal power draw, essential for continuous monitoring and event-based learning. For instance, a neuromorphic vision sensor would only stream salient events (like movement or change), significantly reducing data rates and processing overhead.

8.2 Interplanetary Networking and Space Internet (6G, Laser Communications)

Free Space Optical (FSO) communication will become the standard for high-bandwidth Inter-Satellite Links (ISLs). This technology provides the necessary data rates to support the sophisticated data exchange required for distributed AI systems and federated learning.

Looking further ahead, **Space 6G** systems are being designed with edge intelligence integrated at multiple layers. Future satellites will host AI that jointly optimizes link routing, network resource allocation, and predictive maintenance across interconnected satellite and terrestrial (e.g., 5G) networks, creating seamless, intelligent global connectivity.

8.3 Autonomous Multi-Agent Space Systems (Swarm Coordination)

Future missions will increasingly involve large-scale swarms of small spacecraft or fleets of planetary drones. These require decentralized AI systems to manage coordinated maneuvers and decision-making without a single point of control. Research focuses on developing sophisticated, distributed learning algorithms where each satellite holds part of a neural network ensemble and collectively makes complex decisions in real time.

8.4 AI for Human Spaceflight and Life Support Systems

In crewed exploration, AI will manage complex, dynamic, closed-loop systems. This includes smart habitat monitoring, managing autonomous life support systems (optimizing air, water, and power recycling), and assisting in complex tasks like AI-guided 3D printing and manufacturing operations on orbit or on planetary bases.

8.5 Regulatory, Ethical, and Collaborative Research Initiatives

As AI proliferates, standardization efforts are crucial. There is a pressing need for common APIs, open frameworks, and standardized data formats (such as ONNX adapted for space use) to accelerate deployment and ensure software portability across heterogeneous satellite architectures. Furthermore, defining clear boundaries for onboard autonomy and establishing rigorous testing and evaluation protocols are necessary to ensure the safe, transparent, and ethical operation of AI systems.

8.6 Large Foundation Models in Space

The most significant trend is the deployment of **Large Foundation Models (GeoFMs)**, such as compressed Vision Transformers, in orbit. These models, pretrained on vast amounts of terrestrial data, offer generalized intelligence for diverse tasks (object detection, classification, semantic segmentation).

Future research focuses on **distributed fine-tuning**, leveraging Federated Learning to update and adapt these massive models across satellite constellations. Researchers are exploring architectures where a large model is split into lightweight services running on different satellites and ground nodes, allowing for dynamic updates without massive, raw data transfers.

Federated Learning (FL) is not solely an ML algorithm; in the context of space networks, it functions as a critical communication and data management protocol that fundamentally redefines how data value flows between Earth and space. By using ISLs to transmit aggregated, compressed parameters, FL bypasses the severe ground communication bottleneck, achieving substantial reductions in bandwidth utilization. This means satellites are no longer passive data collectors but active, distributed intellectual assets collectively contributing to a continuously improving, generalized intelligence product. This necessity for collaborative, in-network learning will compel the development of new legal and contractual frameworks governing shared data ownership and synchronized model intellectual property among competing or allied space actors.

CHAPTER 9

CASE STUDIES

Recent demonstrations highlight the rapid operational deployment of edge intelligence in space, moving from theoretical concept to verified capability.

9.1 NASA Perseverance Mars Rover (Onboard AI for Geology)

The Perseverance rover stands as a key example of mission-critical edge intelligence. Utilizing its Enhanced AutoNav and MLNav systems, the rover achieves exceptional autonomy for deep space exploration. The rover autonomously navigates around obstacles, avoids hazards, and identifies scientific targets, executing these decision loops in real time despite communication delays that can exceed ten minutes. This operational success proves the indispensable role of robust, reliable edge AI in guaranteeing mission continuity and safety on planetary surfaces.

9.2 On-Orbit Geospatial Foundation Model (IMAGIN-e)

The IMAGIN-e (ISS Mounted Accessible Global Imaging Nod-e) payload is a state-of-the-art testbed resulting from a collaboration involving the European Space Agency (ESA), Thales Alenia Space, and Microsoft Azure.

Technical Breakdown

Researchers successfully demonstrated the first on-orbit execution of a Geospatial Foundation Model (GeoFM)—a compressed variant of a Vision Transformer (ViT). Foundation models are highly complex, pretrained architectures, previously deemed too large for space hardware.

Achievement

By applying aggressive model compression (quantization and pruning) and domain adaptation techniques, the GeoFM was scaled down to operate reliably within the strict power and memory limits of the IMAGIN-e payload. Despite technical anomalies affecting the onboard imager, a curated test set was uploaded to the compute module, and the model reliably executed inference, successfully classifying land cover, roads, and buildings.

The successful execution of this compressed GeoFM on the ISS validates the operational effectiveness of algorithmic mitigation techniques and provides critical flight heritage. This milestone validates that sophisticated, generalized AI, capable of complex semantic reasoning (classifying roads or water bodies), can operate within genuine orbital constraints, fundamentally shifting the conversation from *if* complex deep learning can be deployed to *how widely* it can be scaled.

9.3 FPGA-based SAR Vessel Detection (Planetek AI-eXpress Satellite Constellation)

This case study demonstrates the current peak efficiency achievable using FPGAs for high-throughput Earth observation.

Implementation and Results

Laganier et al. (2025) designed a customized YOLOv8 deep learning model specifically for detecting ships in Synthetic Aperture Radar (SAR) imagery. They optimized this model to run on an FPGA-based System-on-Module. The results were compelling: the FPGA implementation successfully processed large 700-megapixel SAR images in under one minute, operating within a very tight power budget of less than 10 W. Critically, the model's accuracy remained high, dropping only 2–3% compared to a full GPU reference implementation, while being exponentially more computationally efficient.

This quantitative success validates FPGAs as the superior platform for power-constrained, high-throughput sensing applications in space, enabling autonomous maritime surveillance and near-real-time threat alerts.

9.4 Specialized Hardware Development (GPU@SAT DevKit) and Thermal Monitoring

GPU@SAT DevKit

The GPU@SAT DevKit exemplifies engineering innovation aimed at streamlining space AI development. It utilizes a commercial FPGA (Xilinx Zynq UltraScale+) to implement a GPGPU-inspired parallel core. This modular testbed allows engineers to prototype and test high-performance AI algorithms using standard OpenCL programming on hardware that closely mirrors the architecture of flight-qualified systems. This accelerates the development cycle by bridging the gap between familiar terrestrial software tools and specialized space hardware.

Thermal Anomaly Detection Payload

Moreira et al. (2025) demonstrated a miniaturized, plug-and-play edge computing payload designed for spacecraft health monitoring. The payload comprises a micro-thermal infrared camera feeding data to an FPGA board running an SVM classifier. This system autonomously identifies simulated heat faults (overheating components) in real time on a satellite electronics board. This case highlights the feasibility of creating modular, reliable AI agents dedicated to onboard diagnostics, which can be easily added to existing satellite systems to enhance fault diagnosis and preventative maintenance.

CHAPTER 10

CONCLUSION

10.1 Summary of Key Findings

Edge intelligence is no longer an aspirational concept but a mandatory technological component for scaling future space missions. The limitations inherent in traditional space architecture—high latency, limited bandwidth, and long communication delays—are effectively mitigated by performing AI inference directly on the spacecraft.

Current operational success relies heavily on **heterogeneous architectures**, where highly reliable CPUs manage flight control and robust, energy-efficient FPGAs handle the parallel AI acceleration. This preference for FPGAs is driven by their superior power-to-performance ratio (3–4 times better GFLOPS/W than generic GPUs) and their intrinsic radiation tolerance. Algorithmic efficiency is achieved through essential model compression techniques, such as quantization and pruning, which are necessary to ensure models operate within the strict power budgets (often below 10 W). Recent case studies, including the on-orbit demonstration of compressed Geospatial Foundation Models and high-efficiency SAR vessel detection, confirm that sophisticated AI is now viable in space.

10.2 Implications for Future Space Missions

The future of space systems will be defined by autonomy and collaboration. The shift toward mega-constellations demands distributed intelligence, facilitated by **Federated Learning (FL)** protocols and high-bandwidth **Free Space Optical (FSO)** inter-satellite links. FL is particularly transformative, acting as a data management protocol that minimizes communication load by replacing the transfer of raw data with the synchronization of compressed model parameters.
Looking forward, the development of specialized, radiation-hardened **neuromorphic chips** (SNNs) will revolutionize the power efficiency of onboard systems, potentially enabling continuous, always-on AI monitoring and sensing capabilities that were previously unattainable under tight power constraints.

10.3 Contributions of Edge Intelligence to Exploration

The primary contribution of edge intelligence is the enhancement of situational awareness—both of Earth and the spacecraft itself—in near-real time. It maximizes the scientific return of data by autonomously filtering and prioritizing high-value information, and, critically, it ensures the safety and operational continuity of mission assets, particularly in deep space environments where human intervention is impossible in real time. The successful deployment of generalized AI models (GeoFMs) confirms that future robotic explorers will be capable of complex semantic reasoning, moving beyond simple classification tasks.

10.4 Final Remarks

Edge intelligence is advancing at an explosive pace, driven by necessity and enabled by key architectural and algorithmic innovations. The continued maturation of dedicated space-grade hardware, coupled with robust verification and standardization efforts, will be crucial. The outcome of this engineering focus is clear: the future space explorer—whether an orbiting satellite or a planetary rover—will be inherently "smart," operating with greater autonomy, learning collaboratively, and acting immediately on the data it gathers.

11. References

1. Benelli, G. et al. (2024). GPU@SAT DevKit: Empowering Edge Computing Development Onboard Satellites in the Space-IoT Era. Electronics, 13(19), 3928.
2. BERTEN DSP S.L. (2016). GPU vs FPGA Performance Comparison (White Paper).
3. Chenet, D. et al. (2025). Space Edge Computing for Satellite Systems: Definition and Key Enabling Technologies. In Euro-Par 2024: Parallel Processing Workshops (LNCS vol. 15386, pp. 399–404).
4. Du, A. et al. (2025). First On-Orbit Demonstration of a Geospatial Foundation Model (arXiv preprint 2512.01181).
5. Hawks, B., Duarte, J., Fraser, N. J., Pappalardo, A., Tran, N., & Umuroglu, Y. (2021). Ps and Qs: Quantization-Aware Pruning for Efficient Low Latency Neural Network Inference. Frontiers in Artificial Intelligence, 4, 676564.
6. He, R., Han, D., Shen, X., Han, B., Wu, Z., & Huang, X. (2025). AC-YOLO: A lightweight ship detection model for SAR images based on YOLO11. PLoS One, 20(7), e0327362.
7. Izzo, D. et al. (2022). Neuromorphic Computing and Sensing in Space (arXiv preprint arXiv:2212.05236).
8. Laganier, C. et al. (2025). Efficient SAR Vessel Detection for FPGA-Based On-Satellite Sensing. Proceedings of the ACM/IEEE Symposium on Edge Computing (SEC).
9. Moreira, C. M. et al. (2025). Edge Computing in Space: Design of an FPGA Architecture for Thermal Anomaly Detection Based on a Machine Learning Approach. Advances in Space Research.
10. National Aeronautics and Space Administration (NASA). (2024). AI and Machine Learning Use Cases. [Organization website/report].
11. Razmi, N., Matthiesen, B., Dekorsy, A., & Popovski, P. (2024). On-Board Federated Learning for Satellite Clusters with Inter-Satellite Links. IEEE Transactions on Communications, 72(6), 3408–3424.
12. Sadiku, M. N. O. et al. (2024). Artificial Intelligence in Space Exploration. International Journal of Trend in Scientific Research and Development (IJTSRD), 8(6), 464–473.
13. Shi, Y. et al. (2025). Satellite edge artificial intelligence with large models: architectures and technologies. Science China Information Sciences, 68(170302).
14. The Internet Society (Isoc). (1997). High-Speed TCP and Performance over Satellite Links. [Organization report/paper].

Works cited

1. Edge Computing Use Cases for Space Applications | VORAGO Technologies, https://www.voragotech.com/blog/edge-computing-use-cases-in-space-applications
2. Onboard Data Processing - eoPortal,

https://www.eoportal.org/other-space-activities/onboard-data-processing

3. Ps and Qs: Quantization-Aware Pruning for Efficient Low Latency Neural Network Inference, https://pmc.ncbi.nlm.nih.gov/articles/PMC8299073/
4. Revisiting Edge AI: Opportunities and Challenges - IEEE Computer Society, https://www.computer.org/csdl/magazine/ic/2024/04/10621659/1Z5lGDb639C
5. AI on the Edge of Space | Center for Security and Emerging Technology, https://cset.georgetown.edu/publication/ai-on-the-edge-of-space/
6. Artificial Intelligence - NASA, https://www.nasa.gov/artificial-intelligence/
7. NASA's AI Use Cases: Advancing Space Exploration with Responsibility, https://www.nasa.gov/organizations/ocio/dt/ai/2024-ai-use-cases/
8. The Internet from Space: RFIC Advances in High Capacity Low Latency LEO Satellite User and Ground Terminals | Analog Devices, https://www.analog.com/en/resources/analog-dialogue/articles/internet-from-space-rfic-in-high-capacity-low-latency-leo-satellite-terminals.html
9. Satellite Communications in the Global Internet: Issues, Pitfalls, and Potential, https://www.isoc.org/inet97/proceedings/F5/F5_1.HTM
10. Satellite to satellite communication - Blog - Satsearch, https://blog.satsearch.co/2025-03-14-satellite-to-satellite-communication
11. Free-space optical communication - Wikipedia, https://en.wikipedia.org/wiki/Free-space_optical_communication
12. Advanced Routing Protocols for Satellite and Space Networks Chao Chen - Georgia Institute of Technology, https://repository.gatech.edu/bitstreams/82862bd3-b6db-490f-8e34-10d7c7496d9b/download
13. On-Board Federated Learning for Satellite Clusters With Inter-Satellite Links - IEEE Xplore, https://ieeexplore.ieee.org/document/10409275
14. AI-Enabled Onboard Edge Computing for Satellite Intelligence in Disaster Management, https://www.un-spider.org/news-and-events/news/ai-enabled-onboard-edge-computing-satellite-intelligence-disaster-management%C2%A0
15. Hardware Platforms Enabling Edge AI for Space Applications: A Critical Review - IEEE Xplore, https://ieeexplore.ieee.org/iel8/6287639/10820123/11115053.pdf
16. 16. Pushing Intelligence to the Edge: Satellogic's Vision for AI-Powered Earth Observation, https://satellogic.com/2025/03/20/pushing-intelligence-to-the-edge-satellogics-vision-for-ai-powered-earth-observation/
17. GPU@SAT, the AI enabling ecosystem for on-board satellite applications - IEEE Xplore, https://ieeexplore.ieee.org/document/10396289/
18. OrbitalAI IMAGIN-e - Φ-Lab Challenges, https://platform.ai4eo.eu/orbitalai-imagin-e
19. Review on Hardware Devices and Software Techniques Enabling Neural Network Inference Onboard Satellites - MDPI, https://www.mdpi.com/2072-4292/16/21/3957