engineering proceedings

MDPI

Proceeding Paper

More TOPS, Less Mass: Compute-Density Metrics for Enabling Future AI Missions in Space †

Jeremiah Gayle [1,*] and Josh Kalin [2]

[1] Department of Systems Engineering, Colorado State University, Fort Collins, CO 80523, USA
[2] Integration Innovation Inc. (i3), Huntsville, AL 35808, USA; joshua.kalin@i3-corps.com
* Correspondence: jeremiah.gayle@colostate.edu
† Presented at The 1st International Online Conference on Aerospace (IOCAE 2026), 16–17 April 2026; Available online: https://sciforum.net/event/IOCAE2026.

Abstract

Onboard artificial intelligence (AI) is increasingly limited less by raw processor availability than by spacecraft-level integration constraints. Modern commercial edge-AI processors can provide orders of magnitude more inference capability than traditional spacecraft processors and, when normalized by spacecraft mass, CubeSat-class platforms can exhibit a very high compute density. This paper introduces compute density, expressed as operations per kilogram and paired with operations per watt, as a practical systems metric for comparing onboard AI capabilities across spacecraft classes. A representative 6U CubeSat case study is used to show that Jetson-class and space-adapted edge processors can enable computer vision, data triage, change detection, and autonomy workloads that were previously impractical on small spacecraft. However, compute density alone is incomplete: useful onboard AI capability is constrained by available power, thermal rejection, radiation tolerance, duty cycle, memory and data movement, and downlink strategy. This paper concludes that future AI spacecraft should be architected through compute power thermal communications co-optimization rather than by selecting the highest peak tera-operations-per-second (TOPS) processor.

Keywords: onboard AI; CubeSat; edge computing; compute density; TOPS/kg; spacecraft thermal design; radiation tolerance; verification and validation

1. Introduction

Spacecraft have traditionally been designed around a simple operational assumption: collect data on-orbit, transmit it to the ground, and perform the computationally intensive processing later at ground station. For decades, this approach was reasonable. Space-qualified computers were commonly built around radiation hardened processors with long flight heritage, prioritizing reliability, deterministic behavior, and environmental robustness over peak performance. These processors remain essential for many missions, especially where long duration, high consequence, or deep space radiation environments dominate the design. However, the resulting onboard processing capability is modest relative to modern commercial edge computing hardware, so many advanced processing tasks have historically been deferred to ground systems where power, cooling, storage, and compute resources are effectively unconstrained [1].

That architecture is beginning to show its limits. Modern space missions increasingly depend on timely decisions rather than simply collecting large volumes of raw data. Earth

Academic Editor: Konstantinos Kontis

Published: 20 July 2026

observation spacecraft may need to identify disasters, cloud-free scenes, illegal maritime activity, or environmental change while the satellite is still in a useful observation geometry. Intelligence, surveillance, and reconnaissance missions may need to detect and prioritize targets in near real time. Autonomous spacecraft, distributed constellations, and cislunar or deep space missions may operate during limited communication windows or with latency that make ground in-the- loop processing impractical. In these cases, waiting minutes, hours, or days for downlinked data to be processed can reduce mission effectiveness and can shift the bottleneck from sensing to communications and operations.

At the same time, commercial edge AI hardware has changed what is technically possible onboard a spacecraft. Small, power-efficient accelerators can now provide tera-scale inference performance within volume and power envelopes compatible with small satellites. Devices in the NVIDIA Jetson family, and other emerging space-adapted systems based on similar processors, provide a useful example of this shift because they offer substantial AI throughput in compact modules with software ecosystems familiar to terrestrial robotics and autonomous systems developers [2]. A CubeSat can no longer be assumed to be compute-poor simply because it is physically small.

The resulting design question is not simply whether a small spacecraft can carry an AI processor. It can. The more important question is how onboard AI capability should be measured and bounded across spacecraft classes. Traditional metrics, such as clock speed, MIPS, or even raw floating point operations per second, do not map cleanly to modern machine-learning workloads. For mission architects, the onboard computer is a spacecraft resource that competes with mass, power, thermal margin, radiation tolerance, storage, and downlink capacity. This paper therefore evaluates compute density, expressed as FLOPS/kg or TOPS/kg and paired with TOPS/W, as a systems-level metric for comparing onboard AI architectures.

The central argument is that onboard AI is no longer limited mainly by raw processor capability. Modern edge processors can provide orders-of-magnitude more AI throughput than traditional rad-hard command and data handling processors, and CubeSat-class systems may achieve very high compute density when normalized by mass. However, compute/kg alone is insufficient. Useful onboard AI must be evaluated as compute density constrained by power availability, thermal rejection, radiation survivability, duty cycle, and communication strategy. Future AI spacecraft should therefore be designed around compute power thermal communications co-optimization rather than around the highest advertised TOPS value. This paper makes four contributions. First, it distinguishes peak compute density from mission-effective compute density by showing that TOPS/kg and TOPS/W must be reduced by duty-cycle, thermal, radiation, data-flow, and verification constraints before they represent useful spacecraft capability. Second, it clarifies how the proposed metric differs from conventional SWaP-C, processor TOPS/W, mass-normalized TOPS/kg, and radiation-aware processor evaluations. Third, it provides a representative 6U CubeSat case study with explicit mass, power, duty-cycle, thermal, and processor-configuration assumptions. Fourth, it defines an application-level validation framework for onboard AI workloads so that processor selection can be tied to inference latency, energy per inference, model accuracy, memory demand, data reduction, and fault recovery rather than peak processor specifications alone [3].

2. Background: Traditional Space Compute and CubeSat Constraints

Traditional spacecraft computers have been sized primarily for command and data handling, spacecraft housekeeping, sequencing, telemetry management, fault detection, and execution of flight software rather than for high-rate neural-network inference. Radiation hardened processors, such as RAD750-class architectures, are valued because they

can operate reliably in harsh environments and provide predictable behavior. Their role is not obsolete; rather, their design priorities differ from those of modern edge AI accelerators. They are optimized for assurance, survivability, and control authority, not maximum operations per watt or machine-learning throughput.

The recent onboard-AI literature has moved beyond the question of whether machine-learning inference can execute on spacecraft hardware and toward the question of whether it can produce useful mission products under spacecraft constraints. Reviews of neural-network inference onboard satellites identify both hardware accelerators and software optimization techniques as enabling technologies, but also emphasize the importance of memory movement, power consumption, model compression, and validation on representative workloads [4]. On-orbit demonstrations have further shown that lightweight models can perform useful satellite-image inference and limited training onboard spacecraft, with reported latency values that make application-level benchmarking essential rather than optional. More recent onboard-AI studies have also shown that multi-task Earth-observation workloads, such as cloud segmentation, flood detection, and marine-debris classification, can be evaluated on low-power embedded systems, reinforcing the need to compare AI processors using workload-level metrics rather than peak TOPS alone [5].

CubeSat missions have introduced a different set of constraints. Standardized form factors, such as 1U, 3U, 6U, and 12U, reduce access to space barriers but impose tight limits on volume, surface area, battery capacity, and heat rejection [6]. A CubeSat may be able to host high performance electronics, but the spacecraft must still generate the electrical power, store energy through eclipse, conduct heat from dense electronics to radiating surfaces, maintain safe component temperatures, and survive launch and space environments. NASA small spacecraft guidance emphasizes the importance of power generation, batteries, avionics, and subsystem integration as coupled spacecraft resources rather than independent component choices. Figure 1 summarizes the relative size increase from 1U through 12U and motivates why the 6U class is a practical scale for studying onboard-AI payload integration.

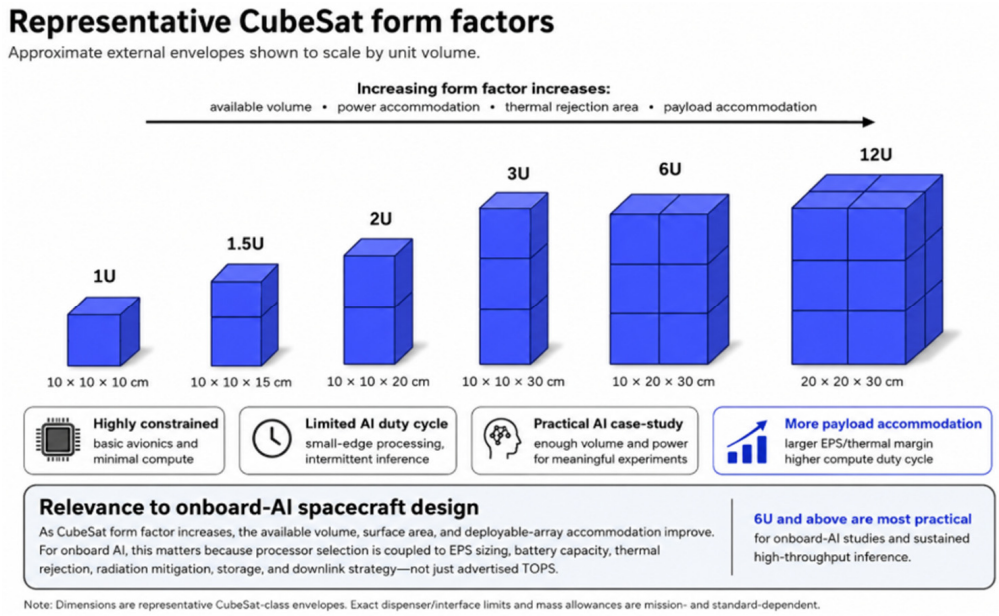

Figure 1. Representative CubeSat-class form-factor scale comparison for 1U, 3U, 6U, and 12U classes. The figure emphasizes the increase in volume, available surface area, and payload accommodation that makes 6U CubeSat attractive for onboard AI mission studies.

The approximate CubeSat power capability illustrates the scale of the constraint. A 3U spacecraft with body-mounted solar panels may provide only tens of watts at peak, while

deployable panels can increase that value substantially. A 6U spacecraft with deployable arrays can support higher peak power, potentially approaching the hundred-watt class for advanced configurations, but orbit average power, eclipse duration, battery depth of discharge, pointing constraints, degradation, and thermal limits reduce how much power is available continuously for the compute. A processor that can operate at 25–60 W may be acceptable as a payload element during short duty cycles but difficult to support as a continuously active subsystem.

Thermal constraints are equally important. Spacecraft dissipate heat by radiation, and small satellites have limited area for radiators and limited internal volume for heat pipes, thermal straps, heat spreaders, and isolators. High density electronics can produce localized hot spots even when total spacecraft power appears feasible. For a CubeSat, rejecting 100 W of concentrated electronics heat is a first-order architecture problem, not a minor packaging issue. The thermal design must therefore be coupled to compute scheduling, payload duty cycle, surface optical properties, spacecraft attitude, and orbit lighting conditions.

Radiation tolerance also differentiates terrestrial edge AI hardware from traditional spacecraft processors. Commercial GPUs and system-on-modules are not inherently radiation hardened. They may experience single-event upsets, latch-up, memory corruption, configuration errors, or total ionizing dose degradation depending on orbit, shielding, mission duration, and device construction. The limiting question is therefore not whether a GPU can run AI, but whether the spacecraft can supply power, reject heat, tolerate radiation, recover from faults, and manage data flow well enough for the GPU to be mission useful. In this paper, "space-adapted" refers to commercial or COTS-derived AI processing hardware that has been packaged, screened, tested, or operationally constrained for use in a space environment, but is not necessarily equivalent to a fully radiation-hardened processor [7]. "Processor availability" refers to the fraction of mission time during which the processor can be powered, thermally maintained, protected from unacceptable radiation or fault risk, and scheduled for useful workload execution. "Co-optimization" refers to simultaneous design of compute, power, thermal, communications, storage, and operations resources rather than sequential selection of the highest-performance processor followed by spacecraft accommodation. "Verification and validation" refers to both demonstrating that the hardware and software meet the specified requirements and showing that the resulting onboard AI products are operationally useful for the mission.

3. Candidate Edge-AI Hardware for CubeSat-Class Missions

The rapid growth of AI at the edge has transformed terrestrial robotics, autonomous vehicles, unmanned systems, industrial automation, and remote sensing. Workloads once dependent on cloud or centralized computing can now be executed directly on the platform, reducing latency and allowing autonomy when communications are degraded. A similar transition is beginning in space: onboard computer vision, object detection, change detection, autonomous tasking, and data triage can reduce downlink burden and increase responsiveness.

This paper uses the NVIDIA Jetson family as a representative commercial edge AI trade space because it spans low power and high-performance configurations and is supported by a mature software ecosystem, including CUDA, TensorRT, and common machine-learning frameworks. The Jetson Xavier NX is advertised at up to 21 TOPS in a compact module, with lower power operating modes around 10–15 W [8]. Jetson Orin NX modules provide higher advertised AI performance, including up to 70 TOPS for the 8 GB configuration and up to 100 TOPS for the 16 GB configuration under sparse INT8 assumptions [9]. Jetson AGX Orin-class systems provide still higher absolute performance

but at power levels more appropriate to larger spacecraft, hosted payloads, or duty-cycled operation [10]. Table 1 summarizes these metrics.

Table 1. Representative edge AI processors for CubeSat-class or hosted-payload mission studies. Values are representative manufacturer advertised or configuration-dependent figures, and should be verified for the selected module, carrier board, software stack, and thermal design.

Processor	Approximate AI Performance	Power Range	Approx. TOPS/W	Representative Use
Jetson Nano	~0.5 TFLOPS FP16	5–10 W	N/A	Low-power vision and autonomy baseline
Jetson Xavier NX	Up to 21 TOPS INT8	10–15 W	1.4–2.1	General embedded AI processing
Jetson Orin NX 8 GB	Up to 70 TOPS INT8	10–25 W	2.8–7.0	Advanced computer vision and autonomy
Jetson Orin NX 16 GB	Up to 100 TOPS INT8	10–25 W	4.0–10.0	High-performance onboard inference
Jetson AGX Orin/space-adapted systems	Up to 248–275 TOPS class	15–60 W or higher	~4–18	Hosted payloads, larger spacecraft, or radiation-tolerant COTS products

A fair comparison between Jetson-class commercial modules and space-adapted AI processors requires common assumptions. Manufacturer peak TOPS values may depend on INT8 precision, sparsity, batch size, software stack, memory bandwidth, and unconstrained thermal operation. A spacecraft-level comparison should therefore state the workload, precision, power mode, thermal boundary condition, radiation posture, and measurement method before using TOPS/kg or TOPS/W as decision metrics. Space-adapted systems should be compared as integrated processor assemblies, including packaging, power conditioning, thermal interface hardware, shielding or radiation-mitigation provisions, and any operational constraints required for the intended orbit [11].

While absolute compute capability is important, power efficiency is often more constraining for spacecraft. The CubeSat class of platforms often operate within power budgets measured in tens of watts, so TOPS/W can be as important as TOPS alone. However, even TOPS/W is not enough because a processor's useful in-flight performance depends on model precision, memory bandwidth, storage access, thermal throttling, duty cycle, and the ability of the spacecraft to keep the processor powered and within qualification limits.

The advantages of Jetson-class devices come with significant engineering risks as shown in Table 2. These processors were developed for terrestrial environments and do not inherently provide the same radiation assurance as traditional space processors. Space-adapted products, such as Aitech's S-A2300 radiation-tolerant COTS AI supercomputer and EDGX Sterna-class onboard processing products, illustrate how the market is beginning to package commercial AI compute for low Earth orbit (LEO) missions [12]. These systems do not remove the need for a mission-specific radiation analysis, thermal design, or verification; rather, they demonstrate a path toward using a high-performance COTS-derived computer in space with explicit environmental controls and operational constraints. This work quantifies these risks into an assessment of likelihood and mission dependence.

Table 2. Representative Jetson-class spaceflight risk assessment. Likelihood and impact are qualitative and mission dependent.

Risk	Likelihood	Impact	Potential Mitigations
Total ionizing dose accumulation	Medium	High	Shielding, mission-duration limits, radiation testing, derating
Single-event upset	High	Medium	ECC memory, checkpointing, watchdog recovery, software restart
Single-event latch-up	Medium	High	Current limiting, fast power switching, latch-up detection
Thermal overload	High	High	Thermal straps, radiators, heat pipes, duty cycling, thermal throttling
Power budget exceedance	Medium	High	Dynamic power modes, compute scheduling, energy-aware autonomy
Software corruption or model fault	Medium	Medium	Containerization, golden images, health monitoring, redundancy
Communication bottleneck	Medium	Medium	Onboard filtering, compression, prioritization, event reports
Compute underutilization	Low	Medium	Mission-specific benchmark cases and workload tuning

4. Compute-Density Metric: Compute per Mass as a Metric

Compute density is defined here as available onboard computational capability divided by spacecraft mass. For floating-point workloads, the metric may be expressed as TFLOPS/kg. For neural-network inference workloads, TOPS/kg is often more relevant because many modern accelerators advertise integer or reduced-precision tensor throughput. A related metric, TOPS/W, expresses compute efficiency relative to electrical power. Neither metric is sufficient by itself, but together they provide a useful first-order view of how much onboard AI capability a spacecraft architecture can carry and operate.

The proposed metric should not be interpreted as a replacement for SWaP-C, TOPS/W, TOPS/kg, or radiation-aware processor evaluation. Instead, it links these views into a mission-level screening framework. SWaP-C captures spacecraft resource burden but does not measure inference value. TOPS/W captures processor electrical efficiency but does not account for spacecraft mass, thermal rejection, radiation posture, or data-product usefulness. TOPS/kg captures mass-normalized peak performance but can overstate capability if the advertised processor throughput cannot be sustained in the selected orbit, thermal boundary, power mode, precision, or workload. A radiation-aware evaluation captures survivability and fault risk but may not quantify how much useful AI processing remains after operational constraints are applied. Mission-effective compute density therefore asks how much onboard inference can be executed, trusted, thermally sustained, and converted into useful mission products [13].

The peak compute density may be written as follows:

$$CD\_peak = Tops\_peak/M\_spacecraft \quad (1)$$

The power-normalized compute efficiency may be written as follows:

$$\eta\_compute = TOPS\_usable/P\_compute \quad (2)$$

The mission-effective compute density may be written as follows:

$$ECD = (TOPS\_usable/M\_spacecraft) \times f\_duty \times f\_thermal \times f\_radiation \times f\_data \times f\_V\&V \quad (3)$$

where f_duty, f_thermal, f_radiation, f_data, and f_V&V are mission-specific availability or usefulness factors between 0 and 1. These factors are not universal constants. They should be estimated from mission analysis, thermal analysis, radiation assessment, benchmark testing, and verification evidence [14]. A comparison of the metrics is highlighted in Table 3.

Table 3. Difference between mission effective compute density and related processor selection metrics.

Metric	What It Captures	Primary Limitation	Role in This Paper
SWaP-C	Size, mass, power, and cost burden.	Does not measure inference value or data-product usefulness.	Input constraint set.
TOPS/W	Mass-normalized peak compute.	Often based on advertised peak TOPS and may ignore thermal or duty-cycle limits.	First-order architecture comparison.
Radiation-aware evaluation	TID, SEE, SEL, SEU, fault modes, and mitigation requirements.	May not quantify mission-useful compute under operating constraints.	Survivability and fault-availability factor.
Mission-effective compute density	Useful compute after duty-cycle, thermal, radiation, data, and V&V constraints.	Requires mission assumptions and benchmark evidence.	Primary proposed metric.

Mass normalized compute is useful because launch mass remains a dominant spacecraft constraint and because small satellites can exploit distributed architectures. A large monolithic spacecraft may carry greater absolute compute capability, but a constellation of small spacecraft can distribute sensing and processing across many orbital vantage points. In that context, a 6U CubeSat with modest absolute AI throughput may outperform a much larger spacecraft in compute per kilogram, especially when the mission value comes from local event detection, onboard triage, or distributed responsiveness rather than centralized processing.

However, compute density must be interpreted carefully. FLOPS do not directly equal useful AI performance. TOPS depends on precision, sparsity assumptions, workload type, memory bandwidth, compiler optimization, model architecture, batch size, and sensor I/O. A processor advertising high sparse INT8 throughput may not achieve that throughput on a specific flight workload. Similarly, spacecraft mass alone is incomplete because a high TOPS processor that cannot be powered or cooled for the required duty cycle has little mission value. For this reason, compute density should be treated as a screening metric rather than a final design criterion.

A more mission-relevant formulation is effective compute density: usable operations per kilogram after accounting for power, thermal, radiation, duty cycle, and data-flow constraints. In simplified form, this can be understood as peak compute density multiplied by availability factors that represent duty cycle, thermal headroom, fault tolerance, and downlink usefulness. This framing shifts attention from peak processor performance to mission effective compute, that is, the amount of onboard inference the spacecraft can execute, trust, store, and communicate under real flight constraints.

https://doi.org/10.3390/engproc2026142011

5. Case Study: Advanced 6U CubeSat AI Compute Architecture

A representative 6U CubeSat provides a useful case study because it is large enough to host meaningful payloads and deployable arrays while still remaining strongly constrained by small spacecraft SWaP. The notional spacecraft considered here uses the standard 6U external envelope, approximately 10 cm by 20 cm by 30 cm, with a mass on the order of 12 kg. The power architecture is assumed to include deployable solar arrays capable of high peak generation relative to earlier CubeSats, but with orbit-average power still limited by attitude, eclipse, degradation, battery capacity, and operational duty cycle. Figure 2 illustrates the resulting spacecraft-level coupling between the AI processor, sensor payload, power system, thermal design, radiation mitigation, operations, and communications architecture and Table 4 is the specific Jetson class compute assessment.

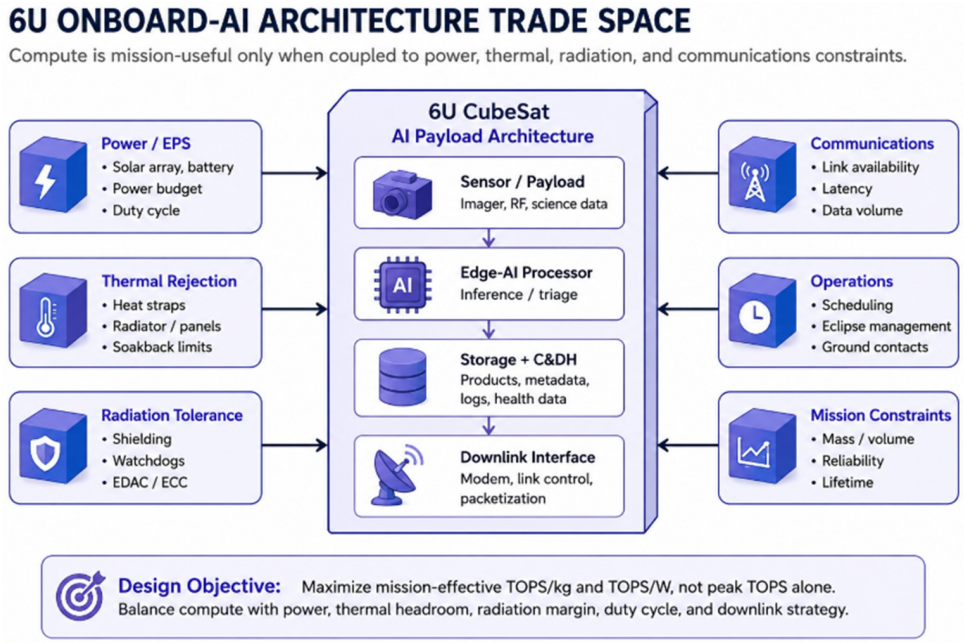

Figure 2. Notional 6U CubeSat-class spacecraft onboard-AI architecture showing how an edge-AI processor must be co-designed with sensor data flow, power generation and storage, thermal rejection, radiation mitigation, communications capacity, and operational duty cycle.

The representative case is intentionally not a claim of flight qualification for a specific processor. It is a systems-engineering example that makes the assumptions visible. The spacecraft is assumed to be a 6U CubeSat with a total mass of approximately 12 kg. A notional 3 kg payload allocation includes the sensor, AI processor, carrier electronics, heat spreader, thermal straps, local structure, and harnessing. The electrical power system is assumed to provide approximately 60–80 W orbit-average power under favorable deployable-array and attitude conditions, but only a fraction of that power is available for AI compute after spacecraft bus loads, payload loads, communications, heaters, battery charging, and margins are considered [15]. High-power inference is therefore treated as a duty-cycled activity rather than a continuous operating mode.

Two candidate compute configurations illustrate the design challenge. The first uses four Orin-class modules operating near 25 W each, producing an approximately 100 W compute load during peak inference periods. The second uses two higher-power modules or space-adapted edge-processing units operating near 40 W each, producing an approximately 80 W compute load but with potentially simpler packaging and reduced harnessing. These examples are intentionally notional. Their purpose is to show that the compute decision immediately becomes a spacecraft architecture decision. The processor selection

drives the EPS sizing, battery margin, radiator area, heat spreader mass, thermal-strap routing, structural packaging, harness design, current limiting, fault protection, and flight software scheduling.

Table 4. Representative Jetson-class spaceflight risk assessment. Likelihood and impact are qualitative and mission dependent.

Parameter	Assumption	Rationale/Impact
Spacecraft class	6U CubeSat, approximately 10 cm × 20 cm × 30 cm.	Large enough for meaningful payload and deployable arrays while remaining SWaP constrained.
Total mass	12 kg.	Representative mass for systems-level normalization.
Payload allocation	3 kg.	Includes sensor, AI compute, carrier electronics, thermal hardware, local support structure, and harnessing.
Orbit-average power	60–80 W.	Not all generated power is available to AI compute; EPS, ADCS, payload, communications, heaters, and margin compete.
AI compute power cases	25 W, 60 W, and 100 W burst cases.	Represents single-module, integrated space-adapted, and multi-module burst-processing cases.
Thermal design assumption	Dedicated heat spreader or thermal strap to radiator surface.	Sustained concentrated loads above approximately 50 W are likely thermal-design drivers.
Duty-cycle assumption	10–30% of orbit for high-power inference.	Represents imaging-pass, sunlit-burst, or event-driven operation rather than continuous peak compute.
Downlink strategy	Transmit detections, metadata, compressed chips, or prioritized scenes.	Data reduction is valuable only if onboard products are trusted and operationally useful.

A 6U AI CubeSat is unlikely to operate at full compute continuously unless specifically designed as a compute payload with substantial thermal accommodation. More realistic operations may involve burst processing during imaging passes, sunlit segments, or periods when downlink is unavailable. The processor can remain in a low-power state during eclipse or during payload inactive portions of the orbit, then execute inference when imagery or sensor data are available. This type of duty-cycled operation aligns AI processing with mission value while limiting sustained thermal load.

The case study therefore supports a key architectural point: AI compute should be treated as a payload-level subsystem rather than as an ordinary avionics upgrade. A high-performance AI processor requires dedicated power modes, thermal interfaces, radiation mitigation, fault containment, and operational rules. The spacecraft should be designed to answer not only how much compute can be installed, but how much can be used without violating energy balance, thermal limits, radiation-risk posture, or communications strategy.

6. Mission Workloads Enabled by Onboard AI

The highest near term value of onboard AI is not general-purpose intelligence in orbit, but selected inference tasks that reduce latency and downlink load. Computer vision is the leading use case because many spacecraft produce image-like data products and because convolutional and transformer-based vision models can be executed efficiently on edge

accelerators. Relevant tasks include object detection, cloud screening, scene classification, change detection, anomaly detection, event detection, and target prioritization.

Earth observation missions provide a clear example. A spacecraft imaging wide areas may collect far more data than it can downlink during available ground contacts. Onboard AI can filter out cloudy, low-value, duplicated, or empty scenes and prioritize images containing floods, fires, ships, vehicles, infrastructure damage, volcanic activity, or other mission-specific features. Instead of downlinking every raw image, the spacecraft can downlink detections, compressed chips, metadata, confidence scores, or a prioritized subset of scenes. The value is not only reduced data volume but improved timeliness.

Distributed constellations add another dimension. If each spacecraft can detect events locally, the constellation can perform autonomous tasking, cue neighboring satellites, or adjust observation priorities without waiting for centralized ground processing. This is particularly useful when ground contact is intermittent or when multiple spacecraft must coordinate sensing over a rapidly changing target. In such missions, compute density combines with communications architecture: onboard AI is valuable when it converts raw sensor data into smaller, more actionable information products.

Autonomy workloads also extend beyond imagery. Spacecraft may use onboard AI or machine learning-derived classifiers for health monitoring, fault diagnosis, communications scheduling, payload tasking, and navigation support. These applications must be implemented carefully because flight critical autonomy requires verification, fault containment, and explainable operational boundaries. Nevertheless, they show that onboard AI is best understood as part of a mission decision chain rather than as a standalone processor benchmark.

7. Application-Level Validation Framework

Processor metrics alone do not verify mission usefulness. A flight-relevant onboard-AI benchmark should use representative sensor data, the intended model architecture, the selected model precision, the flight-like software stack, the selected power mode, and the expected thermal boundary condition. The benchmarks, as seen in Table 5, should report inference latency, energy per inference, model accuracy, memory demand, processor utilization, data-reduction ratio, and recovery behavior after induced faults [16].

Prior onboard-AI work has reported measurable inference latency for satellite imagery and has shown that onboard processing can reduce the need to downlink raw data products. The relevant benchmark is therefore not only peak TOPS, but latency per image or tile, energy per inference, model accuracy, data-reduction ratio, memory-bandwidth demand, thermal response, and fault recovery. These quantities should be measured using representative sensor data, the intended model architecture, the flight-like software stack, the selected power mode, and the expected thermal boundary condition.

For Earth-observation triage, the required output may not be a full image-classification product. It may be a cloud mask, object-detection list, changed-region chip, compressed region of interest, confidence score, or priority flag. Accuracy and data reduction must therefore be evaluated together. A model that reduces downlink volume but discards mission-critical scenes is not useful. Conversely, a highly accurate model that cannot run within the available power, thermal, or timing window is also not mission useful. This is consistent with the recent onboard-AI literature showing that a practical mission value depends on data reduction, latency reduction, and workload-specific validation rather than processor peak performance alone [17]. Table 6 provides an example of how to validate AI workloads on a 6U spacecraft.

Table 5. Application-level validation data required for flight-relevant onboard AI benchmarks.

Validation Quantity	Required Measurement	Why It Matters
Inference latency	ms/image or s/tile at batch size one.	Determines whether processing can keep up with imaging cadence and operations timeline.
Energy per inference	J/image = average compute power × latency.	Connects model selection to battery state, power margin, and duty cycle.
Accuracy/mission utility	Precision, recall, F1, false-alarm rate, missed-event rate.	Determines whether onboard products can be trusted operationally.
Data reduction ratio	GB/s demand, model size, intermediate buffers, storage write rate.	Quantifies the communications benefit of onboard AI.
Memory bandwidth and footprint	25 W, 60 W, and 100 W burst cases.	Represents single-module, integrated space-adapted, and multi-module burst-processing cases.
Thermal response	Temperature rise per inference burst and recovery time.	Determines duty-cycle limits and radiator or thermal-strap requirements.
Fault recovery	Watchdog, restart, checkpoint, and degraded-mode behavior.	Determines whether COTS AI compute can be safely contained.

Table 6. Example validation matrix for representative 6U onboard AI workloads. Values should be treated as planning targets until replaced by measured benchmark data.

Workload	Representative Input	Target Output	Minimum Useful Result	Primary Limiting Resource
Cloud screening	Multispectral or RGB scene tiles.	Cloud/clear classification or mask.	High recall for clear scenes and meaningful downlink reduction.	Model accuracy and memory bandwidth.
Object detection	J Image tiles over maritime, disaster, or infrastructure regions.	Bounding boxes and confidence scores.	Latency compatible with imaging pass and acceptable false-alarm rate.	Inference latency and energy per image.
Change detection	Current scene plus stored or referenced prior scene.	Changed-region chips and metadata.	Transmit only changed or high-value regions.	Storage access and memory movement.
Anomaly/health monitoring	GB/s demand, model size, intermediate buffers, storage write rate.	Event flag or degraded-mode recommendation.	Low false-positive rate for operational use.	V&V evidence and fault containment.
Distributed tasking	Detection report from one spacecraft.	Cueing message to another spacecraft or to ground.	Decision latency below mission threshold.	Communications timing and autonomy rules.

8. Tradeoffs, Constraints, and Verification Considerations

The strongest engineering constraints for onboard AI are communications, power, thermal density, radiation, and scheduling complexity. Communications provide the mission motivation: onboard processing is most valuable when raw data volume exceeds downlink capacity or when latency matters. AI can convert high volume sensor data into detections, event messages, compressed products, or prioritized scene lists. However, this only helps if the onboard products are trusted, if false positives and false negatives are managed, and if the ground system can use the data products operationally.

Power is the first spacecraft-level constraint. The GPU or accelerator duty cycle must be coordinated with payload operations, battery state of charge, eclipse, downlink sessions, transmitter use, heater use, and thermal soak back. A processor that is acceptable in

isolation may become infeasible when operated simultaneously with an imaging payload, high-rate transmitter, and attitude control system. Energy-aware autonomy is therefore required: the spacecraft should schedule compute based on the expected mission benefit, available energy, and thermal state rather than running inference whenever data appear.

Thermal design is often the decisive constraint for CubeSat-class AI. Multiple accelerator modules can create concentrated heat loads that exceed what a small chassis can reject passively. Thermal straps, heat spreaders, radiator panels, phase-change elements, heat pipes, oscillating heat pipes, and deployable radiators may be needed depending on duty cycle and orbit. The relevant design problem is not simply average power, but where the heat is generated, how quickly it can be moved, what surfaces can reject it, and how component temperatures evolve over repeated inference cycles.

Radiation introduces a different risk category. COTS GPUs and AI system-on-modules may be attractive because of their high compute density, but they require system-level mitigation. Options include selective use in LEO, shielding, current limiting, watchdogs, power cycling, memory protection, software checkpointing, redundant processing lanes, graceful degradation, and board-level or system-level radiation testing. Space-adapted products, such as the S-A2300, illustrate that the industry is moving toward radiation-tolerant COTS AI systems, including radiation testing and packaging approaches intended for LEO use. Even so, each mission must assess the orbit, duration, shielding, single-event effects, and acceptable fault modes.

Verification must address AI compute as a mission function, not only as an electronics box. Traditional electrical functional tests are insufficient. A flight-like configuration should demonstrate that representative workloads can execute within the power, thermal, timing, storage, and communications constraints. Engineering development units or engineering models can be used for early workload characterization, thermal correlation, software stability testing, and fault recovery. Flight models should then undergo acceptance testing appropriate to the mission risk posture, including thermal vacuum testing, vibration, electromagnetic compatibility as needed, and processor workload tests that represent expected on-orbit duty cycles.

AI-specific verification also requires benchmark datasets and mission-relevant success criteria. The spacecraft should demonstrate not only that a model runs, but that it produces useful data products at the required cadence and under expected sensor, lighting, and compression conditions. Verification should include throughput, latency, memory use, power draw, thermal response, model accuracy on representative scenes, recovery from software faults, and data product compatibility with the ground segment. In this sense, onboard AI verification bridges avionics testing, payload testing, software validation, and mission operations rehearsal.

9. Discussion

Compute density is becoming a meaningful spacecraft architecture metric because it captures a shift in the relationship between spacecraft size and onboard processing. Small satellites are no longer limited to simple housekeeping and store-and-forward operations. With modern edge processors, a CubeSat-class spacecraft can host inference workloads that were previously associated with much larger platforms. This does not mean that CubeSats replace large spacecraft; rather, it means that some AI-enabled mission value can be distributed into smaller, more numerous platforms.

The principal limitation of compute density is that it can be misleading if treated as a standalone objective. A high TOPS/kg value may reflect an impressive processor, but mission value depends on whether the spacecraft can use that processor at the right time and for the right workload. A processor that throttles thermally, exceeds the power

budget, corrupts data under radiation, or produces outputs that cannot be downlinked effectively does not provide mission-effective autonomy. Therefore, compute density should be paired with power-normalized performance, duty-cycle assumptions, thermal margin, fault tolerance, and communications impact.

This framing changes how AI spacecraft should be designed. Instead of starting with the highest-performance processor and then attempting to package it, mission designers should begin with the sensing problem, latency requirement, downlink limitation, and operational concept. From there, the processor, model, power system, radiator, storage, and communications plan can be co-optimized. For CubeSat-class spacecraft, this approach is especially important because the margins are small and because one subsystem's peak operating mode can dominate the entire spacecraft design.

The distributed architecture implication is significant. A constellation of smaller AI-enabled spacecraft may outperform a single monolithic spacecraft for missions where event detection, revisit rate, geographic coverage, or downlink triage are more important than maximum absolute compute at one node. In such cases, the system metric is not only TOPS/kg per spacecraft, but mission-relevant autonomy per kilogram, per watt, and per downlinked bit across the constellation.

The distributed-autonomy implication also connects onboard AI to multi-agent coordination and consensus-control literature. When onboard detections are used to cue neighboring spacecraft or to support cooperative tasking, the communication delay, asynchronous updates, and cooperation-competition dynamics can influence constellation behavior. Prior work on multiagent consensus tracking and convergence of substochastic matrix products with communication delays is therefore relevant to future distributed-autonomy applications [18,19]. However, those works address networked decision stability and delayed coordination rather than the processor-level compute-density metric itself. They should be treated as enabling context for future constellation autonomy, not as the basis of the compute-density definition.

10. Limitations and Future Work

This paper remains a systems-engineering design study. It does not report new measured flight-test data for a specific processor, sensor, orbit, or spacecraft. The quantitative factors in the representative 6U case are intended to make assumptions explicit and to define the benchmark data required for mission selection. They should not be interpreted as demonstrated flight performance.

Published onboard-AI demonstrations and optimization studies show that latency, data reduction, and low-power embedded performance can be measured directly. Future work should therefore implement the proposed validation matrix on representative hardware using actual satellite imagery, selected neural-network models, and flight-like power and thermal constraints. The highest-value additions would include measured inference latency, energy per inference, thermal response, model accuracy, data-reduction ratio, memory-bandwidth demand, radiation fault behavior, and recovery time. Those data would allow mission-effective compute density to become a calibrated design metric rather than a screening framework.

11. Conclusions

This paper introduced compute density as a useful systems-level metric for evaluating onboard AI capability across spacecraft architectures. The analysis shows that modern edge AI processors can provide vastly greater inference capability than traditional spacecraft command and data handling processors and that CubeSat-class spacecraft can achieve very high compute density when normalized by mass. This shift enables small spacecraft to

perform onboard computer vision, data triage, event detection, and selected autonomy functions that can reduce latency and downlink burden.

However, this paper also shows that compute/kg alone is not a sufficient design metric. Useful onboard AI capability is constrained by power availability, thermal rejection, radiation tolerance, duty cycle, memory movement, communications architecture, and verification strategy. A 6U spacecraft may be able to carry modern AI compute, but it must be designed to operate that compute as a mission subsystem with dedicated power modes, thermal paths, radiation mitigation, and workload-specific test cases.

The main conclusion is that the future question is not whether CubeSats can carry AI processors. They can. The more important question is how much mission-relevant autonomy can be achieved per kilogram, per watt, and per downlinked bit. Future AI spacecraft should therefore be architected around compute power thermal communications co-optimization, not simply around selecting the highest TOPS processor.

Author Contributions: Conceptualization, J.G. and J.K.; methodology, J.G. and J.K.; formal analysis, J.G. and J.K.; investigation, J.G. and J.K.; resources, J.K.; data curation, J.G. and J.K.; writing—original draft preparation, J.G.; writing—review and editing, J.K.; visualization, J.G.; supervision, J.G.; project administration, J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new external dataset was generated for this paper. Mission-level information was compiled from publicly available sources and program documentation cited in the references.

Conflicts of Interest: Author Dr. Josh Kalin was employed by the company Integration Innovation Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
C&DH	Command and data handling
COTS	Commercial off-the-shelf
CV	Computer vision
ECC	Error-correcting code
EDAC	Error detection and correction
EDU	Engineering development unit
EM	Engineering model
EPS	Electrical power system
FLOPS	Floating-point operations per second
GB	Giga-Byte
GPU	Graphics processing unit
INT8	8-bit integer
LEO	Low Earth orbit
SEL	Single-event latch-up
SEU	Single-event upset
SWaP	Size, weight, and power
TID	Total ionizing dose
TOPS	Tera-operations per second
TVAC	Thermal vacuum
U	Unit
V&V	Verification and validation
W	Watt

References

1. Lentaris, G.; Maragos, K.; Stratakos, I.; Papadopoulos, L.; Papanikolaou, O.; Soudris, D.; Lourakis, M.; Zabulis, X.; Gonzalez-Arjona, D.; Furano, G. High-Performance Embedded Computing in Space: Evaluation of Platforms for Vision-Based Navigation. *J. Aerosp. Inf. Syst.* **2018**, *15*, 178–192. [CrossRef]
2. NVIDIA. Jetson Xavier NX Series. Available online: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-xavier-nx/ (accessed on 30 May 2026).
3. NASA Small Spacecraft Systems Virtual Institute. Small Spacecraft Technology State of the Art: Power Subsystems. NASA. Available online: https://www.nasa.gov/smallsat-institute/sst-soa/power/ (accessed on 30 May 2026).
4. Diana, L.; Dini, P. Review on Hardware Devices and Software Techniques Enabling Neural Network Inference Onboard Satellites. *Remote Sens.* **2024**, *16*, 3957. [CrossRef]
5. Inzerillo, G.; Valsesia, D.; Magli, E. Efficient onboard multi-task AI architecture based on self-supervised learning. *arXiv* **2024**, arXiv:2408.09754.
6. NASA CubeSat Launch Initiative. *CubeSat 101: Basic Concepts and Processes for First-Time CubeSat Developers*; NASA/SP-2017-6105 Rev. 1; NASA: Washington, DC, USA, 2017. Available online: https://www.nasa.gov/wp-content/uploads/2017/03/nasa_csli_cubesat_101_508.pdf (accessed on 30 May 2026).
7. EDGX. Sterna: Onboard Data Processing for Constellations. Available online: https://www.edgx.space/product/sterna (accessed on 30 May 2026).
8. Seed Studio NVIDIA Jetson Comparison. Available online: https://www.seeedstudio.com/blog/nvidia-jetson-comparison-nano-tx2-nx-xavier-nx-agx-orin/?srsltid=AfmBOorj49Od-7bU1vWpNxQ6VMOMooaTWDRwhf8KfnrKpaqOM94Y6hKt (accessed on 30 May 2026).
9. NVIDIA Developer Blog. Introducing Jetson Xavier NX, the World's Smallest AI Supercomputer. Available online: https://developer.nvidia.com/blog/jetson-xavier-nx-the-worlds-smallest-ai-supercomputer/ (accessed on 30 May 2026).
10. NVIDIA. Space Computing: On-Orbit AI and Accelerated Computing. Available online: https://www.nvidia.com/en-us/edge-computing/space-computing/ (accessed on 30 May 2026).
11. Aitech Systems. Radiation Test Report: S-A2300 AI Supercomputer for LEO. Available online: https://aitechsystems.com/s-a2300-radiation-tolerant-ai-supercomputer-leo/ (accessed on 30 May 2026).
12. Aitech Systems. S-A2300 Radiation-Tolerant COTS AI Supercomputer. Available online: https://aitechsystems.com/products/space/integrated-systems-space/s-a2300-radiation-tolerant-cots-ai-supercomputer/ (accessed on 30 May 2026).
13. Del Prete, R.; Thind, P.K.; Mazzeo, A.; Whitley, M.; Papa, L.; Longépé, N.; Meoni, G. Optimizing deep learning models for on-orbit deployment through neural architecture search. *Sci. Rep.* **2025**, *15*, 37783. [CrossRef] [PubMed]
14. Růžička, V.; Mateo-García, G.; Bridges, C.; Brunskill, C.; Purcell, C.; Longépé, N.; Markham, A. Fast model inference and training on-board of satellites. *arXiv* **2023**, arXiv:2307.08700.
15. NASA Small Spacecraft Systems Virtual Institute. Small Spacecraft Technology State of the Art: Avionics. NASA. Available online: https://www.nasa.gov/smallsat-institute/sst-soa/small-spacecraft-avionics/ (accessed on 30 May 2026).
16. Pham, H.V.; Tran, T.G.; Le, C.D.; Le, A.D.; Vo, H.B. Benchmarking Jetson Edge Devices with an End-to-End Video-Based Anomaly Detection System. *arXiv* **2023**, arXiv:2307.16834. Available online: https://arxiv.org/abs/2307.16834 (accessed on 30 May 2026).
17. Minott, D.; Siddiqui, S.; Haddad, R.J. Benchmarking Edge AI Platforms: Performance Analysis of NVIDIA Jetson and Raspberry Pi 5 with Coral TPU. In *Proceedings of IEEE SoutheastCon 2025*; IEEE: New York, NY, USA, 2025. Available online: https://scholars.georgiasouthern.edu/en/publications/benchmarking-edge-ai-platforms-performance-analysis-of-nvidia-jet/ (accessed on 30 May 2026).
18. Li, W.; Yan, S.; Shi, L.; Yue, J.; Shi, M.; Lin, B.; Qin, K. Multiagent consensus tracking control over asynchronous cooperation-competition networks. *IEEE Trans. Cybern.* **2025**, *55*, 4347–4360. [CrossRef] [PubMed]
19. Shi, L.; Yan, S.; Li, W. Consensus and products of substochastic matrices: Convergence rate with communication delays. *IEEE Trans. Syst. Man Cybern. Syst.* **2025**, *55*, 4752–4764. [CrossRef]