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

**Shuyi Chen, Zhengchang Hua, Nikos Tziritas, Georgios Theodoropoulos**

*Southern University of Science and Technology & University of Thessaly*

SUSTech
Southern University of Science and Technology

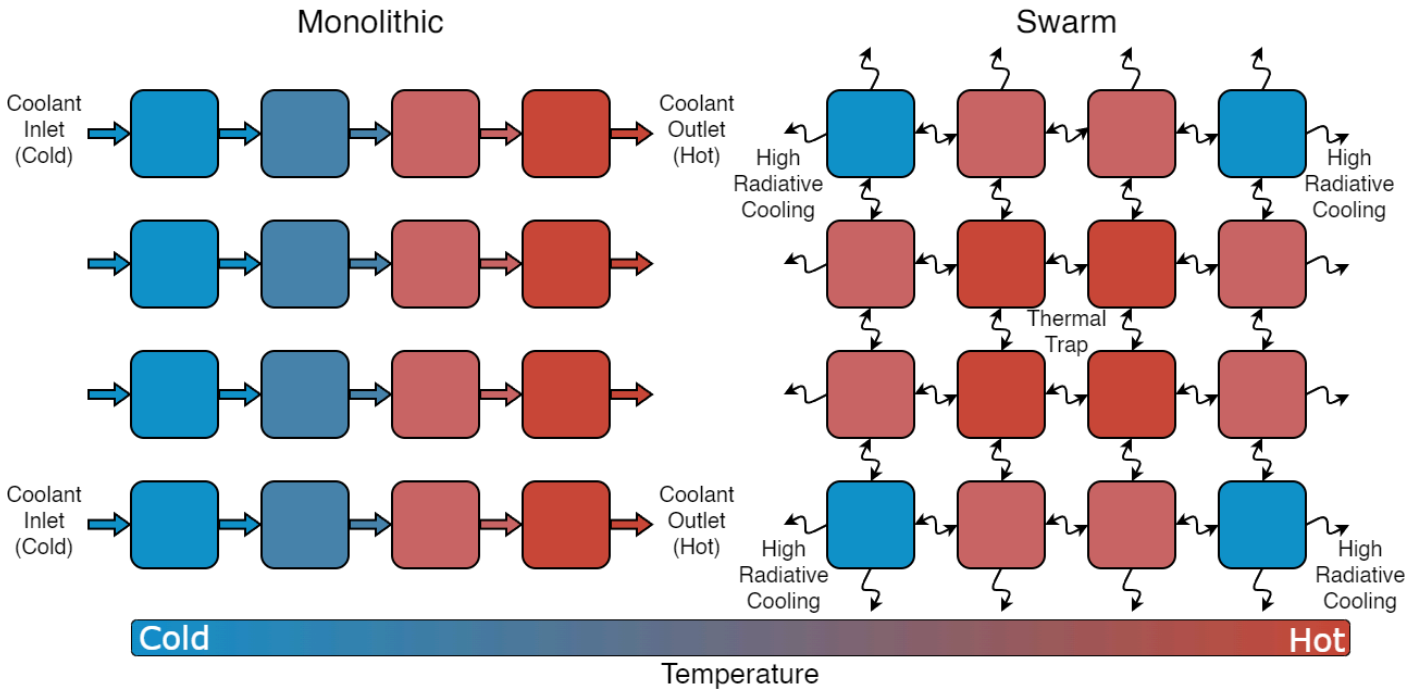
# The AI Crisis & Orbital Data Centers (ODCs)

- **The Problem**: Terrestrial AI training faces an unsustainable energy/water crisis.
- **The Solution**: Orbital Data Centers (ODCs) offer **zero operational carbon** via unlimited solar energy and deep-space cooling.
- **The Catch**: Rocket launches incur massive **embodied carbon**.
- To truly offset this, we must maximize **hardware longevity** and lifetime Model Flops Utilization (MFU); early hardware death defeats the sustainable premise.

# The Synchronous AI Challenge

SUSTech
Southern University of Science and Technology

- Synchronized workloads (e.g., LLM Tensor Parallelism) demand **sub-10$\mu$s latencies**, forcing **extreme high-density topologies**.
- ODCs scale into two paradigms: Monolithic Structures (shared cooling pipes) and Proximity Swarms (dense formations).



# The Proximity-Thermal Paradox

SUSTech
Southern University of Science and Technology

Extreme physical density triggers severe thermal crosstalk, trapping heat:

$$P_{in} = P_{idle} + \Delta P \left( \frac{S_{instant}}{S_{max}} \right)^{\gamma}$$

- **Thermal-Fluid (Monolithic)**: Downstream nodes choked by pre-heated fluid.

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

- **Thermal-Radiative (Swarms)**: Geometric shadowing ($\rho_i$) blocks deep-space radiation.

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

SUSTech
Southern University of Science and Technology

# The Straggler Problem & Space E-Waste

- Distributed training relies on **"All-or-Nothing"** synchronization:

$$t_{step} = \max_{i} \left( t_{comp}(i) + t_{comm} \right)$$

- A single node caught in a thermal trap throttles, stalling the entire gigawatt-scale cluster and collapsing global MFU.
- **The True Cost**: Cyclical thermal fatigue degrades silicon ($MTTF \propto \exp\left(\frac{E_a}{k_B T}\right)$), leading to premature **space e-waste**.

SUSTech
Southern University of Science and Technology

# The Core Thesis: Thermal-Aware Heterogeneity

- Traditional AI load-balancing assumes *uniform* cooling, blindly allocating data evenly to each node.
- **Our Position**: ODCs must explicitly reject uniform load-sharing.
- **The Paradigm Shift**: We must treat spatial and temporal cooling variance as a *primary, schedulable computational resource*.

# Thermal-Load Balancing (TLB)

TLB dynamically migrates computation to the coolest units by sizing micro-batches ($b_i$) proportionally to a real-time **Thermal Capability Score ($w_i$)**:

$$b_i \propto w_i$$

- **Monolithic Priority**: Upstream nodes receive larger workloads.

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

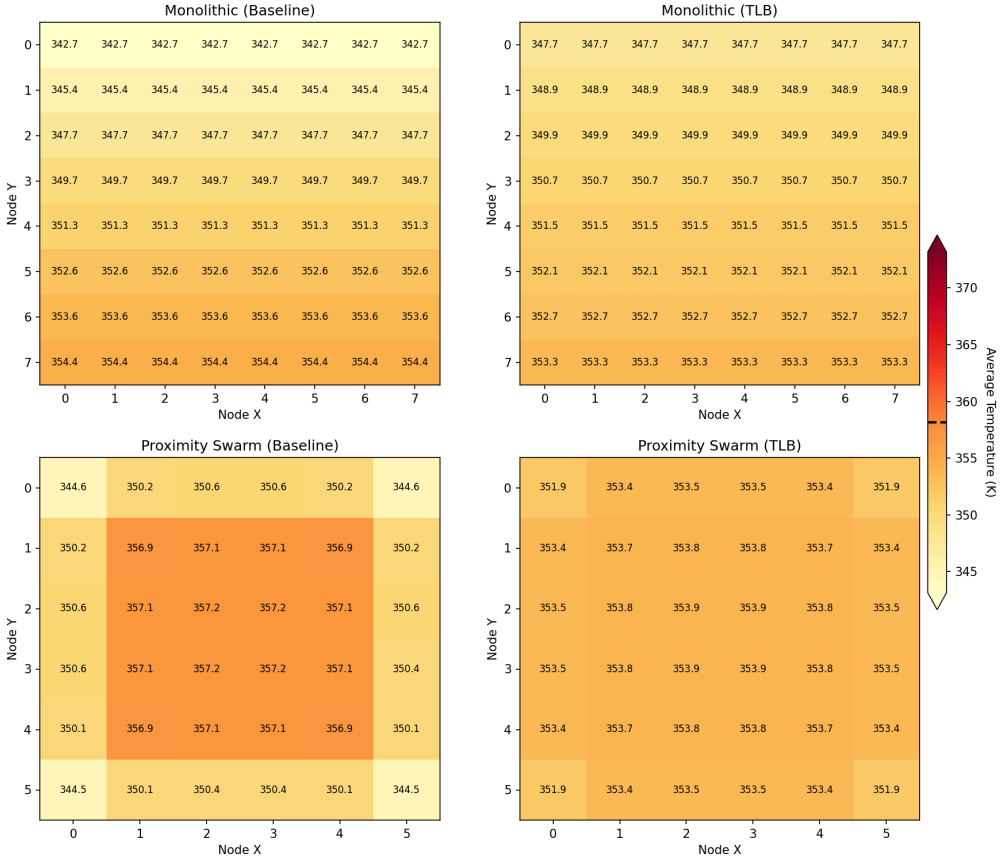
- **Swarm Priority**: Periphery nodes (high $\rho_i$) do the heavy lifting.

$$w_i = 1.0 + \alpha \cdot \rho_i$$

# Mitigating Heat Traps (Spatial Analysis)

SUSTech
Southern University of Science and Technology

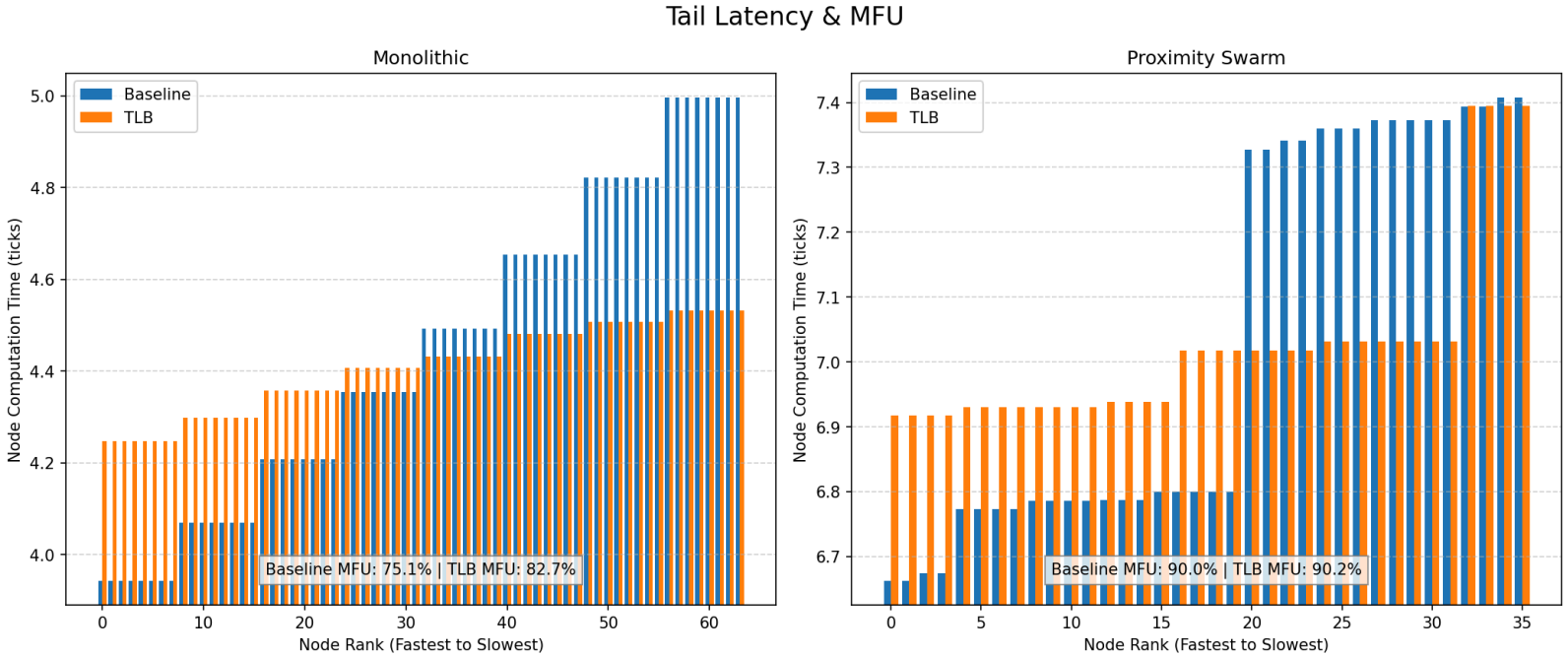
*Baseline uniform scheduling creates severe heat traps. TLB's asymmetric slicing neutralizes the cooling variance.*



# Prioritizing Hardware Survivability Over Throughput

*TLB equalizes computation times, eliminating the straggler effect.*

- **Hardware Survivability**: Extends MTTF (e.g., +6.15% in swarm cores) by capping peak temperatures, successfully amortizing embodied carbon.
- **Throughput**: Restoring MFU is a secondary byproduct of averting throttling.



# Open Challenges for the Community

SUSTech
Southern University of Science and Technology

To realize sustainable orbital computing, we must address:

1. **Dynamic Network-Thermal Co-design**: Realigning laser links physically changes satellite orientations. How do we co-optimize network topology and thermodynamics?
2. **The E-Waste vs. Carbon Trade-off**: Does zero operational carbon mathematically justify a paradigm of "disposable" high-end AI accelerators in orbit?
3. **Federated Thermal Telemetry**: Securely sharing thermal margins across multi-tenant swarms.

SUSTech
Southern University of Science and Technology

# Conclusion

- Orbital Data Centers offer a path to zero-carbon AI, but the **Proximity-Thermal Paradox** threatens their physical viability.
- Extending hardware survival via software scheduling is the indispensable prerequisite for sustainable space AI.
- **Thermal-Aware Heterogeneity** is a foundational concept for any distributed scientific workload constrained by barrier synchronization in space.

# Thank You!

# Questions?