Orbital AI Datacenter: An Architectural Analysis

Barry Lyu [1,*], August Ning [2], Ronald Dreslinski [1], Nathaniel Bleier [1]
[1]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA
[2]EPFL, Lausanne, Switzerland
[*]Corresponding author: barrylyu@umich.edu

***Abstract*—Compute demand for frontier AI training is growing at $\sim 5\times$/year, straining compute hardware, terrestrial power grids and cooling infrastructure. Motivated by abundant solar energy and radiative heat rejection without cooling towers or chillers, several recent proposals have suggested moving AI training into orbit, ranging from disaggregated satellite platforms to distributed constellations and monolithic orbital mega-structures. However, the feasibility and economics of orbital AI training remain largely unexplored. This paper presents an architectural and techno-economic analysis of the benefits and limitations of orbital AI datacenters. We characterize the key constraints that distinguish in-space AI training from its terrestrial counterpart, including thermal management, radiation and deployment logistics. Under plausible launch and maintenance rates, orbital datacenters are unlikely to become a primary substitute for terrestrial capacity. We develop a bottom-up, compute-centric cost model to quantify the overhead of sustaining compute in orbit. Orbital compute remains more expensive than terrestrial capacity across the modeled scenarios, even under optimistic hardware and launch assumptions. Our performance analysis further shows that communication bottlenecks and radiation-induced errors lead to severely underutilized compute, compounding the cost disadvantage. We further show that scaling up individual satellites to improve co-location and amortize structural overhead is necessary to mitigate both cost and performance penalties, but that debris avoidance, construction complexity and single-point-of-failure risk render this approach untenable beyond megawatt scales. Under moderate assumptions, even the cost-optimal satellite scale retains a substantial cost disadvantage relative to terrestrial infrastructure. We conclude that edge inference on space-generated data better matches space computing than frontier training because the interconnect, synchronization and reliability demands that defeat training are absent.**

I. Introduction

The compute used in frontier AI training is growing $\sim 5\times$ per year [1], far outpacing energy-efficiency and chip-density gains. Neural scaling laws [2] suggest this growth will continue because larger models yield predictably better performance. The economic incentive to train them remains strong even as terrestrial data centers face mounting constraints (grid capacity, cooling, water, permitting) that threaten to bottleneck the growth.

Against this backdrop, a wave of proposals has emerged to move AI compute into space (Figure 1). The proposed SpaceX–xAI merger explicitly targets orbital AI compute, with Elon Musk arguing "space-based AI is obviously the only way to scale" [3]. Leaders from OpenAI, Amazon and Google echo the vision [4]. Skeptics counter that the logistics are prohibitive. AWS CEO Matt Garman noted "there are not enough rockets to launch a million satellites yet" and "it is just not economical" [5].

Computing in space has a growing record at smaller scales, including on-board inference demonstrations [14], [15] and architectural studies of single-satellite microdatacenters [16]–[18], alongside learning-based methods for operating constellations autonomously [19]. On the software side, latency-tolerant training (DiLoCo [20]) and distributed frameworks (OrbitalBrain [21]) address wide-area communication constraints. The proposals in Figure 1 envision datacenter-scale training across hundreds to millions of spacecraft at gigawatt power, a qualitative leap. At this scale, system-level interactions among interconnect, deployment logistics and economics add constraints beyond single-node concerns, placing the problem in the domain of datacenter architectural analysis.

This paper presents an architectural analysis of orbital AI datacenters across the spectrum of proposed designs (§II-A). We contribute two tools. A bottom-up, compute-centric total cost of ownership (TCO) model captures the major economic and physical constraints of sustaining AI compute in orbit. ASTRO-SIM, an orbital network simulator, couples Keplerian orbital propagation (via the `hapsira` library) with a modified analytical network backend of ASTRA-sim [22], [23]. It supports heterogeneous topologies and STAGE-generated traces [24] to model inter-satellite networks and quantify attainable training throughput. We use these tools to map the orbital design space and derive concrete design guidelines (§VII-B), yielding the following observations.

*a) **The orbital environment imposes architecture-agnostic physical constraints**:* (§III). Heat rejection dominates satellite mass budgets. Multi-GW systems require kilometer-scale radiator surface; thermal hardware often matches or exceeds compute mass. Demonstrated orbital capabilities cannot service hardware at the cadence terrestrial clusters sustain, causing permanent capacity loss at rates measured in real datacenters. Radiation and total ionizing dose (TID) effects compound this baseline attrition. Launch throughput caps sustainable orbital capacity at $\sim$34.7 GW even at one Starship launch per day, a ceiling terrestrial demand surpasses by $\sim$2028.

*b) **The economics do not close in our modeled design space**:* (§IV). A bottom-up TCO model reveals a U-shaped cost curve. Small satellites devote mass to per-node fixed costs (free-space optical (FSO) terminals, attitude control), while mega-structures require $\sim$91 debris avoidance maneuvers per day and propellant budgets requiring hundreds of annual

Proposal	Organization	Scale	Interconnect	Target workload	Status (2026)	Ref.
Suncatcher	Google / Planet Labs	81 sats, ∼300 m	FSO‡	ML training	Prototype 2027	[6], [7]
SpaceX/xAI	SpaceX	1M sats	ISL 100 Gbps	ML training	S-1 / FCC	[3], [8]
Three-Body	Zhejiang Lab / ADA Space	2,800 sats	Laser 100 Gbps	On-board inference†	12 sats on-orbit	[9]
StarCloud-4	StarCloud Inc.	4 km × 4 km, 5 GW	Wired (NVLink)	ML training	H100 on-orbit	[4], [10]
ASCEND	Thales Alenia / EU	1 GW (2050 target)	Wired	General compute	Feasibility study	[11]
Axiom ODC	Axiom Space	2 nodes	N/A	Edge / nat'l security	On-orbit (ISS)	[12]

†Each satellite carries an 8B-parameter model for on-board data processing, consistent with the edge inference niche we identify in §VI.
‡Google demonstrated 1.6 Tbit/s DWDM optical in ground tests; we analyze both 100 Gbps and 1.6 Tbps.

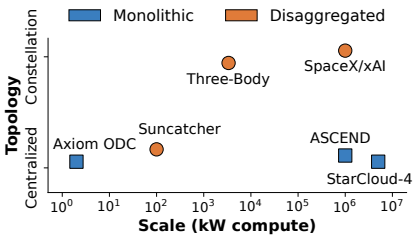


Fig. 1. Space compute proposals and their placement in the orbital datacenter taxonomy [13]. (Left) Proposal summary, grouped by architectural similarity. (Right) Taxonomy mapping by scale (horizontal), topology class (vertical) and physical construction (marker color, monolithic vs. disaggregated). The vertical axis shows categories (centralized vs. constellation); marker offsets only improve legibility. StarCloud-4 (a monolithic mega-structure) and the SpaceX/xAI 1M-sat constellation anchor the two topology classes.

launches. A cost minimum emerges at sub-megawatt-scale satellites under moderate assumptions, but orbital compute remains more expensive than terrestrial capacity before utilization and reliability losses. Even optimistic hardware and launch assumptions do not yield a modeled cost advantage.

*c) **Distributed training hits irreducible performance limits**:* (§V). Limited bandwidth and multihop latency impose substantial overhead on cross-satellite Mixture of Experts (MoE) communication, favoring satellite-local expert groups or dense models. graphics processing unit (GPU) utilization depends almost entirely on per-satellite NVLink domain size. Without a full NVL72 rack, utilization may fall to ∼5%. This creates a tension unique to orbit. Utilization favors ever-larger per-satellite scale-up domains, while the cost and physical limits of §IV cap satellite size. The two pressures meet at a single narrow design point (§VII-B). Independently, hardware failures at rates measured in terrestrial clusters cause permanent capacity loss without physical servicing; low Earth orbit (LEO) radiation compounds this already severe limitation, cutting checkpoint-based training efficiency several-fold at Llama-3 scale and by up to ∼75× at frontier scale (§V-E).

*d) **Edge inference matches the orbital environment**:* (§VI). On-board inference on space-generated data avoids the inter-satellite communication, synchronization and error accumulation across shared weights that make space hostile to training. This workload calls for a different satellite. It maximizes local inference throughput, keeps only the limited inter-satellite links needed to ingest data and receives models trained cheaply on the ground.

*e) **The combined penalties guide the design**:* (§VII-B). The cost, utilization and radiation penalties above are multiplicative and largely independent (Fig. 14), yielding ∼4.8–37× less effective compute per dollar at the best evaluated design point under moderate costs. The analysis identifies a cost-optimal satellite scale and informs parallelism mapping through both the link regime and the physical formation. Our formation analysis favors free-fall (Hill–Clohessy–Wiltshire) strings, rings and cylinders over rigid meshes or tori, coupling network design to station-keeping costs. The analysis also identifies inference as the orbit-matched workload.

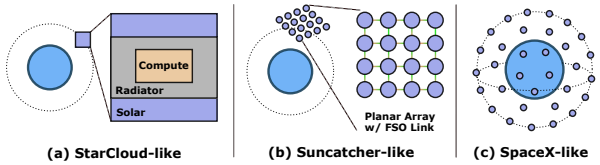


Fig. 2. Three representative orbital AI architectures spanning the design space of Figure 1. (a) Centralized monolithic mega-structure, (b) centralized disaggregated formation, (c) distributed constellation.

II. Background

Proposals for space-based AI infrastructure aim to alleviate several of the most pressing constraints facing terrestrial AI datacenters, including energy supply, heat rejection, site availability and carbon emissions [6], [10], [11].

First, solar panels in space receive ∼5.9× more incident power per unit area than even well-sited terrestrial installations (1361 W/m^2 under the AM0 solar constant versus roughly 230 W/m^2 annual average on the ground [25], [26]). This stronger solar resource makes orbit attractive for power-hungry frontier AI training.

Second, space offers an attractive ultimate heat sink. Radiators reject waste heat to deep space without cooling towers or chillers. Thermal management still requires transporting heat to those radiators, whose area and cooling-loop overhead can be substantial [10], [18].

Finally, space may relieve site scarcity and environmental pressure on Earth. Terrestrial AI infrastructure needs land, grid capacity, cooling and permitting, making large deployments slow [27], [28]. The carbon footprint of fossil-fuel grids adds regulatory and public scrutiny [29], [30]. Solar-powered orbital systems sidestep grid carbon and replace site constraints with launch and satellite cost constraints.

Launch costs have fallen by roughly 20× over two decades, with projections of further reductions [6]. Meanwhile, terrestrial growth is increasingly hard to sustain. U.S. datacenters consumed 183 TWh in 2024 (∼ 4% of national electricity), with projected consumption more than doubling by 2030 [31]. Together these trends make the case more compelling than in prior eras.

A. Orbital AI Taxonomy

Existing orbital AI proposals vary in scale, topology and physical construction (Figure 1). Three proposals target ma-

chine learning (ML) training as their primary workload and anchor our analysis (Figure 2).

Monolithic Mega-Structure (StarCloud-4 [10]). A single structure with centralized solar power and cooling, GPUs co-located and *wired* (NVLink/InfiniBand), effectively forming a terrestrial datacenter in space. StarCloud launched an H100 to orbit in November 2025, training NanoGPT on board [4]. **Disaggregated Satellite Formation (Suncatcher [6]).** A single logical compute node spread across a planar array of satellites in close formation ($\sim$300 m spacing), each with dedicated solar panels and TPU-class compute connected via FSO at up to 1.6 Tbps. Google plans prototype launches in 2027 with Planet Labs [7]. **Satellite Constellation (SpaceX/xAI [3], [8]).** A constellation spread across global LEO via inter-satellite links (ISLs). The proposed SpaceX–xAI merger filed for up to 1 million compute satellites with ISLs up to 100 Gbps, while SpaceX's S-1 describes orbital AI satellites targeted above $100\,\mathrm{kW/t}$, leveraging Starlink's launch and operations infrastructure.

B. ML Training and Datacenter Interconnect

Large-scale ML training distributes work across GPUs using several complementary parallelism strategies, including tensor parallelism (TP), pipeline parallelism (PP), data parallelism (DP) and expert parallelism (EP) [32]–[35]. Training systems map each strategy to the workload and the available interconnect hierarchy [32], [36].

Interconnect hierarchy. Modern GPU clusters exploit a two-tier interconnect hierarchy that matches the communication structure of distributed training. *Scale-up* traffic stays inside a node or rack-scale NVLink domain. NVLink/NVSwitch connects 8 GPUs in DGX A100/H100 nodes, or 72 GPUs in the GB200 NVL72 rack-scale system [37], [38]. NVLink 5 provides $900\,\mathrm{GB/s}$ per direction per GPU at $\sim 600\,\mathrm{ns}$ latency. *Scale-out* traffic crosses the inter-node fabric. InfiniBand (IB) NDR400 provides $400\,\mathrm{Gbit/s}$ per direction per port at $\sim 1.3\,\mu\mathrm{s}$ [39] (Table I). The tiers differ by $18\times$ in per-GPU bandwidth, so efficient training keeps fine-grained tensor-parallel collectives inside the NVLink domain and spreads data-parallel synchronization across many network rails [36], [38], [40]. This rail parallelism does not reduce the aggregate gradient volume that leaves a node. With a G-GPU tensor-parallel domain, data parallelism induces G shard-wise all-reduces of size $1/G$ each; the node still emits a full-gradient-equivalent volume, but roughly one network rail per GPU lets the per-link load fall with G.

Density and the space mismatch. The datacenter industry is moving toward extreme density, from 8 kW per rack in 2015 to 120 kW in 2024 (NVIDIA GB200 NVL72), with 500 kW projected by 2027 [38], [41], [42]. Keeping more GPUs inside the high-bandwidth NVLink domain drives this $15\times$ densification. Orbital proposals move in the *opposite* direction by distributing compute across satellites hundreds of meters to hundreds of kilometers apart, over links one to two orders of magnitude below intra-rack NVLink, undoing precisely the co-design that density buys.

TABLE I
INTERCONNECT TECHNOLOGY COMPARISON (TWO-WAY AGGREGATE BANDWIDTH).

Interconnect	BW (Gbps)	Latency	Switch Time	Type
NVLink 5	14 400	0.6 µs	$\sim$1 ns	Intra-node
IB NDR400	800	1.3 µs	100 ns	Inter-node
800GbE	1600	3 µs	$\sim$1 µs	Inter-node
FSO (Starlink)	100	667 µs†	10–30 s	Inter-sat

†At 200 km (Starlink); 167 µs at 50 km (SpaceX 1M); 1 μs at 300 m (Suncatcher).

C. Free-Space Optical Technology

FSO inter-satellite links use pointed laser terminals to establish high-bandwidth point-to-point connections [43], [44]. As of 2023, Starlink operated over 8,000 laser terminals across 4,827 satellites [45]. Throughout the performance analysis, we treat quoted FSO capacities as two-way aggregate bandwidth unless otherwise stated. A $100\,\mathrm{Gbps}$ Starlink-class link therefore provides $6.25\,\mathrm{GB/s}$ usable bandwidth per direction, while Suncatcher's $1.6\,\mathrm{Tbps}$ bidirectional bench result provides $100\,\mathrm{GB/s}$ per direction.

Per-terminal bandwidth depends strongly on distance. Beyond the Fresnel limit, received optical power falls as $1/d^2$. Close-range designs can use commodity $1.55\,\mu\mathrm{m}$ telecom optics and dense wavelength division multiplexing (DWDM); Suncatcher demonstrated $1.6\,\mathrm{Tbps}$ bidirectional transmission in bench testing [6]. At shorter formation spacings, spatial multiplexing may further increase aggregate throughput [6]. At long range, Starlink v2 claims up to $100\,\mathrm{Gbit/s}$ [46], [47], while the flight-proven SDA-standard class provides $2.5\,\mathrm{Gbit/s}$ [48].

FSO links use *circuit switching*. Changing a connection requires laser repointing, acquisition and tracking over 10–30 seconds [44]. Each terminal is one point-to-point link, so a satellite's terminal count k bounds its network degree. Without dedicated switching nodes, achievable topologies are low-degree graphs, including rings, trees and mesh/torus-like overlays only where geometry permits enough neighbor links. Clos or fat-tree networks have no analogue unless satellites serve solely as switches.

Geometry further restricts which links can exist. Terminals cannot operate near the Sun. Same-orbit links clear intervening spacecraft only when the chord's depth below the orbital arc ($L^2/8R$) exceeds spacecraft dimensions, which occurs at about $40\,\mathrm{km}$ at LEO altitudes. Thus close clusters can support terabit-class DWDM links but require coordinated station-keeping, while dispersed constellations relax formation flying but fall back to lower-rate long-range terminals.

III. FUNDAMENTAL CHALLENGES OF ORBITAL AI

The orbital environment imposes fundamental challenges beyond those of terrestrial datacenters. Space computing has handled many of them for mission computers and small-scale on-board processing, but the scale of orbital AI introduces new dimensions and interactions. This section covers the cross-cutting challenges that affect *all* orbital AI architectures, with implications for both cost (§IV) and performance (§V).

A. Heat Rejection

In space, radiation provides the *only* heat-rejection mechanism [49]. Existing satellites manage modest thermal loads, but a multi-GW datacenter demands unprecedented dissipation. Its radiators and coolant loops may dominate the satellite mass budget.

Stefan–Boltzmann emission at $20\,^{\circ}\mathrm{C}$ yields only $\sim 634\,\mathrm{W/m^2}$ of net rejection after accounting for solar absorption and Earth albedo [10], so a $5\,\mathrm{GW}$ thermal load demands $\sim 7.9\,\mathrm{km^2}$ of radiator surface.

Pumps transport heat from the compute hardware to these radiators at an additional power cost. The radiator panels and coolant loops (pumps, plumbing, working fluid proportional to radiator mass [18]) account for a significant fraction of satellite mass. With sufficient engineering effort and accepted overhead, though, the thermal bottleneck is manageable. Terrestrial liquid-loop cooling provides a starting point; radiator area scales predictably with waste heat, making this a costly but surmountable challenge.

B. Reliability and Serviceability

Hardware failures are a constant operational reality in large-scale GPU clusters. An 11-month study of Meta's clusters (16k and 8k A100 GPUs) attributes 100% of confirmed "lemon"-node causes to physical hardware requiring hands-on repair [50]. Terrestrial clusters absorb this through continuous physical intervention, including health checks, drain-and-repair, lemon replacement and adaptive routing, while sustaining effective training-time ratios of 0.85–0.9 [50].

Serviceability. These mitigations rely on physical access; in orbit, technicians, spare-parts inventory and hot-swap are unavailable at datacenter cadence, so unrepaired failures reduce usable capacity until replacement or servicing. At a representative lemon-node rate of $\sim$1.2%/month [50] (corroborated within a 0.5–2$\times$ band by device-level GPU replacement rates from Titan [51] and NCSA Delta [52]), a substantial fraction of nodes develop irrecoverable faults well within a 5-year mission, with no path to restoration other than launch-and-replacement (§III-C). The absence of servicing makes baseline commercial-silicon failures permanent. §V-E quantifies the training effects of both a *transient* class (radiation upsets, for which checkpointing offers a mitigation) and a *permanent* class (this attrition, which requires servicing).

Radiation effects. The LEO radiation environment introduces additional failure modes that have no terrestrial analogue. TID degrades devices over months to years; single-event effects (SEE) from individual high-energy particles cause instantaneous upsets. Both threaten all orbital compute hardware regardless of architectural class. No space-qualified GPU at $<7\,\mathrm{nm}$ exists today [53], [54].

TID. High Bandwidth Memory (HBM) is particularly sensitive to TID. In ground-based proton-beam tests, Suncatcher reports HBM stress-test irregularities at $2\,\mathrm{krad(Si)}$, about $2.7\times$ its estimated five-year in-orbit dose of $750\,\mathrm{rad(Si)}$ [6]. Sinclair and Dyer [53] model approximately $2.6\,\mathrm{krad(Si)}$ behind $10\,\mathrm{mm}$ Al over five years at $800\,\mathrm{km}$ under conservative radiation assumptions. Although the orbital assumptions differ, their analysis also finds limited shielding benefit beyond about $5\,\mathrm{mm}$ Al. The margin to HBM irregularities therefore requires mission-specific dose assessment.

Single-event effects. Individual high-energy particles flip memory bits. On-die error-correcting code (ECC), typically single-error correction, double-error detection (SECDED), corrects most, but the rest split into *detected-uncorrectable* and *silent* errors. Detected-uncorrectable errors cause the job to fail-stop and roll back; silent errors alias past ECC without a signal [55], [56]. In training, an uncorrectable gradient-buffer error can propagate via all-reduce to every worker and accumulate into weights; silent errors lack the signal to trigger rollback. Edge inference is more tolerant because a corrupted activation affects only one query, with no accumulation into shared state. §V-E quantifies the training impact using measured modern-memory upset rates, treating this transient class separately from the permanent attrition above.

C. Deployment

Every kilowatt of orbital compute requires launch to altitude. Replacing failed or obsolete subsystems uses the same constrained launch pipeline. Deployment is a continuous logistical operation whose throughput bounds sustainable capacity.

Using StarCloud's mass-favorable parameters [10] ($13\,\mathrm{kg/kW}$), we split satellite mass into infrastructure ($M_{\text{infra}} \approx 10.31\,\mathrm{kg/kW}$, 20-year life) and compute ($M_{\text{compute}} \approx 2.69\,\mathrm{kg/kW}$, 5-year refresh). Under *modular servicing*, which swaps individual subsystems in situ, the maintenance rate (m_{maint}) is $\frac{10.31}{20} + \frac{2.69}{5} \approx 1.05\,\mathrm{kg/kW/yr}$. Under *full retirement*, the operator discards the entire satellite at GPU end-of-life, increasing the rate to $13/5 \approx 2.6\,\mathrm{kg/kW/yr}$ (2.5$\times$ higher). Modular servicing demands capabilities that do not yet exist at scale (autonomous rendezvous, robotic swap, standardized interfaces), whereas full retirement is more straightforward but less mass-efficient.

These rates bound maximum sustainable capacity at $C_{\text{max}} = R \times M_{\text{launch}} \;/\; m_{\text{maint}}$ for R launches/year. US datacenter demand is $\sim$21 GW in 2024, growing at $\sim$15% CAGR [31]. We assume Starship becomes fully operational with daily or even multiple daily flights and dedicates *all* launches to orbital compute. Terrestrial demand surpasses the modular-servicing ceiling of $\sim$34.7 GW (one launch/day) by $\sim$2028; under full retirement the ceiling drops to $\sim$14.0 GW (Figure 3). This suggests that without order-of-magnitude improvements in launch capacity, orbital compute will remain a marginal supplement to terrestrial datacenters.

D. Orbital Mechanics

Orbital mechanics constrain site selection for an orbital datacenter. Altitude and inclination jointly determine power availability, launch mass, radiation exposure, drag and the formations the system can physically maintain. The first-order consideration is eclipse. In most low-inclination LEO orbits, the spacecraft enters Earth's shadow on every revolution (up to 37% of each $\sim$96-minute orbit), interrupting power unless

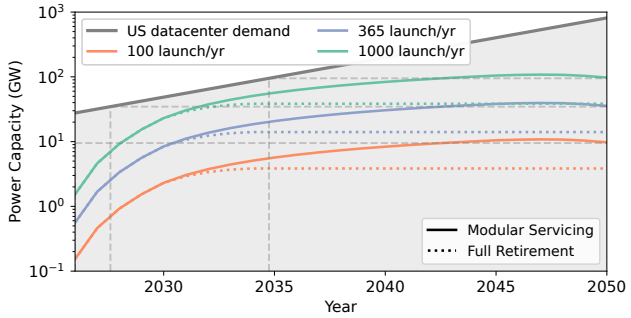


Fig. 3. Orbital compute capacity buildup vs. US data center demand. Modular servicing and full retirement at three launch cadences. Terrestrial demand (gray band, 15% CAGR) surpasses the aspirational 1-launch/day ceiling by ∼2028.

TABLE II
ORBIT TRADE-OFFS.

Orbit	Alt. (km)	Incl.	Max eclipse	Rel. Payload	5yr TID (vs. ISS)
Starcloud-1 (demo)	350	45.4°	40%	95%	∼0.5×
ISS (reference)	400	51.6°	39%	94%	1.0×
SpaceX/xAI	550	53.0°	37%	91%	3.7×
SSO-274	274	96.6°	25%	88%	1.7×
SSO-567	567	97.7°	∼5%	84%	∼5.3×
Suncatcher (dd-SSO)	650	97.8°	∼5%	82%	7.1×
SSO-894	894	99.0°	0	79%	19×
Molniya	39,800 / 530	63.4°	∼8%	45%	55×

the system carries enough energy storage and oversized solar generation to bridge the dark portion of the orbit. A dawn–dusk sun-synchronous orbit (dd-SSO) avoids most eclipse by riding the day–night terminator, but pays for it with a near-polar, slightly retrograde inclination that reduces delivered payload. Hence current proposals converge on sun-synchronous or near-continuous-sunlight orbits [6], [8]. Highly eccentric orbits, including Molniya, can also reduce eclipse through long dwellings at a far apogee, but they incur higher payload penalty and repeatedly traverse the radiation belts, trading the power problem for a radiation environment unsuitable for commodity AI hardware.

Table II shows the resulting constraint space. It reports worst-season eclipse fractions for circular orbits and payload relative to a low-inclination launch baseline. Molniya altitudes denote apogee/perigee. Low-inclination LEO preserves launch payload but pays for it with eclipse. Dawn–dusk SSO suppresses eclipse, but its payload penalty and radiation exposure grow as the orbit moves higher; moving lower reduces dose but increases drag and station-keeping burden, especially for array-dominated datacenters with large cross-sections and low ballistic coefficients. For the remainder of the analysis we adopt a dd-SSO in this narrow operating region, applying its launch penalty in the cost model (§IV) and its radiation environment in §III-B.

E. Formation Geometry

Orbit selection fixes the environment; the cluster's physical arrangement then constrains the interconnect topology, which a terrestrial datacenter architect normally controls through cabling. The formation determines which satellites remain close enough for short FSO links. Here the datacenter is a single compact formation (the disaggregated architecture of §II-A); *sphere* and *torus* therefore describe local cluster shapes, distinct from the planet-scale orbital shells of dispersed constellations. A formation preserves its geometry for free only if every satellite shares one semi-major axis and hence one orbital period, so the cluster returns to itself each revolution instead of drifting apart. We evaluate candidate formations at 550 km with ASTRO-SIM, propagating each under Earth's oblateness and atmospheric drag to measure its relative drift and the station-keeping ΔV it requires. A same-period pair stays bounded indefinitely, whereas a 100 m altitude difference, enough to change the orbital period, drives the two 31 km apart within two days. Small eccentricity and inclination differences provide radial and cross-track extent, so a feasible cluster follows a relative orbit that rotates and breathes once per revolution. These are the free Hill–Clohessy–Wiltshire (HCW) modes [57].

TABLE III
FORMATION GEOMETRY AND STATION-KEEPING.

Free-fall (unforced)		Forced (infeasible)	
Geometry	**ΔV/yr**	**Geometry**	**ΔV/yr**
Along-track string	28 m/s	Rigid 2D mesh	23 km/s
Ring	28 m/s	Rigid 3D mesh	41 km/s
Cartwheel disk	28 m/s	Sphere	63 km/s
Cylinder (helical)	28 m/s	Torus	143 km/s

We evaluate unforced along-track strings, projected-circular rings, cartwheel disks and along-track cylinders. The rigid tori and spheres in our comparison require continuous forcing. These unforced geometries cost only differential-drag makeup (∼28 m/s/yr, set by array area), while a rigid 3D mesh, sphere, or torus demands 41–143 km/s/yr, one to two orders of magnitude beyond a satellite's entire lifetime ΔV budget (∼0.1–1 km/s, Table III). Earth's oblateness adds only a slowly accumulating drift (∼188 m/day without correction) that routine formation-keeping removes [58]; differential atmospheric drag is the irreducible residual.

Dense, high-bisection orbital interconnects require additional terminal hardware, while maintaining the evaluated rigid arrangements incurs prohibitive station-keeping costs. These costs favor strings, rings, disks and cylinders; Suncatcher's formation design similarly yields planar ($N \sim R^2$) scaling [6]. For the neighbor-linked mesh/torus fabrics modeled in §V-C, bisection bandwidth scales as $O(\sqrt{N})$.

IV. ECONOMICS OF ORBITAL AI

A central claim of orbital compute proposals is that space will be *cheaper* than terrestrial data centers [6], [10]. The argument draws on real advantages. Solar irradiance in orbit is ∼5.9× higher per unit area than optimal terrestrial installations. Orbital facilities require no electric grid; waste heat radiates to the 2.7 K cosmic background without cooling towers or chillers. Meanwhile, terrestrial facilities face grid-

capacity, cooling-water, permitting and carbon constraints that add cost and schedule risk.

These advantages, however, are necessary but not sufficient. In orbit, every kilowatt of compute requires solar panels, power conditioning, radiators and structural support. Launching and eventually replacing this hardware introduces expenses with no terrestrial analogue. Both on the ground and in space, energy constrains capacity. Terrestrial facility capacity, cooling budgets and grid allocations use kilowatts of sustained IT load [41], [59], [60]; orbital collection area, radiator surface and structural mass likewise scale with electrical power. We therefore compare orbital and terrestrial systems on amortized $/kW-yr, the cost of sustaining one kilowatt of compute for one year. We evaluate whether the economics close even assuming solutions to the engineering challenges in Section III.

A. Power–Mass Breakeven

We divide total modeled satellite mass, including compute and all supporting subsystems, by total electrical power to obtain the mass-to-power ratio (kg/kW). It sets a first-order cost floor for a given launch price, equipment price and mission life. Using total electrical power favors orbit in this comparison; we normalize the detailed TCO below to GPU power.

Figure 4 plots this floor. We model annualized space cost as (kg/kW) $\times$ ($/kg launch + $/kg equipment) $/$ mission life with a 5-year mission. For comparison, we use a hyperscale facility at $10k/kW capex [41], [59], [60] and assume low-cost electricity at $0.02/kWh, yielding $675/kW-yr. Even under the most favorable space assumptions, with $200/kg launch (below any achieved Starship price) and $100/kg equipment (below the cost of consumer electronics, let alone space-qualified hardware), cost parity with this baseline requires $\leq$11.3 kg/kW. The Starlink v2 Mini, the most mass-optimized satellite in production, has an estimated 67–114 kg/kW [46], placing it well above this threshold.

This back-of-envelope estimate deliberately favors orbit by ignoring hardware refresh, operational overhead and the architectural penalties of Sections III–V. At the midpoint of the Starlink mass-to-power range, even the lowest-cost curve is 8.0$\times$ above the terrestrial baseline. The following subsection builds a more granular TCO model to examine where this gap could narrow.

B. Compute-Centric TCO Model

To move beyond the first-order breakeven, we build a bottom-up TCO model for sustained AI compute in orbit. Traditional parametric spacecraft cost models (SSCM [18], [62]) estimate procurement costs for science and communications satellites without modeling $/kW-year for datacenter workloads. In particular, they have no analogue for GPU refresh cycles, modular servicing versus full satellite retirement, collision-avoidance propellant at km^2 cross sections, or the power/thermal mass cascade that accelerator thermal design power (TDP) creates. We therefore choose a GPU count and TDP as our starting design point. We propagate

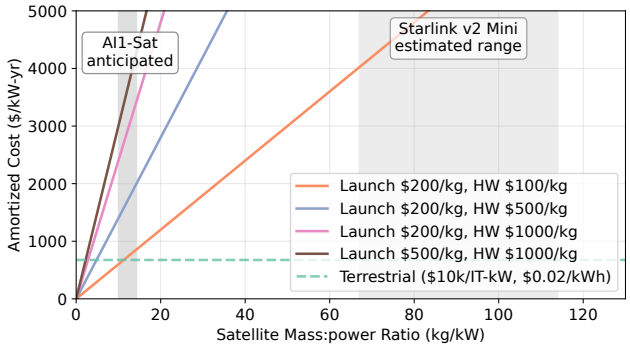


Fig. 4. Amortized cost ($/kW-yr) vs. satellite mass-to-power ratio. Solid lines show space cost under several (launch, equipment) price assumptions with a 5-year mission life. The dashed line shows the hyperscale baseline (facility capex amortized over 20 years plus electricity) [41], [59], [60]. Shading marks the Starlink v2 Mini estimate and the anticipated AI1 range [46], [61].

outward through power, thermal, structure, propulsion, launch, replacement and operations over a 20-year analysis horizon. Where possible, we anchor parameter ranges to public product numbers spanning projected frontier systems, demonstrated near-term hardware and mature heritage subsystems.

The model sizes every subsystem from the GPU power budget outward. A system-level power overhead (the space analogue of power usage effectiveness (PUE)) sets the total electrical load. The model then sizes solar arrays to generate that power and radiators to reject waste heat via Stefan–Boltzmann radiation, with additional thermal-loop mass (pumps, plumbing and working fluid) proportional to panel mass. Structural mass covers the satellite bus and scales with component mass; we scale propulsion and attitude determination and control system (ADCS) hardware from International Space Station (ISS) reference data. Construction cost is total mass times effective $/kg (with an SSO payload penalty) plus hardware costs, adding on-orbit assembly for structures beyond single-launch capacity. Table IV lists parameter ranges across three technology tiers.

Two ongoing cost drivers dominate operations. For *component replacement*, subsystem lifetimes vary widely (Table IV), so the operator chooses between modular servicing and full satellite retirement. We evaluate both at every design point; satellite size, servicing price and lifetime spread determine the cost-optimal choice. For *collision avoidance*, debris encounter rate scales with cross-section (calibrated from ISS), while each maneuver's propellant cost scales with the mass it moves.

The model deliberately excludes non-recurring engineering (NRE) costs (space-qualified GPU packaging, km-scale radiator development, autonomous servicing R&D) and does not account for insurance, regulatory, or spectrum costs. Many input parameters are uncertain or undisclosed; we address this through sensitivity sweeps across technology tiers (§IV-D). Table IV therefore combines reported product values, values derived from public geometry or hardware data and speculative bounds on undisclosed future costs and lifetimes.

TABLE IV
COST-MODEL PARAMETERS.

Parameters		Optimistic	Moderate	Pessimistic
Launch [10], [63]–[65]	Cost ($/kg)	500	1,500	2,700
	Payload (t)		100	
Compute [61]	GPU lifetime (yr)	5	5	3
	GPU cost ($k)	25	40	60
	Per GPU Mass (kg)	4.5	9.5	9.5
Power [10], [61], [66], [67]	Solar W/kg	325	150	80
	Solar Cost ($/W)	2	10	20
	Lifetime (yr)	20	10	5
	PUE overhead	1.15×	1.25×	1.35×
Thermal [49], [61], [66], [68]	Radiator kg/m^2	2.0	10.0	15.0
	Rad temp (°C)	60	20	20
	Loop/panel mass	1.1×	1.5×	1.8×
	Panel ($/m^2)	2,000	5,000	10,000
	Loop ($/kg)		1,000	
Structure [10], [66]	Mass overhead	15%	25%	30%
Propulsion [66]	I_{sp} (s)	1,900	1,500	1,500
	Thrust (N)	0.05	0.3	0.3
	Mass (kg)	3.4	15	15
	Cost ($k)	20	200	200
	Propellant ($/kg)	10	10	3,000
FSO [44], [47], [48]	Mass (kg)	8	20	35
	Cost ($k)	30	150	857
Servicing [69]	Vehicle mass (kg)	300	500	2,300
	Vehicle hw ($/kg)	3,000	5,000	20,000
	Payload (kg)	5,000	5,000	2,000

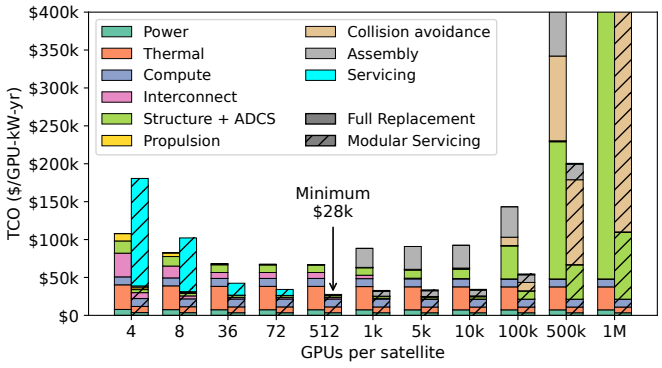


Fig. 5. 20-year TCO ($/GPU-kW-yr) vs. GPUs per satellite for a 1M-GPU fleet under moderate assumptions, showing modular servicing (hatched) and full retirement (solid) strategies.

C. Cost vs. Satellite Scale

Figure 5 sweeps satellite size from 4 to 1M GPUs, evaluating both modular servicing and full retirement at each design point. A U-shaped cost curve emerges, rendering both extremes of the design space untenable.

At the small end, per-node fixed costs, including FSO terminals, reaction wheels, propulsion and servicing, amortize over very few GPUs. For example, each Suncatcher satellite carries only a small TPU-class accelerator payload [6] yet needs a full complement of power, thermal and attitude hardware, so compute is only ∼8% of satellite mass. Such satellites also require full retirement (§III-C); servicing's per-visit cost amortized over so few GPUs exceeds the cost of relaunching.

At the large end, superlinear penalties dominate. A $4\,\mathrm{km}$ $\times$ $4\,\mathrm{km}$ structure like StarCloud [10] presents $\sim 16\,\mathrm{km}^2$ of debris cross-section (3711× the ISS), yielding ∼91 avoidance maneuvers per day and rendering deployment unrealistic. Each maneuver accelerates the full $67\,000\,\mathrm{t}$ mass. Even with argon Hall thrusters ($I_{sp} \approx 1500\,\mathrm{s}$), the annual propellant budget reaches $\sim 75\,855\,\mathrm{t}$, equivalent to 759 Starship launches per year for propellant alone, while the $7.9\,\mathrm{km}^2$ radiator surface adds $39.5\,\mathrm{kt}$ of mass to accelerate. On-orbit assembly at this scale is also far beyond demonstrated capability. The ISS, a fraction of this extent, took 13 years and 40+ crewed missions. Moreover, a single structure concentrates all risk. One debris impact or array failure affects the entire system, whereas losing one satellite in an 81-node constellation costs 1.2% of capacity.

Between these extremes, the moderate cost minimum occurs at 512 GPUs per satellite ($0.61\,\mathrm{MW}$ of GPU power). At this ISS-class scale, the satellite can host a full NVL72 domain and amortize fixed infrastructure while keeping debris avoidance routine (∼7 maneuvers/year) and single-satellite loss non-catastrophic. Existing missions have demonstrated on-orbit assembly at this scale. Modular servicing allows operators to replace shorter-lived compute and solar arrays without retiring longer-lived infrastructure. This infrastructure-lifetime decoupling makes StarCloud attractive without the aggregation penalties that make its scale untenable.

D. Sensitivity Analysis

The TCO model depends on parameters with uneven maturity. Public sources report launch pricing and payload limits, flight hardware masses and satellite geometry. We calibrate array W/kg and estimate radiator temperature using those disclosures. The least certain inputs are future costs, lifetimes, servicing capability and the compute packaging mass of custom orbital accelerators. We therefore define three technology tiers and sweep the GPU-per-satellite design space under each (Table IV).

The tiers bracket mass-to-power regimes. We anchor the optimistic tier to an AI1-class disclosed frontier at $12.3\,\mathrm{kg/kW}$ ($81\,\mathrm{kW/t}$), the most favorable density case in the sweep. An NVL72 domain, however, includes rack-scale liquid cooling and support hardware for its 72 accelerator packages [38]; an orbital implementation must replace commodity rack packaging with space-native compute, power, thermal and structural hardware qualified for launch vibration, vacuum, thermal cycling and radiation. In the model, matching AI1-class density therefore requires solar specific power, radiator areal density and compute-payload mass beyond publicly flown hardware [61], [66]. The moderate tier ($72.6\,\mathrm{kg/kW}$) uses current commercial/deployable assumptions and is consistent with the Starlink-v2-class mass-to-power range used in the first-order estimate. The pessimistic tier ($117.1\,\mathrm{kg/kW}$) uses heavier heritage-like hardware and near-term servicing costs. Launch pricing in all tiers is forward-looking. Even the pessimistic $2,700/kg assumes dedicated Falcon 9 vehicles, while actual 2026 rideshare pricing is ∼$7,000/kg [65].

Figure 6 separates scenario-dependent quantities from trends that are stable across the sweep. The exact mini-

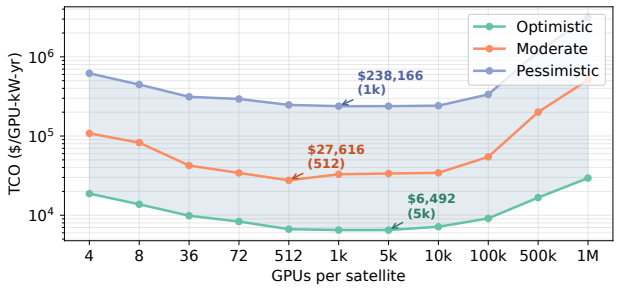


Fig. 6. TCO sensitivity envelope across three technology tiers. The moderate baseline minimizes near 512 GPUs/satellite. The optimistic tier forms a broad 1k–5k plateau because AI1-class density keeps the 5k point below the single-launch threshold.

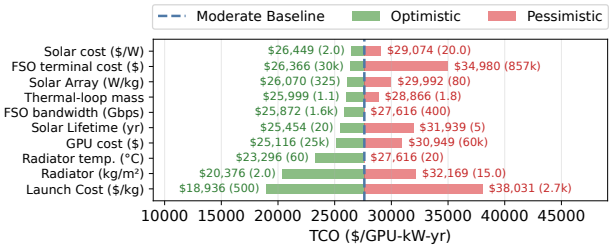


Fig. 7. Parameter sensitivity at the moderate minimum, ranked by cost reduction toward optimistic inputs. Launch price and radiator areal density provide the largest reductions.

mizing satellite size moves across tiers, but the preferred architectural scale remains bounded. All tiers avoid both accelerator-per-satellite designs and single mega-structures, instead favoring ISS-class satellites with hundreds to thousands of GPUs. Under moderate assumptions, the minimum is at 512 GPUs/satellite. Under the optimistic AI1-frontier tier, 1k and 5k GPUs per satellite are nearly tied (<1% apart); the apparent 5k minimum occurs because the satellite remains just under the 100 t single-launch threshold. Under pessimistic assumptions, 1k and 5k are again nearly tied, but at much higher absolute cost. The preferred scale regime therefore remains large enough to amortize per-satellite overhead and host an NVLink domain, yet small enough to avoid mega-structure assembly, maneuver and single-failure penalties.

The mass-to-power ratios determine much of the TCO variation. Minimum TCO is ∼$6.5k/GPU-kW-yr in the optimistic tier, ∼$27.6k/GPU-kW-yr in the moderate tier and ∼$238k/GPU-kW-yr in the pessimistic tier. Relative to terrestrial systems with the same compute prices and refresh lifetimes, these minima are 1.12×, 3.34× and 12.07×, respectively.

Ranking the parameters by cost impact at the moderate minimum identifies which uncertainties matter most. Launch price is the largest cost lever, but it does not determine the scale optimum by itself. With moderate masses, even reducing launch to the Starship target leaves the minimum at 512 GPUs/satellite because larger satellites cross the single-launch threshold. Radiator areal density provides the next-largest reduction, followed by radiator temperature, GPU price and solar-array lifetime. Thermal-loop mass and array specific power also matter, reflecting the coupled mass of powered and cooled compute. Terminal bandwidth remains performance-critical in §V (Figure 7).

Excluding GPU and server procurement and refresh, while retaining their orbital mass, power, delivery and servicing, leaves the optimistic tier at ∼1.9× terrestrial facility costs; the moderate and pessimistic tiers remain much higher. Because the optimistic tier combines upper-envelope mass density with low component costs, we treat it as a lower-bound TCO case, unsuitable as a central production-cost estimate.

V. Training Performance and Reliability

The cost model establishes that orbital compute carries a significant $/kW-yr premium, but cost per kilowatt-year only prices sustained power. We therefore evaluate *training goodput*, the rate at which orbital GPUs produce correct training steps over the mission. Inter-satellite optical links can bottleneck distributed training. Orbital faults further reduce goodput by wasting time, corrupting updates, or permanently removing capacity.

These communication results apply to any architecture that spreads training across spacecraft, whether in disaggregated formations (Suncatcher) or constellations (SpaceX). A monolithic structure like StarCloud keeps the fabric wired and avoids the inter-satellite terms but still faces the reliability losses below. The key design parameter is the per-satellite NVLink domain. A larger domain handles more training traffic locally, reducing communication over ISLs.

We validate the communication models with ASTRO-Sim, which executes synthetic Llama-scale collective traces through ASTRA-sim's analytical network backend [23] on the same NVLink+FSO topologies as the closed-form model. Figures 9, 10 and 11 plot analytical curves with simulation markers; the text calls out the few places where simulation exposes additional overhead beyond the closed form.

A. Inter-Satellite Link Constraints

§II-C gives the optical-link hardware model. Here we use only the training-facing parameters, namely bandwidth, propagation latency, switching time and bandwidth-delay product (BDP). Under the two-way aggregate convention in §II-C, we model the Starlink-class 100 Gbps long-range case at 6.25 GB/s per direction and Suncatcher's 1.6 Tbps bidirectional DWDM case at 100 GB/s per direction.

Bandwidth. Dense training stresses full-gradient-equivalent synchronization across data-parallel ranks. A 100 Gbps two-way aggregate ISL is already 8× below NDR400 InfiniBand and 144× below an NVLink 5 GPU port (Table I). Higher-rate close-formation FSO helps only if enough terminals or optical rails expose that bandwidth to the GPUs. Per-satellite NVLink domain size therefore governs the all-reduce penalty alongside link rate.

Latency. Range matters even before optical power loss reduces bandwidth. Propagation latency equals distance/c, giving 1 µs at 300 m, 167 µs at 50 km and 667 µs at 200 km.

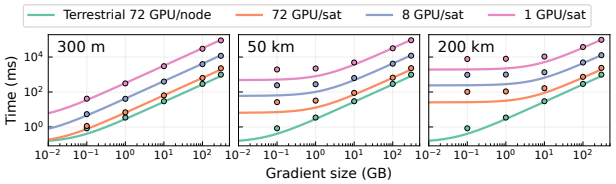


Fig. 8. Ring all-reduce time vs. gradient size across three propagation-latency regimes, with orbital-link bandwidth fixed at $6.25\,\mathrm{GB/s}$ per direction. Lines show the closed-form model; markers show ASTRO-SIM. The terrestrial NVL72 baseline (green) provides a reference.

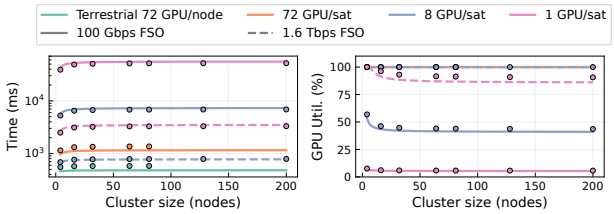


Fig. 9. Ring all-reduce time (left) and GPU utilization (right) vs. cluster size for a 175 GB gradient at 300 m spacing. Lines show the analytical model; markers show ASTRO-SIM (median deviation 6%). Larger NVLink domains split the gradient over more per-GPU FSO links.

We fix bandwidth at $6.25\,\mathrm{GB/s}$ per direction to isolate this effect. Close formation is nearly bandwidth-limited, while wider spacings add 10% or more and inflate BDP to $\sim$2035 kB at 50 km, 16$\times$ larger than InfiniBand. Because real optical bandwidth also changes with range, this comparison isolates latency and BDP without a full link-budget model. In this regime ASTRO-SIM agrees within 6% for bandwidth-dominated 100–300 GB transfers, but tiny-gradient cases at 50–200 km exceed the closed-form line by 2.5–4$\times$ (Figure 8).

Switching. Switching constrains scheduling as well as latency. Schedulers can plan static rings, trees, or pipelines around slowly reconfigurable circuits, but MoE expert routing and attention scatter-gather change sender-receiver pairs at training-step timescales. Re-targeting links takes 10–30 seconds, too slow to adapt the topology at those timescales; §V-C quantifies the bandwidth and latency penalties when expert traffic uses existing paths.

B. All-Reduce and GPU Utilization

We model gradient-only ring all-reduce, the most favorable communication pattern for orbital training. Real frontier runs also use ZeRO/FSDP-style parameter gathering and mixed collectives [33], [70]; our estimates are therefore a lower bound on inter-satellite communication overhead.

Following the hierarchy in §II-B, TP operates within each satellite's NVLink domain while DP all-reduce crosses FSO. We compare per-GPU provisioning (one $100\,\mathrm{Gbps}$ two-way aggregate FSO terminal per GPU, or $6.25\,\mathrm{GB/s}$ per direction) as the orbital analogue of the one-NIC-per-GPU rail networks in terrestrial clusters. This is deliberately the most favorable interconnect assumption for orbit. Per-GPU provisioning maximizes aggregate ISL bandwidth, making the utilization penalties below conservative. The TCO model (§IV) charges for the terminals' mass and cost as bandwidth scales with G. This provisioning does not reduce the total gradient volume that leaves each satellite.

We evaluate NVL72, DGX 8-GPU and flat 1-GPU-per-satellite configurations against a terrestrial NVL72 baseline for a canonical 700B model (PP=8, 175 GB gradient per stage) across 81 nodes. ASTRO-SIM replays the matched hierarchical trace (compute, local reduce-scatter, inter-satellite all-reduce and all-gather); median simulation–model deviation is 6% in the cluster-size sweep and 14% in the bandwidth/speedup sweep.

With an NVL72 rack per satellite, the 8$\times$ FSO bandwidth disadvantage relative to InfiniBand produces only a 2.4$\times$ slowdown. Without any NVLink domain, the full 175 GB gradient funnels through one $6.25\,\mathrm{GB/s}$ per-direction link, yielding a 115.3$\times$ penalty. Even Suncatcher's $1.6\,\mathrm{Tbps}$ link only brings the flat configuration from 115.3$\times$ to 7.2$\times$. The limiting factor is therefore *per-satellite compute density*, which determines how much aggregate ISL bandwidth per-GPU provisioning creates.

GPU utilization follows from how much of the all-reduce hides under compute. We grant communication *perfect* overlap with computation, so a step costs whichever of compute and communication is larger and utilization is $\eta = T_{\text{compute}}/\max(T_{\text{compute}}, T_{\text{comm}})$. The plotted utilization is an upper bound because real training adds the omitted collectives above and exposes more communication under tighter 3D-parallel schedules [36]. Figure 9 shows that terrestrial and orbital NVL72 both fully overlap communication, while the DGX and flat-satellite cases fall to $\sim$41% and $\sim$5% utilization. The utilization gap ranges from 1$\times$ to $\sim$20$\times$ depending almost entirely on NVLink domain size.

Sensitivity to technology improvements. Both GPU compute speed and ISL bandwidth will improve. Figure 10 sweeps 1–4$\times$ faster GPUs and 1–16$\times$ ISL bandwidth for the same 175 GB gradient. Faster GPUs *reduce* utilization unless communication improves too, because the overlap window shrinks while the gradient transfer stays fixed.

Per-satellite NVLink domain size dominates this effect. A flat 1-GPU/satellite system remains below 10% utilization at current ISL bandwidth; even 16$\times$ ISL drops below 25% at 4$\times$ GPU speed. An 8-GPU domain can use 16$\times$ ISL at current GPU speed but falls behind as GPUs improve. Only NVL72 has enough local partitioning for 4$\times$ ISL ($400\,\mathrm{Gbps}$ two-way aggregate per GPU) to sustain near-100% utilization even at 4$\times$ GPU speedup.

Pipeline parallelism introduces a smaller, separate overhead. Each inter-stage activation transfer ($\sim$0.3 GB per micro-batch) takes $\sim$48.3 ms over the modeled $6.25\,\mathrm{GB/s}$ per-direction FSO link, adding 12.9% per stage versus $<$0.1% over NVLink. With PP=8, the exposed pipeline communication reaches $\sim$340 ms per step. End-to-end ASTRO-SIM placement sweeps reinforce the mapping rule. Orbital fabrics penalize cross-satellite all-reduce and all-to-all more than pipeline traffic.

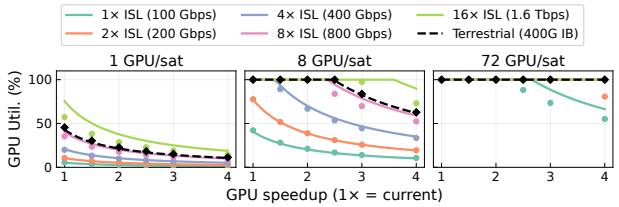


Fig. 10. GPU utilization vs. GPU speedup and ISL bandwidth, by NVLink domain size (50 km spacing, 175 GB gradient, 1,440 total GPUs). Markers show ASTRO-SIM (median deviation 14%). Only the 72-GPU domain sustains near-terrestrial utilization at achievable ISL improvements (4×).

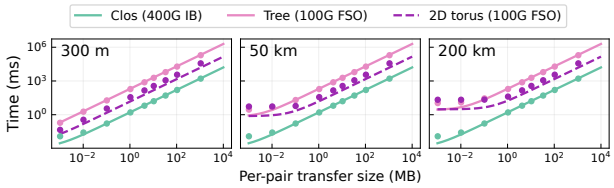


Fig. 11. MoE all-to-all time vs. per-pair transfer size across three distance regimes, comparing a terrestrial Clos (400 Gbps IB), constellation tree (k=4 FSO) and 2D torus (100 Gbps two-way aggregate FSO). Markers show ASTRO-SIM. Clos/tree agree within 8%; the analytical torus is an ideal contention-free bound; simulation places it ∼2.4× higher.

C. Topology and Bisection Bandwidth

Constellation topology compounds the per-link deficit. A terrestrial Clos/fat-tree provides $O(N)$ bisection bandwidth because N nodes each contribute full port bandwidth across the midpoint [71]. In contrast, the modeled neighbor-linked 2D torus with k=4 FSO terminals per satellite achieves only $O(\sqrt{N})$, with $B_{\text{bisection}} = \frac{k \cdot \sqrt{N}}{2} \cdot b_{\text{link}}$.

At N=10 000 nodes with 100 Gbps two-way aggregate links, this yields 20 Tbps two-way bisection versus 4000 Tbps for a terrestrial Clos, a 200× gap that grows as $O(\sqrt{N})$. All-reduce and all-to-all collectives therefore face both per-link and topological bandwidth deficits.

The missing component is a switch tier. Existing proposals equip satellites as identical compute nodes, so forwarded collective traffic competes with local training for link bandwidth. Even rail-only topologies [40] presuppose packet-switched 400 Gbps rails and intra-node NVSwitch, neither of which exists between satellites.

MoE expert routing exposes this concretely. Expert-parallel groups typically span 8–64 GPUs and stay within NVLink on a full NVL72 rack, but scaling EP beyond 72 needs all-to-all across FSO, where limited bandwidth and multihop latency impose substantial communication overhead.

For architectures with fewer GPUs per satellite, including Suncatcher-like small-accelerator nodes or 8-GPU DGX-class satellites, the EP group must span spacecraft. Here we use tree and torus as logical communication graphs to bound all-to-all performance without assuming freely maintainable compact formations; §III-E shows that a rigid physical torus would require continuous forcing. At 300 m spacing with 4–32 MiB per-pair transfers, the simulated tree and 2D torus are approximately 120× and 23× slower than Clos, respectively (Figure 11). Simulation confirms the Clos and tree curves within 8%. The analytical torus provides an optimistic contention-free bound; with real link contention the simulated torus runs ∼2.4× slower still.

Since MoE is a dominant scaling approach for frontier models [34], [35], this imposes a structural constraint. Small satellites force expert groups to span spacecraft; even NVL72 satellites support only EP≤72 before expert traffic spills onto FSO. These bandwidth and latency penalties favor keeping expert groups local to a satellite.

D. Data Distribution

Data must reach the satellites that consume it. Launching 1 PB on SSDs weighs only 1.4 kg, but it arrives at one satellite and must fan out via ISLs. For an 81-node constellation, a tree broadcast (4 terminals/satellite) or linear relay (2 terminals/satellite) takes about 1.9 days at 50 Gbps per direction with persistent links and ideal pipelining, versus 5.6 hours for a terrestrial fat-tree. Subsequent updates still require uplink capacity. Bulk redistribution competes with workload communication for ISL bandwidth.

E. Reliability and Training Efficiency

§III-B separates orbital faults into two classes. Radiation upsets are transient. Detected errors trigger rollback to a checkpoint, while silent errors corrupt model state without a signal. In contrast, hardware failures that a terrestrial cluster would repair become permanent capacity loss without servicing. We quantify both and combine them as training goodput.

Transient faults. At the 650 km SSO design point, AP9 gives an orbital proton flux of $\sim 14\,\mathrm{p/cm^2/s}$ behind 5 mm Al [72]. We estimate the resulting HBM upset rate from beam tests of *modern* memories. We show the legacy DRAM extrapolation, which gives a per-GPU uncorrectable mean time between failures (MTBF) of ∼113 s, only as a reference line. Three modern anchors bracket the rate. HBM2 under plain SECDED gives ∼5 days per GPU [55]. Scaling Trillium beam data by flux gives ∼104 days; a dose-scaled midpoint gives ∼13 days. SECDED corrects the single-bit majority; the remaining ∼26% split into detected-uncorrectable and silent cases [55]. For the detected-uncorrectable class, fail-stop behavior permits checkpoint/restart. We apply the first-order Young/Daly interval [73] $T = \max(C, \sqrt{2CM})$ to this class alone, estimating efficiency as $\eta \approx (T - C)/(T + T^2/2M + TR/M)$, where M is the system mean time between uncorrectable failures (the per-GPU MTBF above divided by GPU count N), $C = 2.5\,\mathrm{s}$ [74] is the checkpoint cost and $R \approx 30 + 12 \log_2 N$ s [75] is recovery time. This estimate assumes instant detection, durable in-memory checkpoints and restart on the same hardware. It excludes additional corruption of checkpoint state stored in irradiated HBM. Against the terrestrial MTBF (∼29,538 GPU-hours; η ∼94%/77% at 16k/100k, matching Meta's Llama 3 [76]), the orbital band costs ∼1.4–10× at Llama-3 scale and ∼3–75× at frontier

scale (Figure 12a); only under the legacy rate does the formula leave its regime ($M \leq C$ at ∼45 GPUs).

The silent subset threatens correctness. ∼13% of uncorrectables alias past ECC undetected [55], [56], so no error signal triggers a rollback. If a corrupted value participates in gradient all-reduce, this collective distributes the corruption to every replica. Averaging can dilute a mantissa flip (a $\leq 2\times$ element change, divided by N), but exponent flips can dominate the normal gradient even after division by $N{=}10^5$ [77]. At these rates, a frontier cluster (10^5 GPUs) injects of order one silent corruption per 10–200 training steps. Fault-injection studies observe degraded converged loss and divergence at the upper end of this frequency range [77], [78]; production SDC studies similarly find model-state drift that may not surface as an immediate loss spike [79].

Mitigations. Mitigations target different parts of the fault chain. For detected errors, lower checkpoint and recovery costs reduce C and R. Asynchronous-island training, including DiLoCo, bounds the failure domain at a convergence cost, but none of these mitigations addresses silent correctness failures. For silent errors, stronger coding (e.g., chipkill ECC, which raises M by $\sim 100\times$) and radiation hardening cut the aliasing rate or make some silent errors detectable, though a residual silent fraction remains under SECDED [55]. Algorithm-based fault tolerance, redundant execution and robust aggregation can mask corruption only by adding compute and communication overhead. The remaining frontier-scale terms, namely silent corruption and the permanent attrition below, remain beyond checkpointing's scope.

Permanent faults and capacity decay. Without servicing, three mechanisms reduce capacity over the mission. *Lemon attrition* at ∼1.2%/month [50] removes ∼13.5% of nodes per year, leaving a 5-year mission at ∼48.5% surviving capacity, ∼71.1% of nameplate GPU-years and an effective $/GPU-year increase of $\sim 1.41\times$ (Figure 12b). At a 650 km SSO dose rate, *TID wear-out* begins near year 4.0, cutting combined 5-year survival to ∼7.7% and average capacity to ∼59.1% ($\times 1.69$). Finally, *remap exhaustion* limits recovery. Each uncorrectable error retires an HBM row against a hard cap of 512 per GPU [52], [80]. Errors exhaust this budget in ∼7.6 years at the pessimistic modern rate.

Mission duration therefore enters the reliability trade-off. With TID included, average delivered capacity is ∼79% (3 yr), 70% (4 yr) and 59% (5 yr). Shorter missions raise launch amortization, while longer ones accumulate more dose-driven wear-out. Maintaining flat capacity through replacement adds +34% (modular servicing) to +67% (full-satellite retirement) over the scheduled mass flow in the cost model (§IV). Training goodput is the product $N(t)\cdot\eta(N(t))$ of surviving capacity and per-step efficiency, so transient and permanent reliability losses combine with the communication and cost penalties above.

F. The Training Constellation Paradox

For the flat one-GPU-per-satellite configuration, training occupies under 6% of each step; the remainder is exposed all-reduce time (Figure 9). Figure 13 bounds its reuse using an

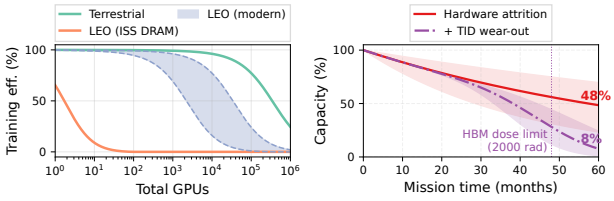


Fig. 12. Reliability losses in orbital training. (a) Training efficiency under a first-order checkpointing model (Young/Daly, $C{=}2.5$ s) for detected-uncorrectable errors only. Terrestrial clusters sustain >94% at 16k GPUs [76]; the orbital band spans independent beam tests of modern memories (∼5–104 d per-GPU MTBF). A reference line shows the legacy DRAM extrapolation. (b) Permanent capacity decay over a 5-year mission without servicing. Lemon attrition alone leaves ∼48.5% of capacity ($\times 1.41$ effective cost); adding TID wear-out past its ∼4.0-year onset cuts survival to ∼7.7% ($\times 1.69$).

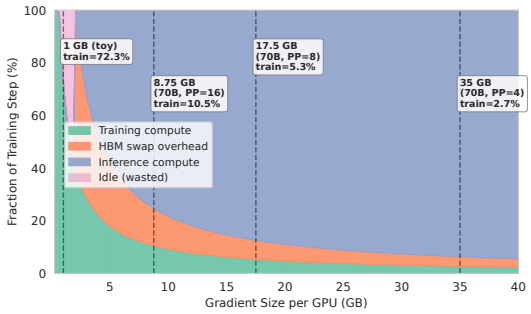


Fig. 13. Optimistic reuse of exposed all-reduce time under the assumptions in §V-F.

illustrative 70B workload (300 ms compute, 67% overlap and an 81-way 100 Gbps ring). Assuming the workload occupies 76 GB of an 80 GB H100 [81], swapping a 7 GB INT8 model at a modeled sustained PCIe bandwidth of 32 GB/s moves 13 GB and takes 406 ms. With a contiguous, contention-free window, the 17.5 GB point spends 5.3% on training and 7.2% on swapping, leaving at most 87.5% for inference. This recovers idle compute without increasing training throughput.

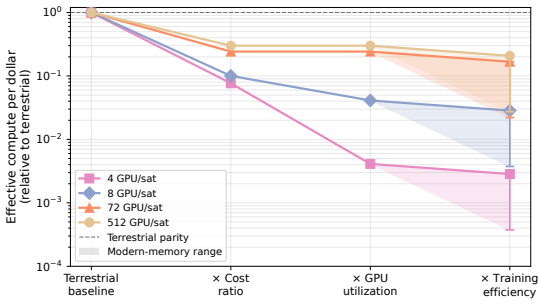


Fig. 14. Effective compute per dollar relative to terrestrial, combining moderate TCO costs with analytical communication and checkpoint/recovery estimates for 18,432 GPUs. All configurations use PP= 8; TP= 1/8/72/64 for 4/8/72/512 GPUs/satellite, respectively, with all eight pipeline stages local in the 512-GPU case. Shading spans the modern-memory range in Figure 12a; lines follow the favorable estimate. We exclude pipeline overhead, permanent attrition and silent-corruption losses.

VI. The Real Niche: Edge Inference

Training asks orbit to supply cheap, tightly synchronized, reliable compute; the preceding sections show that it supplies none of those at terrestrial scale. The better question is which inference workloads actually benefit from being in space. The answer depends on data locality. Serving terrestrial users is a poor fit, while inference on space-generated data is the natural niche.

Interactive serving of terrestrial demand, the workload orbital-datacenter proposals often foreground [10], is the worst fit. It has strict time-to-first-token targets in the low hundreds of milliseconds [82], yet the serving satellite leaves view within ∼150–360 s. Handoff or inter-satellite forwarding therefore lands inside an already-tight latency budget. Multi-turn sessions also carry a KV cache tied to a moving host [83], [84]; if a handoff evicts the cache, the accelerator recomputes context it already served.

Batched serving of terrestrial demand avoids this state problem because each request finishes in one pass. It also removes the reason to be in orbit. The users and data are on the ground, so the system uplinks inputs and downlinks outputs for work that executes just as well in a terrestrial datacenter.

Inference on space-generated data is the workload that belongs in orbit. Satellites already generate data that is expensive to downlink. Next-generation LEO imaging constellations produce ∼1 TB/day per satellite against $\sim 150\,\mathrm{Mbit/s}$ of downlink during ground-station passes [17]; Sun–Earth L2 instruments face 5-second round-trip latency; Mars missions wait 4–24 minutes one way. On-board inference processes data at the source and transmits compact results. The relevant models, including cloud detection, change detection and anomaly flagging, span 10^6–10^9 parameters and fit on a single accelerator, so they need no all-reduce, no cross-satellite pipeline stage and no dense circuit-switched FSO fabric (§V). Operators can also batch requests and throttle execution to match available power and cooling. Radiation remains a concern, but an error in request-local state may affect only one request. In training, corrupted model state can affect the entire cluster (§V-E).

ESA's Φ-Sat-1 demonstrated this pattern with an Intel Movidius VPU that ran cloud-detection inference on board and filtered ∼70% of unusable imagery before downlink [85]. WorldFloods separates on-board flood mapping from ground-based retraining with downlinked images [86]. China's Three-Body Computing Constellation scales in the same direction. Its first 12 satellites each carry an 8B-parameter model for on-board data processing, targeting 2,800 satellites and 1,000 POPS of aggregate inference throughput [15]. At constellation scale, LEOEdge also schedules AI inference with satellite–ground cooperation and sizes each satellite's model to its on-board compute [87].

The resulting satellite prioritizes inference TOPS per watt over interconnect bandwidth. Designers size ISLs for data ingest; operators train models on the ground and upload them periodically. Removing the training fabric and NVLink-scale domains also removes much of the parasitic mass that hurts small training satellites (§IV), leaving a self-contained solar, radiator, accelerator and data-link node (Guideline 5, §VII-B).

VII. Case Studies

Table V applies the TCO model to three proposals that span the U-shaped cost curve of §IV-B. Suncatcher and SpaceX AI1 use the current TCO-model presets; StarCloud uses its claimed thin-film solar (1,000 W/kg) and structural assumptions [10]. The kg/kW row uses total electrical power, matching the TCO sensitivity sweep.

Suncatcher [6], [7] (left slope) uses 81 small satellites in close formation. Our preset models four accelerator-equivalents per satellite, comprising a $4.8\,\mathrm{kW}$ compute payload and $7.4\,\mathrm{kW}$ total load per node, with fixed thermal, bus/structure and FSO hardware accounting for ∼83% of mass and yielding 63.5 kg/kW on the total-power basis (98 kg/kW against compute power alone). Its value is mainly as a formation and link demonstrator. $1.6\,\mathrm{Tbps}$ DWDM can make small nodes viable for dense collectives, but is still no substitute for an on-satellite NVLink domain at training scale. **SpaceX AI1** [8], [61] (mass-density calibration point) offers a contrasting small-satellite reference, with a claimed $150\,\mathrm{kW}$ peak compute payload at roughly $70\,\mathrm{kW/t}$. Our AI1 preset reproduces a $2.16\,\mathrm{t}$ satellite at 12.5 kg/kW, with power and thermal hardware still ∼58% of mass. Its reported $110\,\mathrm{m}^2$ liquid radiator is below the $151\,\mathrm{m}^2$ our conservative $60\,^\circ\mathrm{C}$ assumption predicts but consistent with a hotter operating point. We therefore use AI1 as a claimed frontier for the optimistic tier, pending independent flight validation. **StarCloud-4** [4], [10] (right slope) uses a single $5\,\mathrm{GW}$ structure with wired NVLink and aggressive mass density (13.5 kg/kW), but its $16\,\mathrm{km}^2$ cross-section needs ∼91 debris maneuvers per day and $\sim 75\,855\,\mathrm{t}$ of propellant per year. The architectural design maximizes the wired scale-up domain, but its structural penalties make this scale impractical.

TABLE V
Per-satellite mass breakdown.

	Suncatcher		SpaceX AI1		StarCloud-4	
Component	**Mass**	**%**	**Mass**	**%**	**Mass**	**%**
Compute†	38 kg	8.1	491 kg	22.7	13.9 kt	20.7
Power§	42 kg	9.1	612 kg	28.3	5.0 kt	7.4
Thermal‖	163 kg	34.8	632 kg	29.3	39.5 kt	58.7
FSO/ISL	80 kg	17.1	32 kg	1.5	N/A	N/A
Bus/structure#	145 kg	30.9	393 kg	18.2	8.9 kt	13.2
Total	**468 kg**		**2.16 t**		**67 kt**	
Power/node	7.4 kW		173 kW		5 GW	
kg/kW	63.5		12.5		13.5	
Radiator area	11.6 m^2		151 m^2		7.9 km^2	

†Includes accelerator packages and server overhead. §Includes solar array and power conditioning. ‖Includes radiator panels and coolant loop. #Includes primary structure, ADCS, propulsion, tank and reserve propellant; StarCloud reports only aggregate structural mass.

Across all three, the mass and lifecycle cost of cooled compute in vacuum constrain feasibility alongside launch. Even the optimistic AI1-class tier does not establish a cost advantage over terrestrial compute; moderate assumptions widen the gap.

A. Formation/Network Co-Design

The case studies also reveal an architectural choice terrestrial clusters lack. Designers can choose the physical formation together with the workload mapping. DP creates closed collective groups, while PP creates nearest-neighbor chains, so the relevant question is whether the formation's short physical edges match the heavy logical edges of the training plan.

Figure 15 illustrates a DP×PP mapping. A string matches a single PP chain; a cylinder can map DP rings around its circumference and PP chains along its height, so radius and height instantiate the chosen degrees.

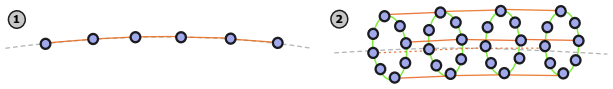


Fig. 15. Formation/parallelism co-design. Green links are data-parallel (DP), red links pipeline-parallel (PP). A string (1) is a one-dimensional pipeline chain; a cylinder (2) maps DP to its circumference and PP to its height, so radius and height instantiate the chosen DP×PP degrees.

ASTRO-SIM makes this distinction visible. In an 8-satellite, 8-GPU-per-satellite Llama-3-70B placement sweep with $100\,\mathrm{Gbit/s}$ ISLs, placing DP all-reduce across satellites is $78\times$ slower than the matching Clos baseline, while placing PP neighbor traffic across the same links is only $1.17\times$ slower. In a 32-GPU formation motivator with $100\,\mathrm{Gbit/s}$ per-direction ISLs, the cylinder mapping (DP=8, PP=4) completes a Llama-3-70B iteration in $52.9\,\mathrm{s}$; the string mapping (DP=1, PP=32) takes $137.1\,\mathrm{s}$.

The cylinder suits only some workloads. Our formation analysis (§III-E) favors free-fall strings and rings/cylinders over rigid meshes or tori. In this design, TP stays on-satellite and only DP and PP map onto the formation. A denser rigid mesh would raise bisection but costs 23 km/s/yr versus 28 m/s/yr for the evaluated free-fall geometries. Designers must therefore choose formation, link budget and parallelism together; the optimal mapping depends on the workload.

B. Design Guidelines

The case studies and co-design constraints suggest five design guidelines. **(1) Build ISS-class satellites.** The moderate TCO minimum occurs at ∼500 GPUs per satellite; smaller nodes amortize terminals over too few GPUs, while larger ones incur debris, propellant and single-failure penalties. **(2) Reduce the mass-to-power ratio.** The first-order breakeven requires ≤11.3 kg/kW, below the Starlink-class range. Thermal hardware contributes a third to half of satellite mass, making lighter heat rejection an important target, as in AI1's claimed liquid radiator [61]. **(3) Keep tensor parallelism on satellite and choose cross-satellite parallelism by link regime.** If cross-satellite bandwidth is scarce, prefer pipeline placement between neighbors. Use data-parallel all-reduce across ISLs only with a close-formation, multi-Tbps fabric. **(4) Keep MoE local or train dense models.** Limited bandwidth and multihop latency impose substantial overhead on cross-satellite expert communication. **(5) For sustained compute, target inference.** Inference needs no cluster-wide synchronization and confines radiation errors to one request, so the satellite carries ingest links and runs ground-trained models without a training fabric.

Guidelines 1–4 describe the least-bad training system; Guideline 5 differs because edge inference sheds the fabric and NVLink domain for a smaller node. Latency-tolerant training (DiLoCo [20]), pipeline-heavy placement, or optical packet switching could soften the interconnect constraints, but none lifts the kg/kW floor.

VIII. Conclusion

Solar energy, radiative cooling and mounting terrestrial constraints motivate orbital AI. However, energy alone does not make a datacenter. Subsystem lifetimes, servicing and thermal/power mass keep orbital compute more expensive than terrestrial capacity in our modeled scenarios. Approaching parity requires optimistic assumptions across mass density, hardware prices and launch costs.

The least-bad training architecture co-locates a full NVL72 domain on each satellite, absorbing enough traffic for near-terrestrial utilization while keeping debris avoidance and single-satellite loss manageable. Mega-structures trade the interconnect bottleneck for debris maneuvers, propellant and single-point failure risk. Limited bandwidth and multihop latency impose substantial overhead on cross-satellite MoE communication. Radiation and hardware failures without servicing at datacenter cadence reduce training efficiency. Launch throughput limits capacity to ∼34.7 GW, a ceiling terrestrial demand surpasses by ∼2028.

Figure 14 quantifies the compound effect. Cost premium, FSO utilization loss and radiation-driven inefficiency are largely independent and multiply, producing $\sim 4.8\times$ to $\sim 2{,}700\times$ less effective compute per dollar across the evaluated designs and modern-memory bounds under moderate costs. Edge inference on space-generated data better matches orbit (§VI). It avoids synchronization, a training fabric and shared-state corruption, using only the ISL capacity needed to ingest data and ground-trained models.

Acknowledgment

We thank Rakesh Kumar, Krisztián Flautner, and Mosharaf Chowdhury for insightful discussions. We thank the reviewers for their valuable input.

References

[1] Epoch AI, "Trends in machine learning compute," https://epochai.org/trends, 2025.

[2] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[3] SpaceX, "SpaceX files FCC application for 1 million satellite orbital data center constellation," FCC filing, 2026, filed January 30, 2026, following the $1.25 trillion SpaceX–xAI merger announced February 2, 2026.

[4] P. Singh, "'Greetings, earthlings': Nvidia-backed Starcloud trains first AI model in space as orbital data center race heats up," CNBC, https://www.cnbc.com/2025/12/10/nvidia-backed-starcloud-trains-first-ai-model-in-space-orbital-data-centers.html, 2025, starcloud-1 launched Nov. 2025 with NVIDIA H100; trained NanoGPT and ran Google Gemma inference on-orbit.

[5] G. Bensinger, "Orbital data centers "pretty far" from reality," Reuters, Feb. 2026, aWS CEO, speaking at Cisco AI Summit: "There are not enough rockets to launch a million satellites yet." "It is just not economical.".

[6] B. A. y Arcas, T. Beals, M. Biggs, J. V. Bloom, T. Fischbacher, K. Gromov, U. Köster, R. Pravahan, and J. Manyika, "Towards a future space-based, highly scalable ai infrastructure system design," 2025. [Online]. Available: https://arxiv.org/abs/2511.19468

[7] Planet Labs, "Planet to build and operate advanced space platform for Google's Project Suncatcher moonshot," Press release, 2025, two prototype satellites planned for early 2027.

[8] Space Exploration Technologies Corp., "Form s-1 registration statement," U.S. Securities and Exchange Commission (EDGAR), CIK 1181412, accession 0001628280-26-036936, 2026, orbital AI compute section: "targeted to generate approximately 100 kW of compute power per ton"; first-gen deployment "as early as 2028." Extracted text archived at sources/spacex_s1.txt. [Online]. Available: https://www.sec.gov/Archives/edgar/data/1181412/000162828026036936/spaceexplorationtechnologi.htm

[9] A. Jones, "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/, 2025, 12 satellites launched May 2025 for Three-Body Computing Constellation by ADA Space and Zhejiang Lab.

[10] E. Feilden, A. Oltean, and P. Johnston, "Why we should train AI in space," Lumen Orbit (now StarCloud) White Paper v1.03, 2024, proposes GW-scale orbital data center with 4 km $\times$ 4 km solar array in dawn-dusk sun-synchronous orbit.

[11] Thales Alenia Space, "ASCEND: Advanced space cloud for european net zero emission and data sovereignty," European Commission Horizon Europe feasibility study, 2024, 50 kW proof of concept targeted for 2031; 1 GW orbital data center by 2050.

[12] Axiom Space, "Axiom space orbital data center nodes," Press release, 2026, two ODC nodes launched to LEO January 2026; AxDCU-1 prototype deployed on ISS fall 2025.

[13] N. Bleier and R. Kumar, "Space datacenters," *IEEE Computer*, 2026. [Online]. Available: https://crucible.eecs.umich.edu/papers.html?id=23

[14] G. Giuffrida, L. Diana, F. de Gioia, G. Benelli, G. Meoni, M. Donati, and L. Fanucci, "Cloudscout: A deep neural network for on-board cloud detection on hyperspectral images," *Remote Sensing*, vol. 12, no. 14, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/14/2205

[15] Space Daily, "China tests AI satellite swarm for space-based computing," Space Daily, https://www.spacedaily.com/reports/China_tests_AI_satellite_swarm_for_space_based_computing_999.html, 2026, on-orbit results: 8B-parameter models, inter-satellite links demonstrated across 6 spacecraft.

[16] B. Denby and B. Lucia, "Orbital edge computing: Nanosatellite constellations as a new class of computer system," in *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, ser. ASPLOS '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 939–954. [Online]. Available: https://doi.org/10.1145/3373376.3378473

[17] N. Bleier, M. H. Mubarik, G. R. Swenson, and R. Kumar, "Space microdatacenters," in *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2023.

[18] N. Bleier, R. Eason, M. Lembeck, and R. Kumar, "Architecting space microdatacenters: A system-level approach," in *Proceedings of the IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2025, pp. 1304–1319.

[19] J. J. Garau-Luis, S. Eiskowitz, N. Pachler, E. Crawley, and B. Cameron, "Towards autonomous satellite communications: An ai-based framework to address system-level challenges," 2021. [Online]. Available: https://arxiv.org/abs/2112.06055

[20] A. Douillard, Q. Feng, A. A. Rusu, R. Chhaparia, Y. Donchev, A. Kuncoro, M. Ranzato, A. Szlam, and J. Shen, "Diloco: Distributed low-communication training of language models," 2024. [Online]. Available: https://arxiv.org/abs/2311.08105

[21] O. Chabra, C. Li, K. Hsieh, S. Segarra, B. Arzani, P. Olsen, and R. Chandra, "OrbitalBrain: A Distributed Framework for Training ML Models in Space," in *1st New Ideas in Networked Systems (NINeS 2026)*, ser. Open Access Series in Informatics (OASIcs), K. Argyraki and A. Panda, Eds., vol. 139. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2026, pp. 5:1–5:32. [Online]. Available: https://drops.dagstuhl.de/entities/document/10.4230/OASIcs.NINeS.2026.5

[22] E. Ding, B. Lyu, B. Kataria, and R. Singh, "Photonic rails in ml datacenters with opus," 2026. [Online]. Available: https://arxiv.org/abs/2602.12521

[23] S. Rashidi, S. Sridharan, S. Srinivasan, and T. Krishna, "Astra-sim: Enabling sw/hw co-design exploration for distributed dl training platforms," in *2020 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2020, pp. 81–92.

[24] C. Man, J. Park, H. Wu, H. Xu, S. Sridharan, and T. Krishna, "Scalable synthesis of distributed llm workloads through symbolic tensor graphs," in *Proceedings of the 53rd Annual IEEE/ACM International Symposium on Computer Architecture (ISCA '26)*, 2026.

[25] G. Kopp and J. L. Lean, "A new, lower value of total solar irradiance: Evidence and climate significance," *Geophysical Research Letters*, vol. 38, no. 1, 2011. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010GL045777

[26] M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby, "The national solar radiation data base (nsrdb)," *Renewable and Sustainable Energy Reviews*, vol. 89, pp. 51–60, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136403211830087X

[27] World Economic Forum, "Nature positive: Role of the technology sector," World Economic Forum, Tech. Rep., dec 2025. [Online]. Available: https://www.weforum.org/publications/nature-positive-role-of-the-technology-sector/

[28] I. Arciniegas Rueda, A. Najera, A. Smidt, R. Wang, H. M. Park, H. van Soest, and D. Gill, "Expanding u.s. net available power capacity by 2030: Barriers and solutions," RAND Corporation, Tech. Rep. RR-A3845-2, 2025. [Online]. Available: https://www.rand.org/pubs/research_reports/RRA3845-2.html

[29] S. Narasimhan, A. Johar, J. Motani, S. Lavin, A. Jain, and S. Mahajan, "Ai data center location attractiveness index," Kearney, Tech. Rep., mar 2025. [Online]. Available: https://www.kearney.com/industry/technology/article/ai-data-center-location-attractiveness-index

[30] C. García-Llorente and I. Olmeda, "The environmental blind spot of ai policy: Energy, infrastructure, and the systematic externalization of sustainability," *Sustainability*, vol. 18, no. 3, 2026. [Online]. Available: https://www.mdpi.com/2071-1050/18/3/1633

[31] Pew Research Center, "What we know about energy use at u.s. data centers amid the ai boom," Report, 2025, uS data centers consumed 183 TWh in 2024 ($\sim$4% of national electricity); projected to more than double (426 TWh) by 2030.

[32] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-lm: Training multi-billion parameter language models using model parallelism," 2020. [Online]. Available: https://arxiv.org/abs/1909.08053

[33] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He, "Zero: memory optimizations toward training trillion parameter models," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '20. IEEE Press, 2020.

[34] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res.*, vol. 23, no. 1, Jan. 2022.

[35] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, "GShard: Scaling giant models with conditional computation and automatic sharding," in *International Conference on Learning Representations (ICLR)*, 2021.

[36] D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley, M. Patwary, V. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro, A. Phanishayee, and M. Zaharia, "Efficient large-scale language model training on gpu clusters using megatron-lm," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3458817.3476209

[37] NVIDIA Corporation, "NVLink and NVSwitch: Fastest HPC data center platform," https://www.nvidia.com/en-us/data-center/nvlink/, 2024,

nVLink 5 provides 1.8 TB/s bidirectional bandwidth per GPU in NVL72 configurations.

[38] ——, "DGX GB200 user guide: Hardware," https://docs.nvidia.com/dgx/dgxgb200-user-guide/hardware.html, 2024, 72 Blackwell GPUs, 120 kW per rack, NVSwitch interconnect.

[39] InfiniBand Trade Association, "InfiniBand NDR400 specification," Specification document, 2024, 400 Gbps per port, sub-microsecond latency.

[40] W. Wang, M. Ghobadi, K. Shakeri, Y. Zhang, and N. Hasani, "Rail-only: A low-cost high-performance network for training llms with trillion parameters," in *2024 IEEE Symposium on High-Performance Interconnects (HOTI)*, 2024, pp. 1–10.

[41] Uptime Institute, "Uptime institute global data center survey 2024," https://uptimeinstitute.com/uptime_assets/7425ec68d479c5d78a743df94a79b114ed9f9c73f13b6460949d2b8e73373209-GA-2024-07-uptime-institute-global-data-center-survey-results-2024.pdf, 2024, average rack density 8 kW across 879 respondents.

[42] M. Blake, M. Hsu, I. Goldwasser, H. Petty, and J. Huntington, "NVIDIA 800 VDC architecture will power the next generation of AI factories," NVIDIA Technical Blog, https://developer.nvidia.com/blog/nvidia-800-v-hvdc-architecture-will-power-the-next-generation-of-ai-factories/, 2025, 800 VDC for 100 kW–1 MW racks; production with NVIDIA Kyber rack-scale systems in 2027.

[43] H. Kaushal and G. Kaddoum, "Optical communication in space: Challenges and mitigation techniques," *Commun. Surveys Tuts.*, vol. 19, no. 1, p. 57–96, Jan. 2017. [Online]. Available: https://doi.org/10.1109/COMST.2016.2603518

[44] M. Gregory, F. Heine, H. Kämpfner, R. Lange, M. Lutzer, and R. Meyer, "Coherent inter-satellite and satellite-ground laser links," vol. 7923, 02 2011, pp. 792 303–792 303.

[45] E. J. Arevalo, "Spacex reveals operation of over 8,000 'space lasers' across starlink satellite constellation – enabling faster internet," Tesmanian, Sep. 2023, over 8,000 space lasers across 4,827 satellites as of September 2023. [Online]. Available: https://www.tesmanian.com/blogs/tesmanian-blog/starlink-video

[46] SpaceX, "Starlink direct-to-cell and v2 mini satellite specifications," FCC filings and press materials, 2024.

[47] A. U. Chaudhry and H. Yanikomeroglu, "Laser intersatellite links in a Starlink constellation: A classification and analysis," *IEEE Vehicular Technology Magazine*, vol. 16, no. 2, pp. 48–56, 2021.

[48] Mynaric AG, "CONDOR Mk3 — optical communication terminal," Product datasheet, 2024, laser ISL terminal; 100 Mbps–100 Gbps; range up to 6,500 km.

[49] D. G. Gilmore, *Spacecraft Thermal Control Handbook*. The Aerospace Press, 2002, radiative cooling is the only heat rejection mechanism in vacuum.

[50] A. Kokolis, M. Kuchnik, J. Hoffman, A. Kumar, P. Malani, F. Ma, Z. DeVito, S. Sengupta, K. Saladi, and C.-J. Wu, "Revisiting reliability in large-scale machine learning research clusters," in *2025 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2025, arXiv:2410.21680.

[51] G. Ostrouchov, D. Maxwell, R. A. Ashraf, C. Engelmann, M. Shankar, and J. H. Rogers, "GPU lifetimes on Titan supercomputer: Survival analysis and reliability," in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020, titan K20X GPU lifetimes; corrosion epidemic affected $\sim$59% of replaced GPUs.

[52] S. Cui, S. J. Mahmud, A. Patke *et al.*, "Characterizing GPU resilience and impact on AI/HPC systems," arXiv:2503.11901, 2025, nCSA Delta A100/H100 replacement rates; confirms 512-row HBM remap cap; in-field row-remap failures at single-digit remap counts.

[53] D. Sinclair and J. Dyer, "Radiation effects and COTS parts in SmallSats," in *Proceedings of the 27th Annual AIAA/USU Conference on Small Satellites*, 2013, "Careful COTS" approach to component selection and radiation testing for LEO missions.

[54] J. R. Srour and J. W. Palko, "Displacement damage effects in irradiated semiconductor devices," *IEEE Transactions on Nuclear Science*, vol. 60, no. 3, pp. 1740–1766, 2013.

[55] M. B. Sullivan, N. R. Saxena, M. O'Connor, D. Lee, P. Racunas, S. Hukerikar, T. Tsai, S. K. S. Hari, and S. W. Keckler, "Characterizing and mitigating soft errors in GPU DRAM," in *54th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-54)*, 2021, hBM2 on V100; 12.51 FIT/Gb; $\sim$26% of events defeat plain SECDED; 216 FIT residual SDC even with SECDED.

[56] Y. Bentoutou, "A follow-up of in-orbit observations of radiation-induced effects in commercial of the shelf memories on-board alsat-1," *Advances in Space Research*, vol. 48, no. 6, pp. 1147–1154, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0273117711003553

[57] W. H. Clohessy and R. S. Wiltshire, "Terminal guidance system for satellite rendezvous," *Journal of the Aerospace Sciences*, vol. 27, no. 9, pp. 653–658, 1960.

[58] H. Schaub and K. T. Alfriend, "J2 invariant relative orbits for spacecraft formations," *Celestial Mechanics and Dynamical Astronomy*, vol. 79, no. 2, pp. 77–95, 2001.

[59] Turner & Townsend, "Data center cost index," Industry report, 2025.

[60] JLL Research, "Global data center outlook," Industry report, 2025.

[61] SpaceX, "AI1 compute satellite," Product reveal video, 2026, 150 kW peak / 120 kW avg compute payload; 70 kW/tonne; 110 m^2 deployable liquid radiator; 150 kW solar array at 250 W/m^2.

[62] E. Mahr, A. Tu, and A. Gupta, "Development of the small satellite cost model 2019 (sscm19)," in *2020 IEEE Aerospace Conference*, 2020, pp. 1–12.

[63] SpaceX, "Falcon payload user's guide," Technical document, 2025, falcon 9 Block 5: 22,800 kg to LEO (28.5°); $\sim$17% payload reduction for polar/SSO orbits.

[64] ——, "Starship users guide," Technical document, 2025, 100+ tonnes to LEO up to 500 km circular, up to 98.9° inclination (mature V3 vehicle).

[65] ——, "SpaceX rideshare program — pricing," https://www.spacex.com/rideshare/, 2026, transporter: from $350k for 50 kg payload ($\approx$$7,000/kg); also reported in NASA SoA-2026 p.286.

[66] B. D. Yost, B. C. Kempa, C. D. Burkhard, L. Santos Soto, D. E. Steinfeld, S. M. Pedrotty, G. F. Benavides, and S. V. Weston, "Small spacecraft technology state-of-the-art report," NASA Ames Research Center, Small Spacecraft Systems Virtual Institute (S3VI), Tech. Rep. NASA/TP–20260003140, May 2026, annual S3VI survey; chapters cited individually by page in our calibration note. [Online]. Available: https://www.nasa.gov/smallsat-institute/sst-soa/

[67] B. Spence, S. White, N. Wilder, T. Gregory, M. Douglas, R. Takeda, N. Mardesich, T. Peterson, B. Hillard, P. Sharps, and N. Fatemi, "Next generation ultraflex solar array for nasa's new millennium program space technology 8," in *2005 IEEE Aerospace Conference*, 2005, pp. 824–836.

[68] NASA Office of the Chief Technologist, "NASA technology roadmap: TA 14—thermal management systems," National Aeronautics and Space Administration, Tech. Rep., 2015, target $\leq$3 kg/m^2 for advanced radiator panels.

[69] Northrop Grumman SpaceLogistics, "Mission extension vehicle (MEV)," Fact sheet, 2021, mEV-2 docked with Intelsat IS-10-02 in April 2021; provides station-keeping and attitude control for 5+ years.

[70] S. Go, J. Park, S. More, H. Wu, I. Wang, A. Jezghani, T. Krishna, and D. Mahajan, "Characterizing the efficiency of distributed training: A power, performance, and thermal perspective," in *Proceedings of the 58th IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 626–642. [Online]. Available: https://doi.org/10.1145/3725843.3756111

[71] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, p. 63–74, Aug. 2008. [Online]. Available: https://doi.org/10.1145/1402946.1402967

[72] G. P. Ginet, T. P. O'Brien, S. L. Huston, W. R. Johnston, T. B. Guild, R. Friedel, C. D. Lindstrom, C. J. Roth, P. Whelan, R. A. Quinn, D. Madden, S. Morley, and Y.-J. Su, "AE9, AP9 and SPM: New models for specifying the trapped energetic particle and space plasma environment," *Space Science Reviews*, vol. 179, pp. 579–615, 2013, trapped-proton flux model behind the 650 km SSO dose/rate estimates.

[73] J. T. Daly, "A higher order estimate of the optimum checkpoint interval for restart dumps," *Future Generation Computer Systems*, vol. 22, no. 3, pp. 303–312, 2006.

[74] Z. Wang, Z. Jia, S. Zheng, Z. Zhang, X. Fu, T. S. E. Ng, and Y. Wang, "Gemini: Fast failure recovery in distributed training with in-memory checkpoints," in *Proceedings of the 29th Symposium on Operating Systems Principles*, ser. SOSP '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 364–381. [Online]. Available: https://doi.org/10.1145/3600006.3613145

[75] H. Zhang, J. Wang, Z. Yu, Y. Zhang, X. Ji, K. Mao, J. Zhang, Y. Zhang, T. Wu, F. Jie, X. Huang, Z. Cai, J. Cheng, S. Wang, W. Li, X. Bao, H. Xu, S. Zhao, J. Li, H. Sun, Z. Zhang,

Y. Xiong, and C. Li, "FlashRecovery: Fast and low-cost recovery from failures for large-scale training of LLMs," 2025. [Online]. Available: https://arxiv.org/abs/2509.03047
[76] A. Grattafiori, A. Dubey, A. Jauhri *et al.*, "The Llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783
[77] A. Altenbernd, P. Wiesner, and O. Kao, "Exploring silent data corruption as a reliability challenge in LLM training," in *26th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, 2026, bit-flip injection during LLaMA (60M–1.3B) training: only FP exponent-bit flips matter; converged eval loss vs. injection frequency (baseline at 1/1000, +11% at 1/10, divergence at every step); backward pass most damaging. arXiv:2604.00726.
[78] A. Tyagi, S. Hukerikar, N. Saxena, Y. Huang, R. Shirvani, T. Tung, and Y. Zhu, "LLM-PRISM: Characterizing silent data corruption from permanent GPU faults in LLM training," arXiv:2604.10390, 2026, 7,664 GPT-2 training runs across FP16/BF16/FP8; outcome distribution (unchanged/changed/crashed); degradation onset at fault rate 0.05.
[79] J. Ma, H. Pei, L. Lausen, and G. Karypis, "Understanding silent data corruption in LLM training," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025, real SDC-afflicted GPUs during LLM pre-training: loss curves look normal while the model drifts to a different minimum; worst case 0.00 downstream accuracy. arXiv:2502.12340.
[80] NVIDIA, "NVIDIA A100 GPU memory error management," https://docs.nvidia.com/deploy/a100-gpu-mem-error-mgmt/, 2023, row-remapping policy: 8 spare rows per bank, hard cap 512 remaps per GPU; generation-invariant.
[81] NVIDIA Corporation, "NVIDIA H100 Tensor Core GPU," https://www.nvidia.com/en-us/data-center/h100/, 2024, h100 specifications, including 80 GB HBM and PCIe Gen5.
[82] Y. Zhong, S. Liu, J. Chen, J. Hu, Y. Zhu, X. Liu, X. Jin, and H. Zhang, "DistServe: Disaggregating prefill and decoding for goodput-optimized large language model serving," in *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2024.
[83] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with PagedAttention," in *29th ACM Symposium on Operating Systems Principles (SOSP)*, 2023.
[84] R. Qin, Z. Li, W. He, J. Cui, F. Ren, M. Zhang, Y. Wu, W. Zheng, and X. Xu, "Mooncake: Trading more storage for less computation — a KVCache-centric architecture for serving LLM chatbot," in *23rd USENIX Conference on File and Storage Technologies (FAST)*, 2025.
[85] G. Giuffrida, L. Fanucci, G. Meoni, M. Batič, L. Buckley, A. Dunne, C. van Dijk, M. Esposito, J. Hefele, N. Vercruyssen, G. Furano, M. Pastena, and J. 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*, vol. 60, pp. 1–14, 2022.
[86] G. Mateo-Garcia, J. Veitch-Michaelis, C. Purcell, N. Longepe, S. Reid, A. Anlind, F. Bruhn, J. Parr, and P. P. Mathieu, "In-orbit demonstration of a re-trainable machine learning payload for processing optical imagery," *Scientific Reports*, 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-34436-w
[87] S. Yao, Y. Lin, M. Wang, K. Xu, M. Xu, C. Xu, and H. Zhang, "Leoedge: A satellite-ground cooperation platform for the ai inference in large leo constellation," *IEEE Journal on Selected Areas in Communications*, 2025. [Online]. Available: https://doi.org/10.1109/jsac.2024.3460083

APPENDIX
ARTIFACT APPENDIX

A. Abstract

This artifact contains the analytical and TCO models, experiment drivers, simulator sources, workloads and validation code for the paper's principal results. Its CPU-only Docker workflow regenerates the analyses, reruns the ASTRA-Sim and STAGE experiments and validates the outputs.

B. Artifact check-list (meta-information)

- **Algorithm:** Analytical mass, TCO, reliability and communication models; ASTRA-Sim/STAGE simulation.
- **Program:** Python/Bash drivers; patched ASTRA-Sim/STAGE.
- **Compilation:** Dockerized GCC, CMake, Make and Ninja.
- **Model:** Llama-3-70B, synthetic 700B and DeepSeek-V3.
- **Data set:** Versioned inputs, compact CSVs and trace generators.
- **Run-time environment:** Docker; x86-64 Ubuntu 24.04 container.
- **Hardware:** CPU; 16 GB RAM recommended.
- **Run-time state:** Deterministic software simulation.
- **Execution:** One host command.
- **Metrics:** TCO, mass, latency, utilization and iteration time.
- **Output:** Validated CSVs, PDF figures, logs and pass/fail report.
- **Experiments:** Full workflow; `smoke` and `core` modes provide shorter checks.
- **Disk space required:** 8 GB recommended.
- **Preparation time:** Approximately 5 minutes to build the Docker image on the reference host (network-dependent).
- **Experiment time:** With a prebuilt image, `smoke`/`core`/`full` took approximately 5/16/25 minutes on a Ryzen 9 9950X.
- **Publicly available:** Zenodo.
- **Code licenses:** MIT; upstream licenses for vendored sources.
- **Data licenses:** MIT License.
- **Workflow automation framework used:** Docker and Bash.
- **Archived:** DOI: https://doi.org/10.5281/zenodo.22698121.

C. Description

1) How to access: Download and unpack the Zenodo archive. The included evaluator README provides instructions; checksummed simulator sources are under `repro/vendor/`.

2) Hardware dependencies: Evaluation requires an x86-64 Linux host with Docker; we recommend 16 GB RAM and 8 GB free disk space.

3) Software dependencies: Docker builds the vendored simulators with pinned Ubuntu and Python packages. Experiments then run with `--network none`.

4) Data sets: The artifact includes model parameters, simulator configurations, targets and compact CSVs; the workflow generates Chakra traces at run time.

5) Models: Workload generators encode Llama-3-70B, a synthetic 700B Llama-class workload and DeepSeek-V3 as communication traces.

D. Installation

Run the following command from the repository root.

```
bash artifact/run.sh full
```

This builds the image, runs the recommended `full` workflow and writes a validated result bundle under `artifact-output/`.

E. Experiment workflow

Choose a mode for `artifact/run.sh`.

Mode	Coverage
`smoke`	Regenerate analytical/TCO outputs, redraw saved collective results and rerun the formation motivator.
`core`	`smoke` plus the DP/PP placement experiments supporting the formation/network co-design claim.
`full`	`core` plus a fresh ASTRA-Sim rerun of all points in Figures 8–11; recommended for evaluation.

F. Evaluation and expected results

Run `bash artifact/run.sh full`. A successful run returns zero and records `status: PASS` in the result bundle's `validation_report.txt`. The validator checks file presence, result structure and numerical tolerances against the expectations below.

Validated result	Expected
Moderate TCO minimum	512 GPUs/satellite
TCO premium vs. terrestrial (O/M/P)	$1.12/3.34/12.07\times$
DP-/PP-over-ISL vs. Clos	$\sim 78/\sim 1.17\times$
Cylinder/string iteration time	52.9/137.1 s
Training/swap/inference at 17.5 GB	5.3/7.2/87.5%
Effective compute/$ penalty	~ 4.8–$\sim 2{,}700\times$

The regenerated result figures should match the corresponding figures in the paper. The included JSON manifest defines all targets and tolerances.

G. Experiment customization

`AE_OUTPUT_ROOT` and `AE_IMAGE` select the output directory and image; scripts under `repro/` expose experiment parameters.

H. Notes

Evaluation covers the listed quantitative results without rebuilding the manuscript.

I. Methodology

Submission, reviewing and badging follow the ACM artifact policy and unified AE methodology.

- ACM artifact policy
- Artifact Evaluation methodology