

Hot AI in Cold Space: Thermal-Crosstalk-Aware Scheduling for Sustainable Orbital AI Clusters

SHUYI CHEN, Southern University of Science and Technology, China

ZHENGCHANG HUA, Southern University of Science and Technology, China

NIKOS TZIRITAS, University of Thessaly, Greece

GEORGIOS THEODOROPOULOS, Southern University of Science and Technology, China

Terrestrial AI training faces an unsustainable energy and water crisis, positioning Orbital Data Centers (ODCs) as a "zero operational carbon" alternative. However, the sub- $10\mu\text{s}$ communication latency required for synchronized scientific workloads, such as distributed Large Language Model (LLM) training, forces ODCs into extreme physical density, triggering a critical "Proximity-Thermal Paradox." As these high-density systems scale into Monolithic Structures or Proximity Swarms, they suffer from intense thermal-fluid crosstalk (heat traps in shared cooling loops) and thermal-radiative crosstalk (mutual heating that blocks deep-space cooling radiators). If left unmitigated, this persistent heat stagnation not only triggers severe thermal throttling that degrades training throughput, but also induces severe thermal fatigue, drastically shortening hardware lifespans and generating premature space e-waste. To make orbital AI truly sustainable, this position paper challenges traditional uniform load-sharing. We propose the Thermal-Aware Heterogeneity Thesis, which treats spatial cooling variances as a primary resource management dimension. Building on this, we introduce Thermal-Load Balancing (TLB), a software framework that dynamically migrates these intensive workloads to the coolest available units based on instantaneous fluid temperatures or absorbed radiation. Our analysis demonstrates that TLB resolves thermal bottlenecks to restore Model Flops Utilization (MFU), while simultaneously reducing physical thermal stress. Extending the operational lifespan of orbital hardware is crucial to amortize the massive embodied carbon of rocket launches, outlining a necessary pathway to scale orbital AI without accelerating e-waste.

CCS Concepts: • **Computing methodologies** → **Distributed computing methodologies**; *Artificial intelligence*; • **Hardware** → **Thermal issues**; *Impact on the environment*.

Additional Key Words and Phrases: orbital data centers, sustainable computing, embodied carbon, thermal-aware scheduling, large language models, satellite swarms

1 Introduction

Terrestrial AI training faces an unsustainable energy and water crisis, making Orbital Data Centers (ODCs) an attractive "zero operational carbon" alternative [5, 21]. However, to truly offset the massive *embodied carbon* of rocket launches [23], ODCs must maximize their lifetime Model Flops Utilization (MFU) and avoid premature hardware failures. This presents a severe challenge for latency-sensitive, synchronized scientific computations. As a primary example, distributed Large Language Model (LLM) training, particularly when utilizing Tensor Parallelism, demands sub- $10\mu\text{s}$ communication latencies [28]. This forces computing nodes into extreme high-density

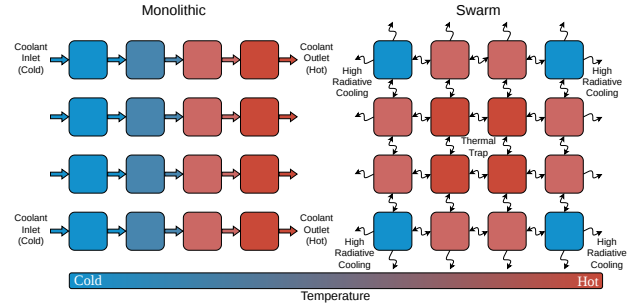


Fig. 1. Two ODC architectural paradigms: Monolithic Structures with centralized cooling loops and Proximity Swarms of free-flying nodes.

configurations (e.g., 3-meter physical proximity), triggering a critical *Proximity-Thermal Paradox*: the spatial compactness required for low-latency synchronization inevitably leads to intense thermal crosstalk.

This crosstalk is particularly destructive because synchronous distributed workloads rely on "All-or-Nothing" synchronization patterns (e.g., All-Reduce operations in LLM training). Under these strict constraints, a single thermally congested "straggler" node stalls the entire gigawatt-scale cluster [10, 15]. If unmitigated, this persistent heat stagnation not only degrades training throughput but also induces severe thermal fatigue, which drastically shortens hardware lifespans and generates premature "space e-waste."

As high-density ODCs scale, they fall into two distinct architectural paradigms, each presenting multi-scale thermal interferences. *Monolithic Structures* are massive self-assembled units that rely on centralized fluid loops, suffering from *Thermal-Fluid Crosstalk* as heat from high-density GPU tiles accumulates along shared cooling paths and creates downstream heat traps [3]. Alternatively, *Proximity Swarms* consist of free-flying nodes maintained in tight formations [33]. These swarms suffer from *Thermal-Radiative Crosstalk*, where geometric shadowing and mutual heating block deep-space cooling windows, particularly for nodes trapped in the cluster's core [20].

In both paradigms, traditional load-balancing frameworks force structural and operational uniformity, which inevitably leads to synchronous bottlenecks. To make orbital AI physically and environmentally sustainable, we introduce the *Thermal-Aware Heterogeneity Thesis*. We argue that ODCs must reject uniform load-sharing and instead treat spatial and temporal cooling variance as a primary, schedulable computational resource.

Authors' Contact Information: Shuyi Chen, Southern University of Science and Technology, Shenzhen, China, chensy8@mail.sustech.edu.cn; Zhengchang Hua, Southern University of Science and Technology, Shenzhen, China, huazc@mail.sustech.edu.cn; Nikos Tziritas, University of Thessaly, Lamia, Greece, nitzirit@uth.gr; Georgios Theodoropoulos, Southern University of Science and Technology, Shenzhen, China, theogeorgios@gmail.com.

Building on this, we propose the Thermal-Load Balancing (TLB) paradigm. TLB is a software orchestration framework that dynamically migrates computation blocks to the coolest available units based on instantaneous fluid temperatures or absorbed radiation. By proactively shifting tasks, TLB actively resolves heat traps, prevents the emergence of thermal stragglers, and mitigates cyclical thermal stress on the silicon.

This paper outlines a pathway to scale orbital AI without accelerating space e-waste. Specifically, we make the following contributions:

- We define the *Proximity-Thermal Paradox* in ODCs, detailing how extreme density induces severe Thermal-Fluid and Thermal-Radiative Crosstalk.
- We introduce the *Thermal-Aware Heterogeneity* thesis, establishing that exploiting cooling variance is essential to preventing thermal fatigue and straggler-induced MFU degradation.
- We propose the *Thermal-Load Balancing* framework, providing workload migration strategies to resolve heat traps in both monolithic flow paths and swarm radiative fields.
- We formulate an evaluation methodology, modeling a dense multi-node mesh, to demonstrate that TLB restores MFU and extends hardware lifespans by reducing cyclic thermal stress, thereby amortizing the embodied carbon of ODCs.

2 Related Work

The development of the TLB paradigm intersects several established research domains. To contextualize the novelty of our approach, this section reviews the state-of-the-art in space-based distributed computing, spacecraft thermal control mechanisms, and terrestrial AI load-balancing strategies.

2.1 Distributed Computing in Orbital Data Centers

The paradigm of satellite operations is rapidly shifting from simple data collection and downlink to on-orbit data processing. To support latency-sensitive and compute-intensive applications such as distributed machine learning and autonomous navigation, researchers have proposed the concept of ODCs [5, 21, 25]. Recent studies have explored various architectures for space-based edge computing, including hierarchical multi-orbit systems, micro datacenters using torus topologies, and satellite-terrestrial integrated networks [4, 29, 31]. Specific to application workloads, frameworks for on-orbit remote sensing and AI task offloading have been proposed to minimize delay and energy consumption [8, 9, 13]. By clustering multiple satellite nodes equipped with commercial-off-the-shelf accelerators, ODCs can circumvent the bandwidth limitations and high latency of terrestrial downlinks.

However, supporting synchronous workloads like distributed LLM training introduces a distinct networking challenge. Terrestrial AI clusters rely on high-speed backplanes or Optical Circuit Switches [26] to achieve the sub-10 μ s latencies required for synchronous operations like Tensor Parallelism. Replicating this in space via Free-Space Optical (FSO) links forces ODCs into extremely tight spatial topologies [12]. Current space-based computing literature primarily focuses on asynchronous tasks (e.g., remote sensing offloading) [8],

largely overlooking the severe thermal consequences of deploying these tightly coupled, synchronous networks.

2.2 Spacecraft Thermal Management

Thermal management in the vacuum of space relies predominantly on radiative heat transfer, supplemented by internal conductive and fluid loop systems, as well as phase change materials and passive thermal coatings [16–18, 30]. Traditional spacecraft thermal design assumes a relatively unobstructed view factor to deep space and isolated internal heat generation. Current literature primarily addresses thermal control for isolated satellites or loosely formed constellations. While terrestrial HPC studies have explored heat accumulation in pumped liquid loops [7], and aerospace studies have modeled radiative mutual heating in simple formations [24], existing space thermal frameworks do not account for the multi-scale, dynamic thermal congestion inherent to GW-scale monoliths or densely packed swarms. Hardware over-provisioning adds prohibitive mass, severely exacerbating the embodied carbon footprint of the mission [23], and mechanical orbital reconfiguration consumes limited propellant. Current space thermal control literature rarely addresses the rapid, cyclic thermal fatigue induced by dynamic computational workloads, which accelerates hardware degradation and generates space e-waste. A software-defined approach to thermal management is highly desirable but remains underexplored.

2.3 Conventional Load Balancing in Distributed AI

In terrestrial data centers, distributed scientific workloads such as AI training via data or pipeline parallelism rely on load-balancing frameworks that optimize for computational throughput, memory capacity, and network bandwidth [10, 19, 34]. Recent advancements in terrestrial networks also employ intelligent, machine learning-based dynamic load balancing to manage heterogeneous resources and reduce energy expenditure in complex ecosystems like cloud services and edge networks [2, 6, 22, 27, 32]. While a distinct body of literature addresses thermal-aware task placement in terrestrial cloud data centers to minimize cooling energy (i.e., operational carbon) [1, 14], these models are heavily anchored in convective air-cooling dynamics (e.g., hot-aisle containment) and active HVAC pump logic. They cannot be translated to ODCs, which rely purely on highly variable radiative heat rejection and closed-loop monolithic flow paths. Since ODCs are powered by solar arrays (effectively zero operational carbon), load balancing must pivot from optimizing electricity expenditure to minimizing thermal stress and maximizing hardware lifespan to amortize embodied carbon.

When applied to ODCs, conventional load balancers distribute tasks symmetrically or based purely on static hardware specifications, completely ignoring the dynamic and spatially heterogeneous thermal environment of the satellite cluster. As a result, nodes with poor cooling margins quickly hit hard temperature thresholds (T_{hard}), triggering aggressive hardware thermal throttling. Because synchronized scientific training requires strict barrier synchronization between nodes (e.g., via All-Reduce operations), the thermal throttling of a single central node cascades, delaying the entire

global step, a phenomenon often referred to as the straggler problem [10, 11]. This highlights the critical need for an orbital-specific thermal-aware scheduling paradigm.

While the vision of ODCs is rapidly maturing, the intersection of spatial topology, thermal dissipation, and distributed AI workload scheduling remains a critical blind spot. Existing literature addresses these challenges in isolation: thermal engineers focus on hardware radiators, network architects optimize for link latency, and AI researchers design load balancers for terrestrial facilities with homogeneous cooling. Our work bridges this gap by proposing that thermal heterogeneity is not just a hardware constraint, but a primary software-schedulable resource.

3 System model

3.1 Orbital Data Center Model

We model ODCs as high-density satellite clusters and consider two distinct architectural paradigms. Let $\mathcal{N} = \{1, \dots, N\}$ be the set of satellite nodes. Each node i is characterized by its thermal capacitance C_i , dynamic power limits (P_{idle} and peak P_{max}), and maximum processing speed S_{max} . Heat dissipation properties differ significantly depending on the architecture:

Monolithic Structures. The monolithic ODC features a rigid, large-scale (e.g., MW-to-GW scale) macro-structure shaped like a massive unified container, similar to conceptual tether-based orbital space stations (Starcloud-4) [3]. Power generation and heat dissipation for the entire ODC are managed globally by the primary structure via externally mounted large-area heat radiators and solar arrays. Standalone compute units are housed within this structure and managed by a centralized liquid cooling system.

As shown in the left part of Figure 1, multiple coolant pipes are routed through the data center, pumping coolant through the compute units before returning it to the heat exchanger. Along the pathway of each coolant loop, upstream nodes receive cold coolant, while downstream nodes receive pre-heated coolant, creating an inherent thermal imbalance across the compute units.

Proximity Swarms. The proximity swarm ODC architecture consists of multiple independent satellite computing nodes flying in a tight formation, similar to Google’s Project Suncatcher concept [33]. These nodes form a dense constellation with inter-node distances typically ranging from hundreds of meters to a few kilometers, communicating via high-bandwidth FSO links, as shown in the right part of Figure 1.

Each satellite in the swarm is equipped with its own solar panels and heat radiators. Although each node utilizes independent passive radiative cooling, the nodes within the constellation are physically close enough to thermally influence one another. Due to view factor occlusion caused by neighboring nodes, their radiators experience varying effective radiating areas. Warmer nodes may also radiate waste heat directly toward their neighbors. Together, these factors establish a dynamic thermal imbalance across the distributed swarm.

3.2 Heat Dissipation and Thermal Crosstalk Models

Let P_{in} and $T_{instant}$ denote the instantaneous chip power and temperature. We link computational load to heat generation non-linearly:

$$P_{in} = P_{idle} + (P_{max} - P_{idle}) \left(\frac{S_{instant}}{S_{max}} \right)^Y \quad (1)$$

The transient thermal state is dictated by the net heat flow $P_{in} - P_{out}$, where heat dissipation P_{out} is architecturally dependent:

Thermal-Fluid Crosstalk (Monolithic). Downstream nodes are cooled by fluid pre-heated by upstream components:

$$T_{fluid}^{(i)} = T_{fluid}^{(i-1)} + \eta P_{out}^{(i-1)} \quad (2)$$

This degrades the downstream convective heat dissipation rate, creating downstream heat traps:

$$P_{out}^{(i)} \propto T_{instant}^{(i)} - T_{fluid}^{(i)} \quad (3)$$

Thermal-Radiative Crosstalk (Swarms). Dense formations rely on deep-space radiation governed by the Stefan-Boltzmann law:

$$P_{out} \propto T_{instant}^4 - T_{amb}^4 \quad (4)$$

Mutual geometric shadowing severely restricts this, which we model via an Effective View Factor $\rho_i \in (0, 1]$. Peripheral nodes retain larger effective radiating areas (higher ρ_i), while core nodes suffer severe occlusion (significantly lower ρ_i), triggering immediate heat traps.

Note that our current formulation employs a time-averaged, steady-state deep space environment abstraction. We deliberately isolate the endogenous thermal crosstalk (heat generated strictly by the AI payload and throttled by topological proximity) from exogenous orbital thermodynamics. Highly transient space environment factors, including direct solar flux, Earth infrared emissions, albedo effects, orbital eclipse cycles, and complex 3D spacecraft attitude geometries, are thus abstracted out of this proof-of-concept simulation. This abstraction is an intentional, conservative baseline. Exogenous variables introduce severe high-frequency thermal transients that will inevitably exacerbate the thermal variance across the cluster. They do not negate the fundamental mechanics of the Proximity-Thermal Paradox; rather, they amplify the necessity of software-level thermal orchestration. Integrating high-fidelity, dynamic orbital mechanics and 3D ray-tracing into the thermal-compute co-simulator represents the immediate trajectory for future work.

3.3 Throttling and Lifespan Degradation Models

To avoid hardware damage, nodes enforce a soft temperature threshold (T_{soft}) and a hard threshold (T_{hard}). When $T_{instant}$ exceeds T_{soft} , the processing speed $S_{instant}$ scales down linearly from S_{max} , capping at a minimum safe speed S_{min} once T_{hard} is reached.

We also model thermal-induced hardware degradation using the standard Arrhenius equation, where the Mean Time To Failure (MTTF) is exponentially dependent on temperature:

$$MTTF \propto \exp\left(\frac{E_a}{k_B T}\right) \quad (5)$$

Here, E_a is the activation energy (e.g., 0.685 eV) and k_B is the Boltzmann constant. Because of this exponential relationship, even minor reductions in peak operating temperatures yield disproportionately large extensions in hardware lifespan.

3.4 Workload Model with Asymmetric Data Parallelism

To accelerate training when memory is sufficient, we apply Data Parallelism. Assuming every computing node holds a full replica of the model, the training data batch is split across different nodes. Each node processes a certain volume of data, determined by the configured micro-batch size.

For distributed workloads executing via Data Parallelism, the global batch (size B_{global}) is partitioned such that each node processes a micro-batch of b_i tokens. The computation time is directly proportional to b_i and inversely proportional to the thermally throttled speed $S_i(T)$.

A global training step strictly requires gradient synchronization (e.g., All-Reduce). This means the total step latency is dictated entirely by the single slowest node in the cluster:

$$t_{step} = \max_i (t_{comp}(i) + t_{comm}) \quad (6)$$

This mathematical “max” operation establishes the severe consequence of the Proximity-Thermal Paradox, where a single node suffering from intense thermal crosstalk will experience degraded $S_i(T)$, thereby stalling the entire ODC cluster and causing a global collapse in MFU.

While the physical density of ODCs is driven by the strict latency bounds of Tensor Parallelism, simulating microsecond-level optical routing alongside macroscopic thermodynamics is computationally intractable. Therefore, we utilize Data Parallelism as a macro-scale surrogate model; since both rely on strict barrier synchronization, the straggler penalty induced by thermal throttling is structurally isomorphic.

4 The Thermal-Load Balancing Framework

4.1 A Generalized Orchestration Architecture

TLB is designed as a generalized, closed-loop orchestration framework, decoupling the load-balancing *mechanism* from any specific scheduling *policy*. It continuously operates through a control loop of thermal telemetry, capability profiling, and asymmetric workload slicing. TLB aggregates real-time physical metrics like instantaneous silicon temperatures, local fluid pre-heating levels, and dynamic view factor occlusions. These raw thermal data are translated into an abstract capability score (w_i) that quantifies each node’s real-time thermal margin. Finally, TLB dynamically adjusts the distribution of the AI workload (e.g., micro-batch sizes in Data Parallelism) to match the profiled capabilities.

This abstraction allows ODC operators to plug in arbitrarily complex decision engines, ranging from convex optimization solvers to Deep Reinforcement Learning agents, without altering the underlying distributed AI communication backends.

4.2 Proof-of-Concept: Proportional Heuristic

For this position paper, rather than deploying a computationally expensive global optimization solver, we implement a lightweight proportional heuristic. Standard Data Parallel frameworks enforce a uniform batch distribution, allocating $B_{global}/|\mathcal{N}|$ tokens to each node. This severely penalizes the entire cluster when nodes with degraded cooling drop their processing speed $S_i(T)$. TLB breaks

this uniformity by calculating the thermal capability score w_i . The workload b_i assigned to node i is strictly proportional to its score:

$$b_i = 1 + \left\lceil (B_{global} - |\mathcal{N}|) \cdot \frac{w_i}{\sum_{k \in \mathcal{N}} w_k} \right\rceil \quad (7)$$

where every node receives at least one baseline unit of work to maintain gradient synchronization participation, and any fractional rounding remainders are allocated to the highest-scoring nodes.

We define an aggressiveness multiplier α and compute the capability scores based on the architectural paradigm:

Monolithic Flow-Path Priority. We prioritize upstream nodes that receive the coldest fluid. Let L_{pipe} be the total number of nodes on a cooling pipe, and idx_i be the zero-based index of node i along that pipe. The capability score is:

$$w_i = 1.0 + \alpha \cdot (L_{pipe} - idx_i) \quad (8)$$

Swarm Radiative Priority. We prioritize nodes on the periphery of the cluster with an unobstructed view of deep space. By directly utilizing the Effective View Factor (ρ_i) introduced in our thermal model, which inherently accounts for geometric shadowing, the capability score is:

$$w_i = 1.0 + \alpha \cdot \rho_i \quad (9)$$

By proactively shifting the heaviest computational burdens to nodes with high w_i , TLB inherently limits the heat generation of central or downstream nodes, preventing them from hitting T_{soft} and triggering the straggler effect.

We explicitly state that this proportional heuristic is a *baseline feasibility demonstration*, not a terminal algorithmic solution. While the heuristics defined in Equations 7, 8, and 9 do not provide strict mathematical guarantees for boundary safety under highly non-linear thermal transients, they serve as a necessary diagnostic probe. Their purpose is to empirically validate the solvability of the Proximity-Thermal Paradox via asymmetric workload slicing, without obscuring the fundamental system dynamics behind a black-box solver.

4.3 Integration with the Distributed AI Stack

To transition TLB from a theoretical model to a deployable system, it must bridge the gap between hardware telemetry and the AI application layer. At the hardware level, TLB leverages standard space-grade baseboard management controllers to poll instantaneous silicon temperatures and coolant states. At the application layer, TLB interfaces with distributed AI frameworks (e.g., PyTorch DistributedDataParallel or Megatron-LM). Instead of relying on a static DataLoader that evenly shards the dataset, TLB implements a *Thermal-Aware Data Sampler*. When the profiling engine updates the capability scores w_i , the sampler dynamically adjusts the micro-batch size b_i for the upcoming training steps.

Dynamic workload slicing introduces overheads from altering micro-batch sizes, such as memory pre-allocation and graph recompilation penalties. While hardware throttling might appear as a zero-overhead alternative terrestrially, it induces severe cyclical thermal fatigue in ODCs. This thermal fatigue drastically shortens hardware lifespan, undercutting the launch’s massive embodied carbon investment.

To ensure that the overhead of dynamic batching does not cannibalize these critical lifespan gains, TLB relies on emerging dynamic-shape compilers (e.g., PyTorch 2.0 `torch.compile`) or pre-compiled computational graph buckets corresponding to a discrete set of allowable batch sizes. This guarantees that thermal-induced workload shifts are executed with near-zero software overhead, making asymmetric scheduling strictly beneficial.

Executing the TLB control loop at every training step would introduce unacceptable synchronization delays. Therefore, TLB employs a hybrid dual-loop policy. First, workload redistribution occurs at coarse-grained intervals (e.g., epoch boundaries) to handle slow-moving thermal dynamics like orbital eclipses. Second, because the static proportional heuristic lacks mathematical guarantees to prevent over-allocation in unpredictable thermal states, TLB deploys a rigid, event-driven safety net. This net operates asynchronously from the training loop. If hardware telemetry detects a node rapidly approaching T_{soft} , the safety net forcibly triggers an immediate asynchronous task evacuation, acting as an absolute physical backstop. This completely prevents catastrophic threshold violations, ensuring that the absence of algorithmic optimality never compromises physical hardware safety.

5 Performance Evaluation

To validate the Thermal-Aware Heterogeneity thesis, we implement a proof-of-concept simulation of the TLB framework. Since this paper focuses on the necessity of a paradigm shift rather than proposing an optimal algorithmic solver, we benchmark a proportional variation of TLB against a standard homogeneous load balancer.

5.1 Simulation Methodology

To model the multi-scale thermal dynamics of ODCs, we developed a time-stepped thermal-compute co-simulator¹ that evaluates distributed scientific workloads executing via Data Parallelism.

For the **Monolithic** architecture, 64 nodes are distributed across 8 liquid cooling pipes (8 nodes daisy-chained per pipe). Downstream nodes process coolant that has been pre-heated by upstream nodes. For the **Proximity Swarm** architecture, 36 independent satellites fly in a dense 6×6 planar grid formation. We implement a radiative view factor model where nodes dynamically cast thermal shadows on each other; edge nodes retain larger effective radiating areas (higher ρ_i), while the core nodes suffer severe radiative occlusion (significantly lower ρ_i).

We compare two scheduling paradigms: **Baseline (Uniform)** shards the global micro-batch evenly across all nodes, mimicking standard PyTorch DDP behavior; **TLB (Proportional)** implements the proportional allocation heuristic. We track instantaneous hardware temperatures, dynamic processing speeds (throttling), and global step synchronization latency.

5.2 Mitigating Heat Traps (Spatial Analysis)

The spatial thermal distribution reveals the fundamental flaw of homogeneous scheduling in ODCs. Figure 2 shows the average node chip temperatures in the Monolithic and Proximity Swarm architectures. Under the Baseline policy, the *Proximity-Thermal*

¹<https://github.com/Sukiichan/sdc-sim>

Average Node Temperatures Over the Iteration

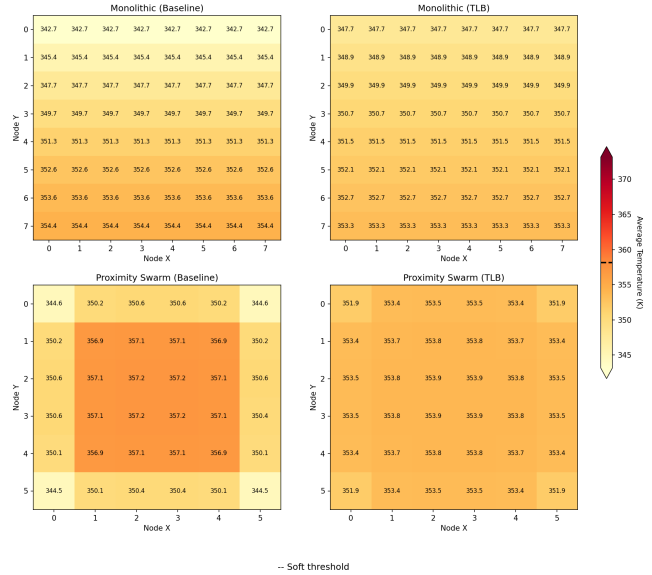


Fig. 2. Average Node Temperatures Over the Iteration (Baseline vs. TLB)

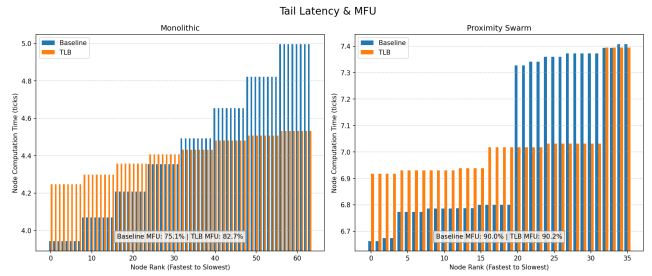


Fig. 3. Tail Latency & MFU of node computation times (Baseline vs. TLB).

Paradox manifests rapidly. In the Monolithic configuration, heat progressively accumulates along the cooling loops, pushing the downstream nodes (Row 7) into thermal saturation (up to 354.4 K / 81.3°C). In the Swarm configuration, the core nodes become severe heat traps reaching 357.2 K (84.1°C) due to high radiative view factor occlusion, while the periphery satellites remain significantly cooler (344.5 K / 71.4°C).

TLB dynamically maps the thermal topology and deliberately shifts the heaviest computational burdens to the upstream Monolithic nodes and the periphery Swarm nodes. This asymmetric workload slicing neutralizes the cooling variance. For the Monolithic setup, the maximum temperature drops to 353.3 K (80.2°C). For the Swarm, the extreme thermal gradient is flattened, with core temperatures reducing to 353.9 K (80.8°C) and edge temperatures rising to 351.9 K (78.8°C), effectively eliminating spatial heat traps.

5.3 Prioritizing Hardware Survivability Over Throughput

The resolution of spatial heat traps shifts the Return on Investment metric for ODCs. Terrestrial data center optimizations traditionally prioritize instantaneous computational throughput (e.g., MFU). However, in the zero-operational-carbon environment of solar-powered orbital swarms, energy is abundant, but hardware lifespan is strictly finite. Thus, a primary metric of success is *hardware survivability* to amortize the massive embodied carbon of the rocket launch.

Extending Lifespan to Amortize Embodied Carbon. Under the uniform Baseline policy, central nodes experience severe thermal fatigue due to relentless temperature oscillation and sustained peak temperatures (e.g., 357.2 K). This cyclical thermal stress rapidly degrades silicon dies and solder joints, accelerating premature space e-waste. By smoothing out thermal gradients and capping peak temperatures at safe operational limits, TLB prevents these nodes from entering cyclical throttling regimes.

Using an apparent activation energy (E_a) of 0.685 eV, our Arrhenius-based degradation analysis demonstrates that TLB directly extends the MTTF of the most thermally constrained nodes. We observe a 6.15% lifespan increase for the core node in the Proximity Swarm and a 1.71% increase for the downstream outlet node in the Monolithic cluster. This tangible extension of physical durability is the primary mathematical pathway to achieving sustainable orbital AI.

Throughput Stabilization as a Secondary Byproduct. Evaluated strictly through a terrestrial performance lens, the MFU gains, particularly the marginal 0.2% increase in the Swarm configuration (from 90.0% to 90.2%), may appear susceptible to system overhead. However, this marginality validates our thesis: the Swarm does not fail due to a lack of aggregate computational capacity, but rather due to structural latency variance caused by localized thermal throttling.

Figure 3 demonstrates this straggler effect. While the Baseline Monolithic architecture exhibits severe step-like latency variance capping the MFU at 75.1% (which TLB decisively improves to 82.7%), the absolute MFU of the Swarm was inherently high. TLB's true contribution in the Swarm is perfectly equalizing the computation time across all 36 nodes, indicated by the entirely flat TLB latency distribution. The 0.2% MFU gain is merely a secondary mathematical byproduct of eliminating synchronization wait times. The primary objective of averting hardware destruction via thermal load balancing has been successfully achieved.

6 Discussion and Conclusion

Terrestrial AI's voracious appetite for energy and water is driving the industry toward ODCs. However, this transition shifts the root cause of unsustainability: while operations become strictly zero-carbon, the environmental burden is entirely front-loaded into the massive *embodied carbon* of rocket launches and the impending accumulation of space e-waste. This paper highlights that the sub-10 μ s latency demands of synchronized workloads force ODCs into extreme physical density, triggering a Proximity-Thermal Paradox that causes premature hardware failure and collapses MFU.

By introducing the Thermal-Aware Heterogeneity Thesis and the TLB framework, we demonstrated that resolving spatial heat traps through asymmetric software scheduling can restore synchronous

throughput and mitigate cyclic thermal fatigue. Extending hardware lifespans via software orchestration is not merely a performance optimization; it is the indispensable prerequisite for amortizing the embodied carbon of space-based AI. As ODCs evolve into generalized orbital supercomputers, Thermal-Aware Heterogeneity becomes a foundational concept for any distributed scientific workload constrained by barrier synchronization, extending the relevance of TLB beyond generative AI.

To fully realize sustainable orbital computing, the sustainable computing community must confront several critical open challenges:

Dynamic Network-Thermal Co-design. Terrestrial AI clusters rely on dynamic network reconfiguration (e.g., Optical Circuit Switches [26]) to optimize all-to-all communication. In space, reconfiguring FSO links within a proximity swarm requires mechanical or optical steering. This directly alters a node's physical orientation, instantly changing its radiative view factors and thermal state. Co-optimizing dynamic network topologies with instantaneous orbital thermodynamics remains an entirely unsolved challenge.

The E-Waste vs. Carbon Trade-off. While software schedulers like TLB can extend hardware survival, orbital chips currently cannot be repaired. The community must conduct rigorous lifecycle assessments to answer a provocative question: does the "zero operational carbon" benefit of ODCs mathematically justify a paradigm of "disposable" high-end AI accelerators? Addressing this may eventually necessitate designing orbital hardware explicitly for in-space recycling.

Federated Thermal Telemetry for Heterogeneous Swarms. Future GW-scale ODCs will likely be multi-tenant constellations composed of hardware from diverse vendors. Executing cluster-wide thermal scheduling requires nodes to share their physical statuses. Developing open standards to broadcast "thermal margins" without exposing proprietary silicon thermal layouts or payload capabilities is essential to building cooperative, sustainable orbital ecosystems.

Acknowledgements

This work was supported by the Research Institute of Trustworthy Autonomous Systems (RITAS), Shenzhen Science and Technology Program (No. GJHZ20210705141807022), and Guangdong Province Innovative and Entrepreneurial Team Programme (No. 2017ZT07X386). Georgios Theodoropoulos is the corresponding author.

References

- [1] Rui Chen, Bo Liu, WeiWei Lin, JianPeng Lin, HuiWen Cheng, and KeQin Li. 2023. Power and thermal-aware virtual machine scheduling optimization in cloud data center. *Future Generation Computer Systems* 145 (2023), 578–589.
- [2] Gowtham Reddy Enjam. 2022. Energy-Efficient Load Balancing in Distributed Insurance Systems Using AI-Optimized Switching Techniques. *International Journal of Artificial Intelligence, Data Science, and Machine Learning* 3, 4 (2022), 68–76.
- [3] Ezra Feilden, Adi Oltean, and Philip Johnston. 2024. Why we should train AI in space. *Lumen Orbit Inc* (2024).
- [4] Ran Ginosar and David Steenari. 2025. Beyond Traditional Payload Data Handling: Micro-Datacenter in Space for Converged Software-Defined Storage and Payload Processing. In *2025 European Data Handling & Data Processing Conference (EDHPC)*. 1–7.
- [5] Carlos Guimarães, Alessio Netti, Markus Sauer, Florian Zeiger, Hans-Peter Huth, and Elizaveta Boriskova. 2026. A Survey on Satellite Computing: Connecting

- the Dots Between Networks and Applications. *IEEE Communications Surveys & Tutorials* 28 (2026), 567–592. doi:10.1109/COMST.2025.3579525
- [6] Emre Gures, Ibraheem Shayea, Mustafa Ergen, Marwan Hadri Azmi, and Ayman A El-Saleh. 2022. Machine learning-based load balancing algorithms in future heterogeneous networks: A survey. *IEEE Access* 10 (2022), 37689–37717.
- [7] Mohamad Hnainy, Ali Chehade, Henryk Klab, Hadrien Bauduin, Guillaume Polidori, and Chadi Maalouf. 2022. Performance analysis of new liquid cooling topology and its impact on data centres. *Applied Thermal Engineering* 213 (2022), 118733.
- [8] Yifei Hu and Wenbin Gong. 2023. An On-Orbit Task-Offloading Strategy Based on Satellite Edge Computing. *Sensors* 23, 9 (2023). doi:10.3390/s23094271
- [9] Qiangqiang Jiang, Lujie Zheng, Yu Zhou, Hao Liu, Qinglei Kong, Yamin Zhang, and Bo Chen. 2025. Efficient On-Orbit Remote Sensing Imagery Processing via Satellite Edge Computing Resource Scheduling Optimization. *IEEE Transactions on Geoscience and Remote Sensing* 63 (2025), 1–19. doi:10.1109/TGRS.2025.3528015
- [10] Mingwei Li, Jilin Zhang, Jian Wan, Yongjian Ren, Li Zhou, Baofu Wu, Rui Yang, and Jue Wang. 2020. Distributed machine learning load balancing strategy in cloud computing services. *Wireless Networks* 26, 8 (2020), 5517–5533.
- [11] Shigang Li, Tal Ben-Nun, Salvatore Di Girolamo, Dan Alistarh, and Torsten Hoefler. 2020. Taming unbalanced training workloads in deep learning with partial collective operations. In *Proceedings of the 25th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*. 45–61.
- [12] Yingzhi Li, Baisong Chen, Haolun Du, Ziming Wang, Heming Hu, Xuetong Li, Huan Qu, Jie Li, Weipeng Wang, Min Tao, et al. 2026. Integrated Optical Wireless Communication Featured With Optical Phased Array Transceivers for Full-Duplex and NonLine-of-Sight Transmission. *Laser & Photonics Reviews* 20, 7 (2026), e00822.
- [13] Yuejin Li, Mi Wang, Kai Hwang, Zhengdao Li, and Tongkai Ji. 2023. LEO Satellite Constellation for Global-Scale Remote Sensing With On-Orbit Cloud AI Computing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16 (2023), 9369–9381. doi:10.1109/JSTARS.2023.3316298
- [14] Jianpeng Lin, Weiwei Lin, Wentai Wu, Wenjun Lin, and Keqin Li. 2024. Energy-aware virtual machine placement based on a holistic thermal model for cloud data centers. *Future Generation Computer Systems* 161 (2024), 302–314.
- [15] Rui Lu and Dan Wang. 2025. A Thermal-Aware Workload Scheduler for High-Performance LLM Inference in Cooling-Regulated Datacenters. *SIGENERGY Energy Inform. Rev.* 5, 2 (Aug. 2025), 98–104. doi:10.1145/3757892.3757906
- [16] Yi-Gao Lv, Yao-Ting Wang, Tong Meng, Qiu-Wang Wang, and Wen-Xiao Chu. 2024. Review on thermal management technologies for electronics in spacecraft environment. *Energy Storage and Saving* 3, 3 (2024), 153–189. doi:10.1016/j.ens.2024.03.001
- [17] Erdinç Mermer and Rahmi Ünal. 2023. Passive thermal control systems in spacecrafts. *Journal of the Brazilian Society of Mechanical Sciences and Engineering* 45, 3 (2023), 160.
- [18] Jianyin Miao, Qi Zhong, Qiwei Zhao, and Xin Zhao. 2021. *Spacecraft thermal control technologies*. Springer.
- [19] Sergio Moreno-Alvarez, Juan M Haut, Mercedes E Paoletti, Juan A Rico-Gallego, Juan C Diaz-Martin, and Javier Plaza. 2020. Training deep neural networks: a static load balancing approach: S. Moreno-Álvarez et al. *The Journal of Supercomputing* 76, 12 (2020), 9739–9754.
- [20] Yunus Murat, Hatice Mercan, Nedim Sözbir, and Ahmet Selim Dalkilic. 2025. Thermal design for a communications satellite payload module. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 239, 18 (2025), 7629–7646.
- [21] Shima Naser, Maryam Tariq, Raneem Abdel-Rahim, De Mi, Azzam Mourad, Hadi Otrok, Mahmoud Al-Qutayri, Ayman Elnashar, and Sami Muhaidat. 2026. From Connectivity to Multi-Orbit Intelligence: Space-Based Data Center Architectures for 6G and Beyond. arXiv:2603.18601 [cs.ET] <https://arxiv.org/abs/2603.18601>
- [22] Aravind Nuthalapati et al. 2024. Advanced techniques for distributing and timing artificial intelligence based heavy tasks in cloud ecosystems. (2024).
- [23] Robin Ohs, Gregory F. Stock, Andreas Schmidt, Juan A. Fraire, and Holger Hermanns. 2025. Dirty Bits in Low-Earth Orbit: The Carbon Footprint of Launching Computers. *SIGENERGY Energy Inform. Rev.* 5, 2 (Aug. 2025), 26–33. doi:10.1145/3757892.3757896
- [24] Yeon-Kyu Park, Geuk-Nam Kim, and Sang-Young Park. 2021. Novel structure and thermal design and analysis for cubesats in formation flying. *Aerospace* 8, 6 (2021), 150.
- [25] Cong Peng, Yuanzhi He, Shanghong Zhao, Lingyang Song, and Boyu Deng. 2023. Integration of Data Center into the Distributed Satellite Cluster Networks: Challenges, Techniques, and Trends. *IEEE Network* 37, 3 (2023), 52–58. doi:10.1109/MNET.105.2100614
- [26] Leon Poutievski, Omid Mashayekhi, Joon Ong, Arjun Singh, Mukarram Tariq, Rui Wang, Jianan Zhang, Virginia Beaugard, Patrick Conner, Steve Gribble, et al. 2022. Jupiter evolving: transforming google’s datacenter network via optical circuit switches and software-defined networking. In *Proceedings of the ACM SIGCOMM 2022 Conference*. 66–85.
- [27] Arvind R Singh, R Seshu Kumar, K Reddy Madhavi, Faisal Alsaif, Mohit Bajaj, and Ievgen Zaitsev. 2024. Optimizing demand response and load balancing in smart EV charging networks using AI integrated blockchain framework. *Scientific Reports* 14, 1 (2024), 31768.
- [28] Jesmin Jahan Tithi, Hanjiang Wu, Avishai Abuhatzera, and Fabrizio Petrini. 2025. Scaling intelligence: Designing data centers for next-gen language models. *arXiv preprint arXiv:2506.15006* (2025).
- [29] Guoping Wang, Gang Wan, Zhijuan Su, Yang Wang, Yutong Jia, Gong Li, and Shi Liang. 2025. High-Performance On-Orbit Intelligent Computing and Real-Time Services for Remote Sensing Satellites Based on Large-Scale Computing Power in Space. *IEEE Access* 13 (2025), 92114–92133. doi:10.1109/ACCESS.2025.3573932
- [30] Wan-fan Wu, Na Liu, Wen-long Cheng, and Yi Liu. 2013. Study on the effect of shape-stabilized phase change materials on spacecraft thermal control in extreme thermal environment. *Energy conversion and management* 69 (2013), 174–180.
- [31] Renchao Xie, Qinqin Tang, Qiuning Wang, Xu Liu, F. Richard Yu, and Tao Huang. 2020. Satellite-Terrestrial Integrated Edge Computing Networks: Architecture, Challenges, and Open Issues. *IEEE Network* 34, 3 (2020), 224–231. doi:10.1109/MNET.011.1900369
- [32] Yuzhe Xu, Thaha Mohammed, Mario Di Francesco, and Carlo Fischione. 2022. Distributed assignment with load balancing for DNN inference at the edge. *IEEE Internet of Things Journal* 10, 2 (2022), 1053–1065.
- [33] Blaise Agüera y Arcas, Travis Beals, Maria Biggs, Jessica V Bloom, Thomas Fischbacher, Konstantin Gromov, Urs Köster, Rishiraj Pravahan, and James Manyika. 2025. Towards a future space-based, highly scalable AI infrastructure system design. *arXiv preprint arXiv:2511.19468* 4 (2025).
- [34] Qing Ye, Yuhao Zhou, Mingjia Shi, Yanan Sun, and Jiancheng Lv. 2022. DLB: A dynamic load balance strategy for distributed training of deep neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence* 7, 4 (2022), 1217–1227.