The Space Edge AI Computer: A Decoupled, Zero-Virtualization Payload for In-Orbit Generative AI under 12W

Anonymous CVPR submission

Paper ID 32

Abstract

Deploying advanced cognitive capabilities on nanosatellites faces extreme Size, Weight, and Power (SWaP) constraints, cosmic radiation, and software rigidity. Current solutions relying on power-hungry data-center GPUs, heavy cloud-native virtualization, or rigid monolithic cold-spares are physically and operationally unviable for 10W-class nodes. In this paper, we propose the Space Edge AI Computer (EAC), a deeply decoupled 12W/703.5g COTS-centric payload engineered for the nanosatellite SWaP-C sweet spot. Hardware-wise, its physical M.2 decoupling allows for agile pre-launch accelerator swaps to resist chip obsolescence, while a mixed-reliability, Open-Collector (OC)-driven intervention mechanism provides granular intra-module storage switching to prevent entire-board failure. Software-wise, we design AIOS, a zero-virtualization spaceborne middleware that eschews heavy containers for OS-level POSIX process derivation, achieving extreme reconfiguration elasticity over weak telemetry links. We report a flawless initial 124-day bare-LEO flight validation campaign. Our in-orbit evaluations validated the first bare-LEO configurable SLM and RAG architecture for medical consultation (achieving a 7.7× latency reduction), concurrent heterogeneous vision pipelines via a strict isolated-frame protocol, and a pioneering feasibility study of an offline Chain-of-Thought (CoT) reasoning model (DeepSeek-R1-Distill) under strict power envelopes. This establishes a pragmatic, flight-validated foundation for future distributed spaceborne autonomous agents.

1. Introduction

The paradigm of space exploration is rapidly shifting towards Spaceborne Edge Computing (SEC) to alleviate the unprecedented telemetry bottlenecks caused by the orbital data deluge [9, 18], and to enable true spacecraft autonomy [23, 24]. Consequently, deploying Artificial Intelligence (AI) accelerators and Foundation Models to the space edge has become a critical frontier. However, deploying these capabilities in Low Earth Orbit (LEO) presents a persistent SWaP-C (Size, Weight, Power, and Cost) trilemma and extreme software rigidity, which fundamentally restricts dynamic on-orbit cognitive reconfiguration [15].

Current high-compute solutions either pursue massive performance using data-center GPUs (20W–70W+) exceeding nanosatellite budgets, or employ highly integrated monolithic SoCs facing severe chip obsolescence. To combat Single Event Upsets (SEUs), they predominantly rely on heavy board-level cold-spares, causing prohibitive resource waste. Furthermore, agile in-orbit updates typically demand heavy cloud-native containerization (e.g., Docker), incurring a severe "virtualization tax" on constrained nodes. Lastly, pioneering space LLM missions are often confined to pressurized, power-rich ISS environments; deploying offline cognitive architectures on independent 10W-class bare-LEO nodes remains a critical gap.

To address these coupled challenges, we design, implement, and validate the Space Edge AI Computer (EAC). Operating at a typical power of 12W and weighing 703.5g, it leverages a deeply decoupled architecture—pairing an industrial COTS low-power SoC with an M.2 COTS AI accelerator to deliver up to 32 TOPS. The payload was launched into LEO in 2025 and successfully completed its initial 124-day active validation campaign without a single catastrophic failure.

In summary, our main contributions are three-fold:

- **Decoupled Hardware & Mixed-Reliability:** We engineered the COTS-based EAC targeting the nanosatellite SWaP-C sweet spot. Its physical M.2 decoupled architecture enables agile pre-launch accelerator updates to combat chip obsolescence. Concurrently, an OC-driven hardware intervention mechanism provides granular intra-module fallback boot selection, preventing entire-board obsolescence without heavy cold-spares.
- **Zero-Virtualization Middleware (AIOS):** We developed an ultra-lightweight spaceborne OS that utilizes native POSIX process derivation instead of heavy containers, achieving extreme dynamic reconfiguration elastic-

ity and deterministic isolation over weak communication links.

- **124-Day In-Orbit Validation:** We report a flawless flight heritage, empirically validating the first domain-specific medical SLM with a configurable RAG architecture on a 12W bare-LEO nanosatellite (7.7× latency reduction), concurrent vision pipelines, and, to our knowledge, the first in-orbit offline Chain-of-Thought (CoT) deployment using DeepSeek-R1-Distill.

2. Related work

2.1. Spaceborne Edge Computing Hardware

The transition from traditional Radiation-Hardened By Design (RHBD) processors—which offer extreme reliability but typically deliver sub-TOPS performance—to Commercial-Off-The-Shelf (COTS) AI accelerators has redefined spaceborne computing. Recent missions have aggressively pushed the computational boundaries in orbit. For instance, the Sidus FeatherEdge Gen-2 payload leverages space-hardened NVIDIA Jetson Orin NX modules to deliver 100 sparse INT8 TOPS [3], while computing payloads such as the Xingxi 04Pro (275 TOPS) [27] and the recent Three-Body Computing Constellation nodes (744 TOPS peak per node) [1] deliver massive computational peaks. At the extreme end, the Starcloud-1 mission recently launched an entire 60-kg satellite dedicated to operating a data-center-grade NVIDIA H100 GPU [22].

While pioneering missions like Φ-Sat-1 and Φ-Sat-2 successfully demonstrated low-power modular COTS AI (e.g., Intel VPUs) for specific vision tasks [10, 16], the recent trend of 'brute-forcing' compute performance highlights a persistent SWaP-C struggle. Data-center GPUs and highly integrated payloads typically demand 20W–70W+ and weigh substantially (¿1.5kg for thermal and radiation shielding), which is prohibitive for strictly constrained nanosatellites. In contrast, our proposed EAC targets the nanosatellite SWaP "sweet spot," delivering 32 TOPS of heterogenous compute at a typical 12W power draw and 703.5g mass. Furthermore, unlike deeply integrated payloads, the EAC adopts a deeply decoupled M.2 architecture, enabling agile pre-launch accelerator swaps to future-proof the compute node.

2.2. Spaceborne Software Resilience and Architectures

Cosmic radiation, notably Single Event Upsets (SEUs), poses a critical threat to COTS memory and computational integrity. The aerospace community is currently divided into two predominant architectural factions to mitigate these risks while managing application updates.

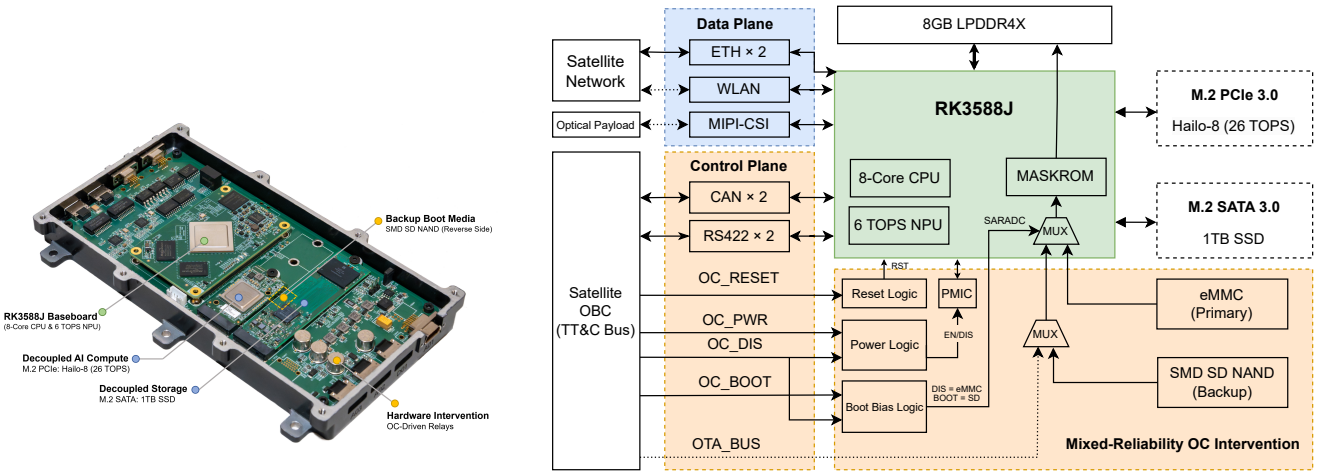
The first faction advocates for "heavy cloud-native" orchestrations. Systems like the Tiansuan Constellation and the ISS-based SBC-2 rely on containerization (e.g., Docker or KubeEdge) to achieve agile service discovery and isolation [7, 26]. However, the virtualization tax and persistent memory footprint of container engines are exceptionally detrimental to power-constrained edge nodes [11, 12]. The second faction relies on "deeply customized embedded" or highly integrated SoC designs, such as the Leopard DPU (Zynq UltraScale+) and the CogniSAT-XE2 (Myriad X VPU) [21, 29]. While power-efficient, these architectures are often monolithic, rendering dynamic in-orbit logic reconfiguration cumbersome and inflexible.

Deviating from both extremes, our work introduces AIOS: a zero-virtualization, runtime-reconfigurable spaceborne system. By utilizing native-process orchestration rather than rigid monolithic firmware or heavy container engines, we achieve dynamic, on-the-fly application updates with a fraction of the memory footprint and processing overhead associated with traditional cloud-native stacks. Additionally, we lay the structural groundwork for system-level fault tolerance through a natively designed multi-level storage redundancy mechanism (eMMC and SD NAND).

2.3. Generative AI and Foundation Models in Orbit

The deployment of Large Language Models (LLMs) and Foundation Models to the space edge represents the latest frontier in spacecraft autonomy. Pioneering efforts such as the "Space Llama" project (a modified Llama 3.2 model) and the ASTREA project (Qwen 2.5) have successfully demonstrated generative AI and agentic reasoning in space [4, 17]. However, these milestones were achieved onboard the International Space Station (ISS), benefiting from a pressurized, temperature-controlled, and power-rich "greenhouse" environment.

Deploying LLMs on independent, bare-LEO satellites has proven far more challenging. Although recent studies extensively explore LLMs for autonomous satellite operations, task scheduling, and network routing [6, 13, 28], these architectures remain predominantly theoretical or rely on ground-based agents. While the aforementioned Starcloud-1 achieved in-orbit LLM training and inference, it essentially functions as a flying data center rather than an edge node. Other recent breakthroughs, such as the deployment of the Qwen-3 model in the Xingsuan Constellation [2], mark significant progress. Yet, these deployments predominantly rely on massive platforms with generous power budgets to accommodate large parameter counts. For strictly constrained (e.g., 10W-class) independent bare-LEO nodes, relying solely on parametric scaling is physically unviable. Consequently, there remains a critical gap in exploring how Small Language Models (SLMs), when synergistically coupled with configurable Retrieval-Augmented Generation (RAG), can deliver reliable, domain-specific cognitive capabilities without exceeding strict SWaP envelopes.



(a) Physical board highlighting the decoupled M.2 expansion slots.

(b) Logical architecture illustrating the isolation of control/data planes and the OC-driven storage switching mechanism.

Figure 1. Hardware architecture of the Space Edge AI Computer (EAC).

Our 124-day in-orbit evaluation provides the foundational empirical data to bridge this gap.

3. System Architecture

Traditional spaceborne electronics typically adopt rigid, monolithic architectures heavily reliant on customized backplanes, which severely limits both hardware iteration and payload scalability. To address the unique challenges of Low Earth Orbit (LEO)—tight Size, Weight, and Power (SWaP) constraints, severe radiation, and dynamic mission requirements—we introduce the Space Edge AI Computer (EAC). Rather than a monolithic payload, the EAC is architected from the ground up as a network-ready, deeply decoupled computing node. It is designed to function either as an independent, high-performance powerhouse for nanosatellites or as a seamless, plug-and-play unit within a distributed space computing network.

3.1. Hardware Architecture: The SWaP-C Sweet Spot and Decoupled Resilience

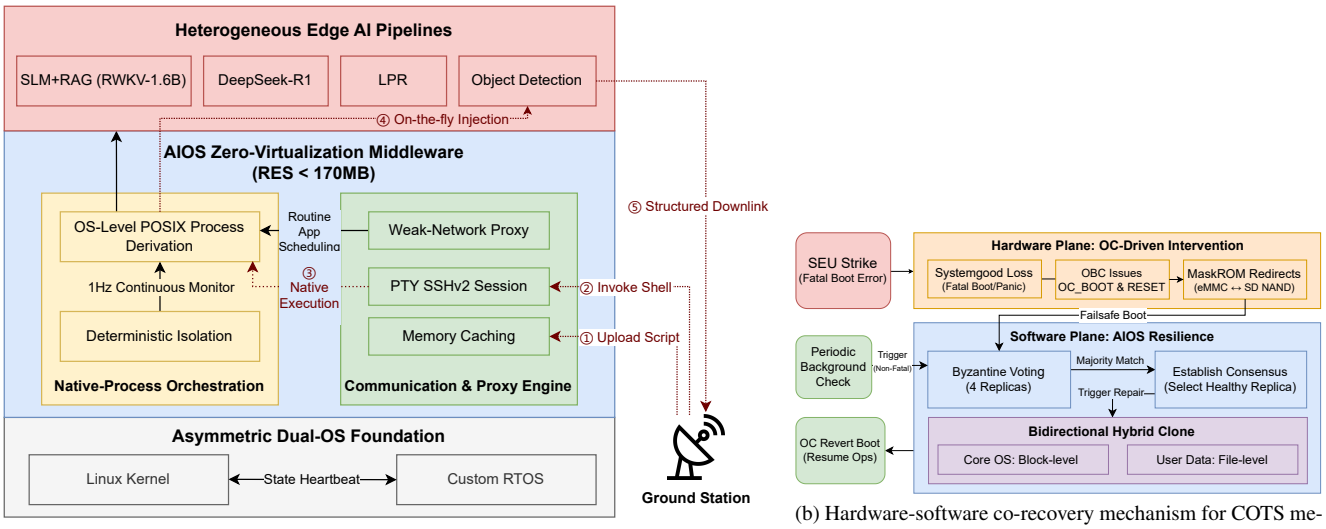
The physical design of the EAC breaks away from traditional rigid integration. As detailed in Table 1 and Figure 1, the system employs a multi-dimensional decoupling strategy across compute, communication, and fault tolerance, successfully targeting the elusive SWaP-C (Size, Weight, Power, and Cost) sweet spot for nanosatellites.

Balancing Compute and Constraints (The COTS SoC and M.2 Decoupling): Selecting hardware for LEO edge computing typically presents a dilemma: data-center-grade Edge GPUs deliver immense compute but demand 20W to 60W of power—prohibitive for nanosatellites—while ultra-low-power FPGAs or VPUs lack the capacity for modern foundation models. To resolve this, the EAC utilizes a highly integrated industrial COTS Internet-of-Things (IoT) SoC (RK3588J). Crucially, this SoC natively supports PCIe 3.0, SATA 3.0, and Dual Gigabit Ethernet without requiring external bridge chips. This integration condenses the footprint to under half the size of an A4 sheet and constrains typical power to 12W. To endure the vacuum thermal environment under an 18.7W peak load, the system establishes a passive thermal pathway linking the silicon directly to the aluminum chassis via carbon-fiber-filled gap pads. Furthermore, while this specific silicon lacks traditional flight heritage, the extensive LEO flight heritage of ARM-based COTS processors instills confidence in its orbital survivability.

Table 1. EAC Hardware Specifications.

Component / Metric	Specification
Main SoC (COTS)	Rockchip RK3588J (8-core ARM)
Base AI Compute	6 TOPS NPU (INT8) + Mali-G610 GPU
Decoupled Expansion	M.2 PCIe 3.0 (26 TOPS) & M.2 SATA 3.0
Memory & Storage	8 GB LPDDR4X / 64 GB eMMC + 1 TB SSD
Backup Boot Media	64 GB SMD SD NAND + Hardware MUX
Data Plane Interfaces	ETH*2, MIPI-CSI, WLAN
Control Plane Interfaces	CAN*2, RS422*2, OC Commands, Systemgood
Power Consumption	12 W (Typical) / 18.7 W (Peak)
Mass & Dimensions	703.5 g, 208 × 127 × 24 mm

To combat the severe discrepancy between multi-year satellite development cycles and rapid AI chip iteration, the EAC employs a deeply decoupled architecture. Rather than a monolithic board, the baseboard features two distinct M.2 expansion slots. In our 124-day flight mission, the M.2 PCIe 3.0 slot housed a 26-TOPS Hailo-8 AI accelerator, while the M.2 SATA 3.0 slot accommodated a 1TB SSD. This physical decoupling allows aerospace engineers



(a) AIOS zero-virtualization architecture for agile payload reconfiguration.

(b) Hardware-software co-recovery mechanism for COTS media repair.

Figure 2. The AIOS software stack and cross-layer fault tolerance.

to swap in the latest generation of compute or storage cards mere months before launch without redesigning the complex, radiation-tested baseboard.

Constellation-Ready Interface Design: Anticipating the transition towards distributed spaceborne clusters, the EAC strictly separates the control and data planes to prevent AI workloads from congesting telemetry buses. The Control Plane relies on dual-redundant CAN and RS422 interfaces for high-priority Telemetry, Tracking, and Command (TT&C). Conversely, the Data Plane is provisioned with Dual Gigabit Ethernet for high-bandwidth, intra-satellite large file distribution (e.g., model weights), MIPI interfaces for zero-copy sensor data ingestion, and WLAN modules providing electrically isolated, cable-free backup links. Furthermore, while auxiliary I/O (UART, GPIO, PPS) and an OTA bus are fully implemented for future multi-node system, they remained unexposed in this mission due to satellite-level constraints. These pre-integrated interfaces ensure the hardware is inherently "constellation-ready" to support future multi-node compute pooling.

Mixed-Reliability Hardware Fault Tolerance: The most critical threat to COTS deployment in space is Single Event Upsets (SEUs) corrupting primary boot media (e.g., eMMC) [14, 25], rendering software-level watchdogs useless. To establish an unbreachable physical safety net, the EAC utilizes a "mixed-reliability" architecture. While the compute core relies on COTS silicon, the ultimate hardware reset circuit employs aerospace-grade magnetic latching relays (governing BOOT and PWR states) and DC relays (for RESET and DIS triggers).

Driven by four Open-Collector (OC) commands directly from the satellite's core bus, these relays bypass the operating system entirely. Upon detecting a catastrophic boot failure (indicated by the loss of the physical *Systemgood* pin state), the satellite bus issues OC pulses that physically alter the SARADC input voltage on the SoC. This forces the lowest-level MaskROM to sever the boot link to the corrupted eMMC and redirect the boot sequence to a 64GB surface-mount industrial SD NAND (utilizing SMD soldering rather than a mechanical slot to ensure vibrational resilience).

Crucially, this toggle mechanism is bidirectional: after booting from the backup SD NAND, the system can repair the corrupted eMMC in-orbit, and subsequently be commanded via OC pulses to revert to the primary eMMC. In the extreme scenario where both internal media fail, a hardware Multiplexer (MUX) can switch the boot interface to an external OTA bus. This ensures a highly reliable, zero-compute mechanical relay acts as the ultimate guardian for the COTS brain. Unlike traditional architectures relying on heavy board-level redundancy (e.g., switching to a completely separate cold-spare payload) upon a single memory failure, this granular intra-module recovery prevents entire-board obsolescence.

3.2. AIOS Kernel & Resilience: Dual-OS Pathfinder and Hardware-Level Recovery

Space radiation poses a severe threat to unmodified COTS processors. Rather than relying on rigid, radiation-hardened components, we developed the AI Operating System (AIOS)—a deeply customized spaceborne OS distribution tailored for software-defined COTS resilience.

Hardware-Level Boot Intervention: Beyond standard watchdogs, AIOS relies on the baseboard's hardware-level response to the four OC commands. While the SoC's MaskROM physically alters the initial boot source via OC-

driven SARADC voltage changes, default COTS bootloaders often exhibit hardcoded medium preferences. To prevent these from overriding the physical selection, AIOS employs a deeply customized U-Boot SPL, ensuring a deterministic, end-to-end hardware-driven boot switch (between eMMC and SD NAND) completely independent of the EAC's software state.

Asymmetric Dual-OS Pathfinder: AIOS is engineered upon an asymmetric dual-OS architecture to separate the highly reliable monitoring plane from the data-intensive computing plane. A deeply customized Linux kernel handles complex AI tasks, while a deeply customized Real-Time OS (RTOS) runs concurrently on a dedicated low-power core. Crucially, with these external I/O channels physically unexposed for this flight, the RTOS currently functions as an isolated monitoring island, exchanging state heartbeats with the Linux kernel. This 124-day concurrent deployment successfully validates the stability of the dual-kernel foundation in space, acting as a flight-proven pathfinder for our future constellation phase.

Byzantine Self-Healing and Hybrid Clone: To combat silent data corruption in COTS memory, AIOS maintains four redundant OS replicas across heterogeneous mediums (eMMC and SSD), governed by a Byzantine fault-tolerant voting mechanism using pre-calculated checksums. Should the primary eMMC fail catastrophically, the hardware triggers a boot from the SMD industrial SD NAND backup. AIOS then initiates a "hybrid clone" reverse repair strategy: utilizing block-level `dd` for critical OS partitions and file-level `rsync` for massive user data. This significantly accelerates in-orbit recovery while preserving the limited erase-cycle lifespan of the COTS flash memory.

3.3. AIOS Middleware: Zero-Virtualization Orchestration and Agile Reconfiguration

Building upon the aforementioned hardware-software co-recovery foundation, AIOS incorporates a zero-virtualization middleware stack engineered to achieve agile reconfiguration over severe space-to-ground transmission bottlenecks.

Native-Process Orchestration and Deterministic Isolation: To circumvent the prohibitive virtualization tax and persistent memory footprint of modern containers (e.g., Docker), AIOS utilizes native-process orchestration (via OS-level POSIX process derivation) to instantiate AI workloads. This minimalist approach achieves equivalent deployment agility with near-zero overhead. Furthermore, AIOS enforces deterministic fault isolation: each AI process is mandated to emit a 1Hz heartbeat. Complementing the hardware-level *Systemgood* pin used for catastrophic kernel panics, AIOS manages application-level soft errors via a 1Hz software heartbeat. If a high-energy particle causes a COTS tensor operation to hang, AIOS bypasses the kernel's traditional OOM killer and issues a precise `SIGINT` soft interrupt to gracefully terminate and restart only the affected micro-process, preventing local errors from cascading into kernel panics.

Weak-Network Proxy and Memory Caching: LEO satellites face extremely brief and error-prone communication windows. AIOS implements a custom space-ground session protocol featuring a high-density 14-byte security header, nested integrity checks, and a sliding window. To facilitate application updates without exacerbating eMMC wear from fragmented writes, AIOS employs memory caching for chunked uploads. Payloads are written to the physical disk in a single contiguous I/O operation only after sequential verification in RAM.

Dynamic On-the-Fly Reconfiguration: Leveraging this weak-network proxy, the EAC achieves true dynamic runtime reconfiguration. Ground operators first transfer lightweight, interpretable scripts (e.g., Python) utilizing the aforementioned memory-cached upload mechanism. Subsequently, these scripts are executed and managed via an AIOS-exposed Pseudo-Terminal (PTY) SSHv2 session. By invoking native shell commands, entirely new AI pipelines can be rolled out instantaneously, effectively bypassing the need for monolithic firmware flashes and transforming a narrowband link into a highly agile deployment channel

3.4. Heterogeneous Edge AI Pipelines

The EAC validates the feasibility of deploying both Large Language Models and concurrent vision tasks on a severely constrained edge node, employing strict resource scheduling strategies.

Hardware-Decoupled Heterogeneous Concurrency: Deploying massive models concurrently naturally exceeds the strict 8GB RAM and NPU bandwidth limits of a 12W node. Rather than employing heavy software locks, AIOS gracefully handles resource starvation via its 1Hz heartbeat mechanism: stalled processes are swiftly reset via soft interrupts. To bypass shared bottlenecks entirely and achieve true concurrency, AIOS aggressively exploits the EAC's physically decoupled M.2 architecture. By routing workloads to dedicated hardware lanes, it seamlessly orchestrates a CPU-bound SLM (RWKV), an NPU-bound vision task (LPR), and a Hailo-bound YOLOv7 simultaneously without intra-board interference.

Constrained Edge LLM Deployment: The primary generative AI payload is a custom "Space-to-Earth Medical Consultation" LLM—to our knowledge, the first in-orbit domain-specific medical SLM—based on the linear-attention RWKV-1.6B architecture [20] running efficiently on the CPU. Due to the unreliability of full-duplex links, the system eschews streaming generation in favor of full-response downlinks. To strictly manage the 12W power budget, it utilizes a dynamically configurable Retrieval-

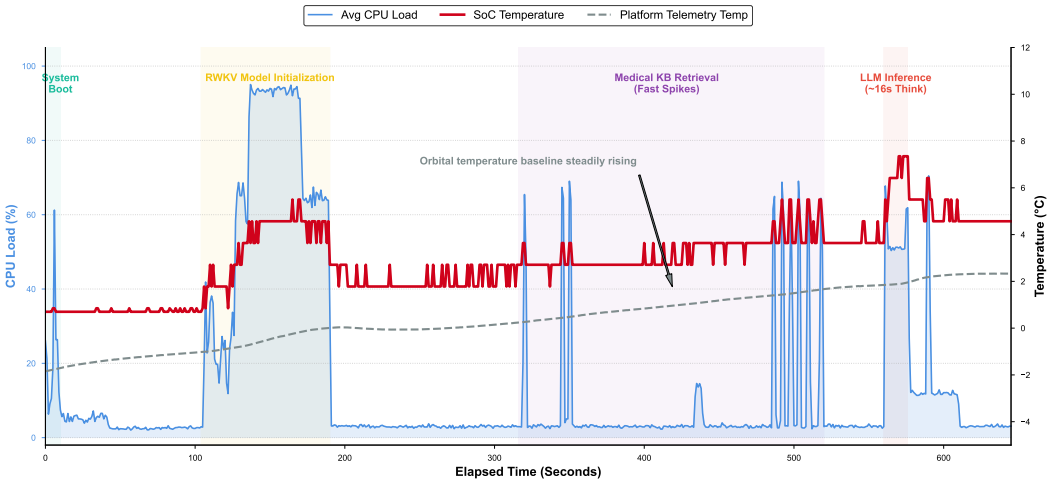


Figure 3. In-Orbit Thermal Profile and CPU Load Across LLM Execution Phases

Augmented Generation (RAG) architecture: queries are first intercepted by a lightweight SQLite vector database [19]. Configured in a power-saving 'cache-intercept' mode for this flight, it directly returns matches to prevent LLM awakening; otherwise, un-cached queries trigger the power-intensive LLM for standalone inference. Furthermore, the architecture pre-integrates a 3-State Flag system (Unverified, Doubtful, Reliable) to facilitate dynamic, closed-loop knowledge base evolution via ground-station overrides. Furthermore, to explore the boundaries of onboard cognitive limits within our strict 12W budget, we successfully deployed the quantized DeepSeek-R1-Distill-Qwen-1.5B on the NPU. Rather than asserting its immediate readiness for mission-critical spacecraft autonomy, this deployment serves as a pioneering feasibility study. It investigates how deeply distilled architectures can introduce offline Chain-of-Thought (CoT) exploration to severely power-constrained bare-LEO nodes, trading off absolute parameter scale for crucial SWaP-C viability.

Concurrent Vision and Structured Downlink: Vision workloads are mapped to dedicated accelerators. The LPR task operates on the NPU and employs a "structured sparse downlink" strategy: instead of transmitting hundreds of megabytes of raw imagery, the system crops the region of interest and downlinks a highly compressed JSON dictionary (bounding boxes, confidence scores, and plaintext license numbers), reducing bandwidth consumption to mere kilobytes. Concurrently, YOLOv7 object detection runs on the Hailo M.2 accelerator. While initial in-orbit functional validation utilized a baseline script yielding approximately 3 FPS, AIOS's agile reconfiguration channel is designed to seamlessly inject fully optimized, memory-resident asynchronous pipelines to unleash the accelerator's nominal peak throughput.

4. In-Orbit Evaluation & Flight Results

The EAC payload was launched into Low Earth Orbit (LEO) aboard an experimental satellite in 2025. As of the latest telemetry, the EAC has successfully completed 124 days of environmental validation in a bare-LEO environment. Operating as an independent payload, the system accumulated approximately 4.5 hours of high-load, on-demand execution with a 100% boot success rate. This section evaluates the system's survivability, reconfiguration agility under weak links, generative AI throughput, and the dynamic hot-updating of concurrent vision pipelines.

4.1. Flight Heritage and COTS Survivability

Despite the absence of active cooling and radiation-hardened shielding in the vacuum of space, the COTS-based EAC demonstrated exceptional thermal and operational stability under its 18.7W peak power draw. Telemetry data confirmed that the system experienced zero catastrophic system-level failures over the 124-day period. Consequently, the OC-driven hardware boot intervention and AIOS's deep SEU recovery mechanisms (e.g., SD NAND reverse repair) remained strictly dormant, acting as an unbreached safety net.

As illustrated in Figure 3, we profiled the system's thermal behavior during an extended high-load medical LLM session. Benefiting from the highly efficient carbon-fiber thermal pathways and a cold-biased initial state (starting near 0°C before exhibiting natural orbital thermal drift), the SoC temperature exhibited a remarkably stable plateau. Even when the CPU load remained saturated (averaging >60%) during the intensive 90-second LLM initialization phases, the absolute SoC temperature peaked at merely 6°C. This indicates that the passive dissipation mechanism effectively prevented localized thermal runaway in the vacuum.

Identical thermal safety margins were observed during intensive NPU-bound tasks.

4.2. Zero-Virtualization Reconfiguration over Weak Links

Relying on a newly-developed experimental full-duplex payload, our satellite-to-ground link suffers from high latency and unstable error rates. Consequently, across 38 in-orbit file upload missions via the AIOS proxy, the effective throughput averaged merely 26.53 KB/s. Extensive offline profiling revealed that this severe degradation was compounded by physical link noise, a mandatory 1024-byte frame padding mechanism, and the overhead of the custom Stop-and-Wait ARQ protocol.

Despite these harsh network realities, AIOS's zero-virtualization PTY mechanism proved its absolute superiority. While traditional cloud-native orchestration demands tens of megabytes for Docker images—rendering updates practically impossible over such a degrading link—the entire AIOS middleware stack operates with a static Resident Set Size (RES) of under 170MB. Leveraging this minimalist footprint, we successfully transmitted a mere 2KB Python update script and seamlessly executed it via the remote terminal. This enabled true on-the-fly hot-updates of both the object detection and LPR pipelines without a single kernel reboot.

To counter the sluggish downlink, we implemented a Structured Sparse Downlink strategy. Instead of transmitting raw imagery, the updated applications crop regions of interest and generate lightweight JSON dictionaries (bounding boxes and confidence scores). This reduced the typical LPR downlink payload from 37.6 KB to 1,348 Bytes (a 96.4% reduction), and the object detection payload from 42.5 KB to 452 Bytes (a 98.9% reduction), transforming multi-second transmission delays into sub-second blips.

4.3. Edge Generative AI Performance

The EAC successfully demonstrates the deployment and validation of an offline, domain-specific LLM on a severely constrained, independent bare-LEO node. We conducted extensive testing of the "Space-to-Earth Medical Consultation" application, processing 45 interactive queries.

The configurable RAG architecture played a pivotal role in compute peak-shaving. To maximize test coverage during short orbital passes and establish a minimal-power baseline, the LLM generation toggle was bypassed. Out of 45 queries, 34 successfully hit the local SQLite vector database (a 75.5% interception rate). For RAG cache intercepts, the end-to-end latency (including link overhead) plummeted to an average of 2.47s, yielding an impressive effective throughput of 50.73 tokens/s. Conversely, the 11 un-cached queries triggered the CPU-bound RWKV-1.6B autoregressive inference, resulting in an average latency of 19.14s and a throughput of 3.92 tokens/s. This 7.7x systemic speedup underscores the absolute necessity of lightweight RAG interception in power-constrained edge nodes.

Furthermore, we evaluated the platform's ecosystem compatibility with advanced reasoning paradigms by successfully deploying the DeepSeek-R1-Distill-Qwen-1.5B quantized model. While we acknowledge the ongoing industry debate regarding the reliability and hallucination rates of distilled SLMs in mission-critical space tasks, this 5-session in-orbit test was primarily designed to validate hardware-software interoperability. To our knowledge, it represents the first successful bare-LEO execution of a distilled Chain-of-Thought (CoT) architecture on a 12W node, yielding an average end-to-end latency of 61.25s and providing foundational telemetry for future, more rigorous onboard cognitive benchmarks.

Table 2. In-orbit performance evaluation of edge generative models under heterogeneous deployment. (DeepSeek-R1 throughput excludes discarded onboard CoT tokens.)

Model / Workload	Phase (Hardware)	Latency (s)	Tokens/s
Space-Medical SLM (RWKV-1.6B)	RAG Cache Intercept (CPU)	2.47	50.73
Space-Medical SLM (RWKV-1.6B)	Full Inference (CPU)	19.14	3.92
DeepSeek-R1-Distill (Qwen-1.5B)	Offline CoT (NPU)	61.25	4.11

4.4. Vision Concurrency and Throughput Analysis

During the concurrent vision tests, the payload processed 100 images for both object detection and LPR. The returned confidence scores aligned 100% with ground-truth bit-level simulations, confirming zero soft-error tensor corruption.

A performance disparity was observed: the NPU-based LPR application maintained a healthy 17.18 FPS (58.2 ms/img). Conversely, the 26-TOPS Hailo YOLOv7 application was evaluated using a strict "isolated-frame" validation protocol. By intentionally tearing down and re-initializing the hardware VDevice for each frame, we ensured that potential SEU-induced tensor corruptions would not accumulate in the accelerator's memory across continuous runs. While this diagnostic approach inherently bottlenecked the throughput to 2.79 FPS (compared to the module's 45 FPS nominal peak), it guaranteed absolute test determinism.

Crucially, for the upcoming operational phase, the fully optimized, memory-resident asynchronous pipeline can now be seamlessly injected on-the-fly via AIOS's agile reconfiguration channel.

5. Discussion and Future Work

While the 124-day deployment of the EAC successfully validated edge generative AI under extreme constraints, it represents merely the first stepping stone. Grounded in our in-orbit engineering realities, our future work spans near-term protocol optimizations, mid-term distributed scaling, and a

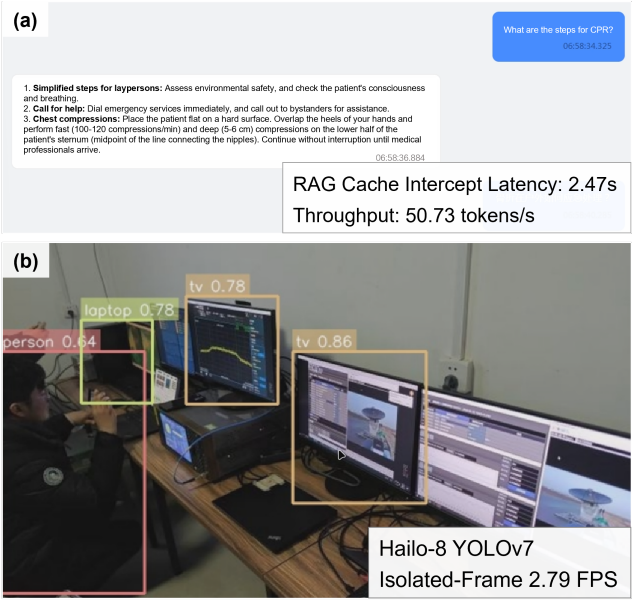


Figure 4. In-orbit functional validation of concurrent workloads. (a) Interface of the "Space-to-Earth Medical Consultation" LLM. (b) Concurrent YOLOv7 object detection executing on the decoupled Hailo-8 accelerator.

long-term paradigm shift toward spaceborne autonomous agents.

5.1. Limitations and Near-Term Enhancements

The current flight mission establishes a clear roadmap for immediate development. First, due to the prioritization of TT&C windows, the closed-loop updating of the knowledge base and the activation of the full RAG generation toggle were deferred. Activating this toggle via AIOS's agile reconfiguration remains a primary objective for the next operational phase. Second, the exercising of auxiliary I/O and OTA bus is pending the deployment of the Network and Management Module (NMM) in our upcoming constellation phase, which will serve as the gateway for multi-node clustering. Finally, leveraging the validated zero-virtualization PTY channel, our immediate software focus is injecting the fully optimized, memory-resident vision pipelines to completely unleash the 26-TOPS Hailo accelerator.

5.2. Scaling Out: Heterogeneous Distributed Topology

To transcend the compute ceiling of a single EAC node, our architectural roadmap introduces a highly resilient, fully distributed constellation topology. To avoid theoretical ambiguity, this vision is anchored in specific cyber-physical integrations. At the physical layer, multiple EAC nodes will interconnect via the pre-provisioned, triple-redundant communication backbone (dual Gigabit Ethernet, high-speed SerDes, and WLAN), with the aforementioned NMMs providing hardware-level survival monitoring. At the software layer, AIOS's decentralized service discovery mechanism will enable elastic compute scaling. Application microservices will autonomously discover newly joined nodes, balance inference workloads across the cluster, and achieve seamless fault self-healing without relying on a single point of failure.

5.3. Paradigm Shift: Towards Spaceborne Autonomous Agents

The ultimate vision extends far beyond executing isolated, command-driven vision or LLM pipelines. The future of aerospace engineering demands a transition toward Spaceborne Autonomous Agents [5, 8, 15, 17]. Operating under the same strict SWaP constraints, future onboard reasoning LLMs—empowered by the CoT capabilities successfully tested here—will act as the central "cognitive engine" to process fuzzy, high-level natural language instructions. They will perform dynamic task planning through *Tool Calling*—e.g., autonomously dispatching the decoupled Hailo accelerator for target acquisition, maneuvering the satellite's attitude control via the CAN bus, and synthesizing the intelligence via the RAG knowledge base—thereby realizing true, closed-loop spacecraft autonomy in deep space exploration.

6. Conclusion

We presented the Space Edge AI Computer (EAC), a hardware-decoupled payload engineered to bridge the critical gap between high-compute COTS accelerators and stringent nanosatellite SWaP limits. Operating at a typical power of merely 12W and weighing 703.5g, the EAC delivers up to 32 TOPS of heterogeneous compute density. By eschewing heavy cloud-native containerization in favor of AIOS's zero-virtualization native-process orchestration, and by physically embedding SEU recovery mechanisms via OC-driven hardware intervention, the EAC achieved a flawless initial 124-day independent bare-LEO validation phase.

Furthermore, this mission successfully demonstrated the viability of deploying cutting-edge generative AI on severely power-constrained edge nodes. The implementation of a configurable SLM+RAG architecture effectively slashed retrieval latency, while concurrent vision pipelines were successfully evaluated. By proving the system's extreme software elasticity—enabling dynamic, on-the-fly algorithm hot-updates over weak links without kernel reboots—this work establishes a highly pragmatic, flight-validated foundation for the impending era of distributed spaceborne agent networks.

References

[1] China launches first of 2,800 satellites for AI space computing constellation - SpaceNews. https://spacenews.com/china-launches-first-of-2800-satellites-for-ai-space-computing-constellation/, 2023. Accessed: 2026-03-03. 2

[2] Alibaba's Qwen-3 becomes one of world's first AI models to operate in orbit | South China Morning Post. https://www.scmp.com/tech/tech-trends/article/3341388/alibabas-qwen-3-becomes-worlds-first-ai-model-operate-orbit, 2024. Accessed: 2026-03-03. 2

[3] Sidus Space Announces Successful On-Orbit Operation of FeatherEdge™ Gen-2 Aboard LizzieSat®-3. https://investors.sidusspace.com/news-events/press-releases/detail/237/sidus-space-announces-successful-on-orbit-operation-of, 2025. Accessed: 2026-03-03. 2

[4] alecniedenthal. Space Llama: Meta's Open Source AI Model Is Heading Into Orbit. Meta Newsroom, 2025. Accessed: 2026-03-03. 2

[5] Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. Large Language Models as Tool Makers. *arXiv preprint arXiv:2305.17126*, 2024. 8

[6] Xinye Cao, Guoshun Nan, Hongcan Guo, Hanqing Mu, Long Wang, Yihan Lin, Qinchuan Zhou, Jiayi Li, Baohua Qin, Qimei Cui, Xiaofeng Tao, He Fang, Haitao Du, and Tony Q. S. Quek. Exploring LLM-Based Multi-Agent Situation Awareness for Zero-Trust Space-Air-Ground Integrated Network. *IEEE Journal on Selected Areas in Communications*, 43(6):2230–2247, 2025. 2

[7] James A. Casaletto, Patrick Foley, Mark Fernandez, Lauren M. Sanders, Ryan T. Scott, Shubha Ranjan, Shashi Jain, Nate Haynes, Marjan Boerma, Sylvain V. Costes, and Graham Mackintosh. Bridging Earth and Space: A Flexible and Resilient Federated Learning Framework Deployed on the International Space Station. *bioRxiv preprint doi:10.1101/2025.01.14.633017*, 2025. 2

[8] Xiaozheng Gao, Yichen Wang, Bosen Liu, Xiao Zhou, Ruichen Zhang, Jiacheng Wang, Dusit Niyato, Dong In Kim, Abbas Jamalipour, Chau Yuen, Jianping An, and Kai Yang. Agentic Satellite-Augmented Low-Altitude Economy and Terrestrial Networks: A Survey on Generative Approaches. *arXiv preprint arXiv:2507.14633*, 2025. 8

[9] Laura Parra García, Gianluca Furano, Max Ghiglione, Valentina Zancan, Ernesto Imbembo, Christos Ilioudis, Carmine Clemente, and Paolo Trucco. Advancements in Onboard Processing of Synthetic Aperture Radar (SAR) Data: Enhancing Efficiency and Real-Time Capabilities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:16625–16645, 2024. 1

[10] Gianluca Giuffrida, Luca Fanucci, Gabriele Meoni, Matej Batič, Léonie Buckley, Aubrey Dunne, Chris van Dijk, Marco Esposito, John Hefele, Nathan Vercruyssen, Gianluca Furano, Massimiliano Pastena, and Josef Aschbacher. The Φ-Sat-1 Mission: The First On-Board Deep Neural Network Demonstrator for Satellite Earth Observation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 2

[11] Ragini Gupta and Klara Nahrstedt. Performance Characterization of Containers in Edge Computing. *arXiv preprint arXiv:2505.02082*, 2025. 2

[12] Raluca Maria Hampau, Maurits Kaptein, Robin Van Emden, Thomas Rost, and Ivano Malavolta. An empirical study on the Performance and Energy Consumption of AI Containerization Strategies for Computer-Vision Tasks on the Edge. In *The International Conference on Evaluation and Assessment in Software Engineering 2022*, pages 50–59, Gothenburg Sweden, 2022. ACM. 2

[13] Sung Jun Kim, Minchae Kim, and Han Lim Choi. ChatGPT-based Mission Scheduling for Earth Observation Satellites. *Journal of Institute of Control, Robotics and Systems*, 31(5): 510–516, 2025. 2

[14] Toshinori Kuwahara, Yoshihiro Tomioka, Kazufumi Fukuda, Nobuo Sugimura, and Yuji Sakamoto. Radiation effect mitigation methods for electronic systems. In *2012 IEEE/SICE International Symposium on System Integration (SII)*, pages 307–312, 2012. 4

[15] Zhenping Li. Developing AI Agents for Satellite Operations. *AIAA Journal*, 21(3), 2025. 1, 8

[16] N. Melega, Bernardo A. Carnicero, N. Longepe, A. Paskeviciute, V. Marchese, Oriol Aragon, Irina Babkina, Alessandro Marin, Jakub Nalepa, Leonie Buckley, Giorgia Guerrisi, Sofia Oliviera, and Hano Stein. Implementation of the Φsat-2 on board image processing chain. In *Sensors, Systems, and Next-Generation Satellites XXVII*, pages 264–276. SPIE, 2023. 2

[17] Alejandro D. Mousist. ASTREA: Introducing Agentic Intelligence for Orbital Thermal Autonomy. *arXiv preprint arXiv:2509.13380*, 2025. 2, 8

[18] NSR. NSR Report: Space Traffic Data Volumes Increase 14x Over the Next Ten Years. https://www.globenewswire.com/news-release/2021/12/06/2346194/0/en/NSR-Report-Space-Traffic-Data-Volumes-Increase-14x-Over-the-Next-Ten-Years.html, 2021. Accessed: 2025-06-12. 1

[19] Taehwan Park, Geonho Lee, and Min-Soo Kim. MobileRAG: A Fast, Memory-Efficient, and Energy-Efficient Method for On-Device RAG. *arXiv preprint arXiv:2507.01079*, 2025. 6

[20] Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Xingjian Du, Teddy Ferdinan, Haowen Hou, Przemysław Kazienko, Kranthi Kiran GV, Jan Kocoń, Bartłomiej Koptyra, Satyapriya Krishna, Ronald McClelland Jr, Jiaju Lin, Niklas Muennighoff, Fares Obeid, Atsushi Saito, Guangyu Song, Haoqin Tu, Cahya Wirawan, Stanisław Woźniak, Ruichong Zhang, Bingchen Zhao, Qihang Zhao, Peng Zhou, Jian Zhu, and Rui-Jie Zhu. Eagle and Finch: RWKV with Matrix-Valued States and Dynamic Recurrence. *arXiv preprint arXiv:2404.05892*, 2024. 5

[21] David Rijlaarsdam, Tom Hendrix, Pablo T Toledano Gonzalez, Alberto Velasco-Mata, Leonie Buckley, Juan Puig,

Oriol Aragon Casaled, and Aubrey Dunne. The Next Era for Earth Observation Spacecraft: An Overview of CogniSAT-6. *Preprint*, 2024. 2

[22] Pia Singh. 'Greetings, earthlings': Nvidia-backed Starcloud trains first AI model in space as orbital data center race heats up. https://www.cnbc.com/2025/12/10/nvidia-backed-starcloud-trains-first-ai-model-in-space-orbital-data-centers.html, 2025. Accessed: 2026-03-03. 2

[23] Kathiravan Thangavel, Dario Spiller, Roberto Sabatini, Stefania Amici, Nicolas Longepe, Pablo Servidia, Pier Marzocca, Haytham Fayek, and Luigi Ansalone. Trusted Autonomous Operations of Distributed Satellite Systems Using Optical Sensors. *Sensors*, 23(6):3344, 2023. 1

[24] Kathiravan Thangavel, Roberto Sabatini, Alessandro Gardi, Kavindu Ranasinghe, Samuel Hilton, Pablo Servidia, and Dario Spiller. Artificial Intelligence for Trusted Autonomous Satellite Operations. *Progress in Aerospace Sciences*, 144: 100960, 2024. 1

[25] B. Todd and S. Uznanski. Radiation Risks and Mitigation in Electronic Systems. *arXiv preprint arXiv:1607.01573*, 2015. 4

[26] Chao Wang, Yiran Zhang, Qing Li, Ao Zhou, and Shangguang Wang. Satellite Computing: A Case Study of Cloud-Native Satellites. In *2023 IEEE International Conference on Edge Computing and Communications (EDGE)*, pages 262–270, Chicago, IL, USA, 2023. IEEE. 2

[27] Guoping Wang, Gang Wan, Zhijuan Su, Yang Wang, Yutong Jia, Gong Li, and Shi Liang. High-performance On-orbit Intelligent Computing and Real-time Services for Remote Sensing Satellites Based on Space Large-scale Computing Power. *Preprints.org*, 2024. 2

[28] Ruichen Zhang, Hongyang Du, Yinqiu Liu, Dusit Niyato, Jiawen Kang, Zehui Xiong, Abbas Jamalipour, and Dong In Kim. Generative AI Agents With Large Language Model for Satellite Networks via a Mixture of Experts Transmission. *IEEE Journal on Selected Areas in Communications*, 42(12):3581–3596, 2024. 2

[29] Maciej Ziaja, Piotr Bosowski, Michal Myller, Grzegorz Gajoch, Michal Gumiela, Jennifer Protich, Katherine Borda, Dhivya Jayaraman, Renata Dividino, and Jakub Nalepa. Benchmarking Deep Learning for On-Board Space Applications. *Remote Sensing*, 13(19):3981, 2021. 2