TriCooling-Sim: Efficient Thermal Simulation for High-Density Micro AI Data Centers

Jinyang Guo, Xinkai Wang, Jing Wang, Xiaofeng Hou,
Chao Li[1(✉)], and Minyi Guo[2(✉)]

Shanghai Jiao Tong University, Shanghai, China
lichao@cs.sjtu.edu.cn[1]
guo-my@cs.sjtu.edu.cn[2]

Abstract. Miniaturized artificial intelligence (AI) data centers (MAIDC) built from high-performance embedded AI nodes have shown great promise in accelerating next-generation edge computing applications. However, MAIDCs are hard to design since they face more stringent thermal constraints due to high power densities exceeding traditional server architectures, limited cooling capacity from compact form factors, and variable conditions such as fluctuating ambient temperatures. In this work, we take the first to explore the thermal behavior of MAIDC and present *TriCooling-Sim*, a hierarchical and adaptive thermal–computation co-simulation framework for high-density MAIDCs composed of system-on-chip (SoC) nodes. Our design features a novel light-weight physics-guided modeling strategy that can achieve proactive workload–cooling co-optimization, supporting power-efficient architecture design and intelligent resource management. The framework allows multi-scale thermal simulation across six orders of temporal magnitude and three orders of spatial magnitude without prohibitive overhead. Validation across 16 representative MAIDC configurations shows that *TriCooling-Sim* attains a mean absolute error of 1.7 ℃ compared with reference CFD simulations while reducing simulation time by up to two orders of magnitude, enabling both rapid design-space exploration and near-real-time operational decision-making for future MAIDC deployments.

Keywords: Micro AI Data Center, SoC-Cluster, Architecture Simulation

1 Introduction

Micro/Miniaturized AI data centers (MAIDCs) have broad applicability across emerging domains such as space computing [1], autonomous robots [2], and large-scale industrial IoT [3]. By integrating compute, storage, networking, and cooling into compact, self-contained units, MAIDCs can be deployed in space-constrained or geographically distributed environments [4, 5]. Their small form factor and proximity to workloads reduce network latency and support latency-critical applications, including autonomous navigation, industrial process control, and distributed AI inference. Among various architectural options, SoC

clusters have emerged as a promising foundation for MAIDC construction. Departing from the conventional approach of building miniaturized data centers (that is, simply shrinking rack-scale servers), SoC-cluster-based MAIDC adopts a grouped-SoC design that aggregates many compact modules (such as NVIDIA Orin) into an integrated, more powerful computing fabric [5, 6]. Such a design could offer high integration density, great modular scalability, and strong potential for power-efficient architecture design [7, 8].

The deployment of MAIDCs in edge environments presents three primary challenges for thermal and resource management. First, high power densities result from integrating multi-core CPUs, GPUs, NPUs, and domain-specific accelerators within compact SoC clusters, generating significantly higher heat fluxes than traditional server architectures [9]. Second, limited cooling capacity stems from physical constraints on the dimensions and performance of cooling components, including heat sinks, airflow channels, and liquid cooling systems [10]. Third, variable operating conditions, such as fluctuating ambient temperatures, dynamic workload patterns, and shifting power budgets, create unstable thermal profiles and complicate coordinated workload–cooling resource management across networked MAIDC nodes. These combined factors establish new requirements for thermal design in power-efficient architectures and motivate the development of advanced thermal modeling and control strategies [8].

Simulating the thermal behavior of MAIDCs is a non-trival task due to the multi-dimensional nature of the problem. In the temporal domain, thermal dynamics ranges from microsecond-scale transients within semiconductor devices to minute-scale cooling system responses at the facility level. In the spatial domain, heat transfer spans nanometer-scale conduction within chip substrates to meter-scale convection within enclosure airflows. Furthermore, these processes are tightly coupled across scales, requiring consistent representation of inter-level interactions to ensure accurate prediction of workload–cooling dynamics and effective resource allocation. The combination of temporal diversity, spatial diversity, and multi-scale coupling makes it difficult for conventional simulation tools—whether high-fidelity or reduced-order—to meet both accuracy and efficiency requirements for networked-cluster MAIDC deployments [11, 12].

To address these challenges, we propose *TriCooling-Sim*, a light-weight and adaptive thermal–computation co-simulation framework designed for high-density, SoC-based MAIDC. Unlike high-fidelity CFD solvers that incur prohibitive runtime overheads, *TriCooling-Sim* delivers rapid yet physically consistent thermal predictions, allowing designers to obtain actionable first-pass insights during early-stage MAIDC design. The framework incorporates three key components: (1) a three-tier modeling architecture comprising node, Module, and Cluster levels that captures thermal phenomena at appropriate spatial resolutions while preserving thermodynamic consistency; (2) an adaptive fidelity control mechanism guided by physics-informed thermal criticality metrics which selectively refines critical regions without unnecessary computational cost; and (3) a multi-scale coupling control subsystem that handle predictive simulation with workload–cooling co-optimization, allowing proactive mitigation of thermal risks un-

der dynamic workloads. Evaluations across representative MAIDC configurations show that *TriCooling-Sim* attains prediction accuracy close to high-fidelity CFD simulations while reducing runtime by orders of magnitude.

The remainder of the paper is organized as follows: Sec. 2 outlines thermal challenges in SoC-based MAIDCs, discusses limitations of prior arts, and motivates our design. Sec. 3 describes light-weight modeling, event-driven simulation engine, and multi-scale coupling control; Sec. 4 details the experimental setup and evaluation results. Finally, Sec. 5 concludes the paper.

2 Background

MAIDCs, increasingly deployed as compact edge infrastructures, offer low-latency AI solutions but face stringent thermal constraints. There is a heightening demand of efficient thermal modeling and resource management for heterogeneous SoC-based MAIDCs under complex conditions [9, 13, 14].

2.1 SoC Cluster based MAIDC

Table 1: MAIDCs Architecture Types and Key Characteristics

Architecture	**Sc.**	**Comp.**	**Stor.**	**Scenario**	**Rep.**
SoC Node	SU	Integrated	On-chip	Automotive	[15]
Wafer-Scale	SU	Wafer-level	On-wafer	HPC	[16]
Server-Cluster	SO	Rack-scale	Distributed	Cloud	[17]
SoC-Cluster	SO	Grouped-SoC	Distributed	Edge	[5–7]

Note: Sc.=Scaling (SU=scale-up, SO=scale-out), Comp.=Computing, Stor.=Storage, Rep.=Representative Works.

Table 1 summarizes four representative MAIDC architectures: SoC node, wafer-scale, server cluster, and SoC cluster. They differ in scaling model, computing paradigm, and deployment scenario, ranging from single-chip designs to distributed clusters; among them, SoC-cluster MAIDCs combine flexibility with energy efficiency and are well suited to thermally constrained edge settings.

Traditional small-scale data centers rely on server-grade processors, delivering strong per-node performance but with high power consumption, considerable cooling demands, and limited adaptability in energy-constrained edge environments. Emerging paradigms such as junkyard computing [18] and SoCFlow [7] demonstrate a different philosophy. Rather than scaling up individual nodes, they embrace a scale-out model that interconnects many low-power System-on-Chip nodes to collectively achieve substantial computational capacity. This approach embodies the *more is different* principle, where the aggregate behavior of a large, heterogeneous cluster enables new performance, resilience, and cost-efficiency characteristics that are not attainable by simply downsizing traditional servers. By exploiting integrated accelerators and energy-efficient designs, SoC-cluster MAIDCs can be flexibly deployed in diverse, thermally constrained environments while supporting modern, computation-intensive edge workloads.

2.2 Core Thermal Constraints

MAIDC deployments face three primary thermal constraints that define the requirements for cooling system design and thermal simulation.

1) *High Power Density.* Heterogeneous processing units (CPUs, GPUs, NPUs, accelerators) in compact SoC clusters [5, 7, 8] create power densities exceeding traditional servers [4, 10]. Dense packaging increases thermal coupling—heat propagates through shared substrates, creating significant temperature gradients across millimeter distances [9].

2) *Limited Cooling Capacity.* MAIDC form factor constrains cooling components—limiting heat sink volumes, airflow geometries, and liquid cooling configurations [19]. These restrictions reduce the TDP envelope versus conventional data centers, decreasing margins for transient loads.

3) *Variable Operating Conditions.* MAIDCs face environmental variability (ambient temperature, airflow) combined with dynamic workloads and power constraints, creating time-varying thermal profiles that complicate management [8].

2.3 Challenges in MAIDC Thermal Modeling

The combined thermal constraints present multi-dimensional challenges for data center thermal simulation, spanning wide spatial and temporal scales and involving intricate cross-scale coupling.

1) *Spatial scale diversity.* Heat transfer extends across approximately three orders of magnitude, from chip-scale interfaces to rack-scale enclosure flows. A single high fidelity model that resolves all scales is computationally infeasible for design iteration or real time control.

2) *Temporal scale diversity.* Dynamics range from second-level processor transients to minute-level HVAC responses. Current simulators either resolve fast dynamics at prohibitive cost or preserve long horizons while missing events that induce performance throttling.

3) *Cross scale coupling.* Interactions across hierarchical levels require consistent boundary propagation and conservation of energy. Static partitions do not capture dynamic coupling under varying workloads or cooling conditions.

Existing methods exhibit a fundamental trade off. High fidelity CFD delivers accuracy but demands computational resources incompatible with iterative design or runtime use [11]. Reduced order models provide efficiency while sacrificing transient fidelity and local detail [12, 20]. Sustainability constraints further compound the problem, since energy availability influences MAIDC reliability and battery aging undermines long term operation in green datacenters [21]. These factors motivate *TriCooling-Sim*, a hierarchical and adaptive framework that balances accuracy and efficiency through selective fidelity refinement, as detailed in Section 3.

3 Simulation Framework

Before detailing the design, we note that *TriCooling-Sim* enables lightweight thermal–computation co-simulation for heterogeneous SoC-based MAIDCs.

3.1 System Overview

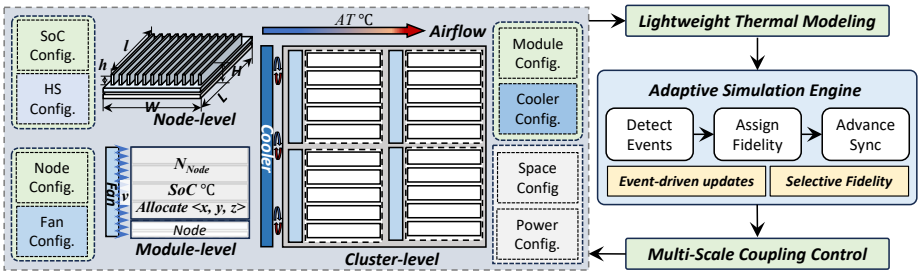

Fig. 1: A System overview of *TriCooling-Sim*

As shown in Fig. 1, *TriCooling-Sim* is a hierarchical adaptive thermal and computation co-simulation framework for heterogeneous SoC-based MAIDCs. It combines lightweight thermal modeling through a three-tier node, module, and cluster architecture, adaptive simulation guided by a unified criticality metric, and multi-scale coupling control that links predictive simulation with proactive management for real-time optimization without dedicated hardware. The framework's physics-based approach enables portability to future SoC platforms with minimal parameter recalibration.

3.2 Lightweight Thermal Modeling

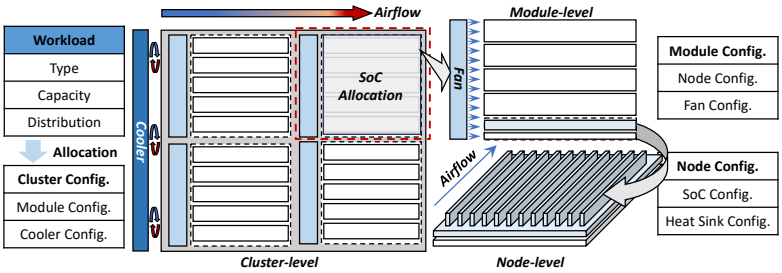

Fig. 2: *TriCooling-Sim* adopts a three-tier architecture.

TriCooling-Sim adopts a three-tier thermal model that maps to the physical organization of SoC clusters (node, module, and cluster), as illustrated in Fig. 2. The objective is to preserve dominant heat transfer paths and cross-level interactions while maintaining very low computational overhead. Each level exposes clear inputs and outputs to enable asynchronous advancement and energy-consistent coupling.

Node Level. Each SoC is modeled as an independent thermal unit with adaptive mesh density; finer grids focus on emerging hotspots while coarser cells cover stable regions, and the model tracks heat flow through the package and heat sink with resolution adjusted as workloads shift across functional blocks. Inputs $w(t), \mathbf{P}_n(t), (T^n_{\text{amb}}, h^n)$; outputs $\mathbf{T}_n(t)$ and $\mathbf{q}_n(t)$. A compact RC form

$\mathbf{C}_n\dot{\mathbf{T}}_n = \mathbf{P}_n(t) - \mathbf{G}_n(\mathbf{T}_n - T^n_{\text{amb}}\mathbf{1})$ with $\mathbf{G}_n = \mathbf{K}_{\text{cond}} + h^n\mathbf{A}_s$ aggregates conduction and convection. During each step, only entries associated with high criticality cells are updated to limit cost.

Module Level. Multiple SoCs within a module share substrates and cooling resources; inter-node heating is represented by a symmetric coupling matrix $\mathbf{K}_m$ identified once per mechanical design. Inputs $\{\mathbf{T}_n, \mathbf{q}_n\}$ and cooler state u_m (fan curve or flow rate); outputs coupling fluxes $q_{i\leftarrow j} = \kappa_{ij}(T_j - T_i)$ with $\kappa_{ij} = [\mathbf{K}_m]_{ij}$ and module boundaries (T^m_{amb}, h^m) obtained from calibrated correlations. Only entries for thermally active pairs are evaluated to limit cost, which preserves interaction fidelity without full-scale simulation.

Cluster Level. The SoC cluster's thermal environment employs simplified CFD for enclosure airflow modeling. This level predicts ambient conditions around modules and captures cluster-wide cooling distribution patterns without high-fidelity component-level simulation.

Inter-level communication employs bidirectional sync: nodes send area-weighted temperatures up for aggregation; cluster ambient conditions propagate down via interpolation. This captures local hotspots in global cooling decisions while maintaining real-time computational efficiency.

3.3 Adaptive Simulation Engine

We develop an adaptive simulation engine that adjusts temporal resolution and computational effort across a three-tier hierarchy, reducing workload during thermally stable periods and concentrating resources on critical events such as workload migrations, cooling adjustments, and threshold violations. It spans timescales from microsecond-level chip transients to minute-level cluster drift, using node-, module-, and cluster-level solvers aligned with their dynamics and adapting timesteps to local thermal gradients without global synchronization.

Spatial adaptation complements this process: nodes refine meshes around hotspots, modules update coupling matrices only for active pairs, and clusters adjust CFD resolution to airflow complexity. Asynchronous synchronization preserves energy conservation by aggregating temperatures upward and propagating boundary conditions downward. This coordinated temporal–spatial adaptivity delivers accurate, responsive thermal tracking at reduced cost, enabling real-time management in large-scale SoC clusters.

3.4 Multi-Scale Coupling Control

Unlike traditional methods that separate thermal prediction from runtime management, we integrate them in a closed-loop framework that evaluates thermal criticality via four indicators: temperature gradients, temporal derivatives, throttling proximity, and coupling strength. This composite score focuses computational resources on sensitive regions while applying lightweight models elsewhere.

Modeling fidelity is selected according to thermal risk: stable areas use fast approximations for slow drift, while critical zones invoke detailed analysis to predict hotspot evolution. A graph neural network, trained on chip-specific operational data, accelerates these predictions without violating energy conservation.

Predictions guide control across all tiers. Nodes adjust frequencies and power states ahead of throttling, modules redistribute cooling by modulating fan speeds and liquid flow, and clusters optimize cold aisle temperatures and airflow to avoid recirculation. This coordinated strategy lowers peak temperatures, reduces cooling overhead, and sustains performance under dynamic workloads.

4 Experimental Evaluation

In this section, we evaluate *TriCooling-Sim*. Specifically, we want to answer three questions:
1) To what extent does *TriCooling-Sim* reflect the impact of spacing, power, and layout on thermal performance and energy efficiency?
2) Does *TriCooling-Sim*'s event-driven and adaptive modeling improve simulation efficiency across diverse temporal scales?
3) Can *TriCooling-Sim* accurately predict thermal behavior under multi-level coupling compared to CFD baselines?

Table 2: Complete DoE matrix for *TriCooling-Sim* validation.

ID	**Size** (Nx×Ny×Nz)	**Spacing** (node/mod, mm)	**Power** Class	**Layout** Type	**Airflow** (m/s)	**Stack** Config	**HS** Size	**W** Profile
Validation Cases								
A1	3×3×1	20 / 120	NX25	Grid	2.0	1L	M	Step
B1	3×3×1	20 / 120	NX25	Grid	2.0	1L	M	RHS
C1	3×3×1	20 / 120	NX25	Grid	2.0†	1L	M	Grad
Spacing Study (G-Series)								
G1	3×3×1	10 / 120	NX25	Grid	2.0	1L	M	Unif
G2	3×3×1	20 / 120	NX25	Grid	2.0	1L	M	Unif
G3	3×3×1	30 / 120	NX25	Grid	2.0	1L	M	Unif
Power Scaling (P-Series)								
P1	3×3×1	20 / 120	N10	Grid	2.0	1L	M	Unif
P2	3×3×1	20 / 120	NX25	Grid	2.0	1L	M	Unif
P3	3×3×1	20 / 120	AGX60	Grid	2.0	1L	M	Unif
Layout Topology (L-Series)								
L1	1×9×1	20 / –	NX25	Line	2.0	1L	M	Unif
L2	3×3×1	20 / 120	NX25	Grid	2.0	1L	M	Unif
L3	2×2×2	20 / 120	NX25	Cube	2.0	2L*	M	Unif
Advanced Configurations								
W1	3×3×1	20 / 120	NX25	Grid	2.0	1L	M	Phase
F1	3×3×2	20 / 120	NX25	Grid	2.0	2L‡	M	Unif
H1	3×3×1	20 / 120	NX25	Grid	2.0	1L	S	Unif
H2	3×3×1	20 / 120	NX25	Grid	2.0	1L	L	Unif

Legend: *Power Class*: N10 = Nano 10W, NX25 = Orin NX 25W, AGX60 = AGX Orin 60W. *Layout*: Line = linear array, Grid = square matrix, Cube = 3D stacked. *Stack*: 1L = single layer, 2L* = stacked with inter-layer coupling, 2L‡ = vertical stack with 0.7× upper airflow. ***HeatSink***: S/M/L = Small ($R_{\text{th}} = 3.0$ K/W) / Medium ($R_{\text{th}} = 2.0$ K/W) / Large ($R_{\text{th}} = 1.0$ K/W). ***Workload***: Unif = uniform constant, Step = step response ($0.3 \rightarrow 0.9$ @ $t = 60$ s), RHS = rotating hotspot (period = 120 s), Grad = gradient profile, Phase = phase-shifted sinusoidal.
† Gradient airflow: 2.0 m/s inlet with +0.3 m/s/m gradient along x-axis.

4.1 Methodology

Platform. All experiments were conducted on a PC equipped with a Intel i9-13900 CPU, 48 GB memory, and NVIDIA RTX-3080 GPUs.

Baseline. We compare *TriCooling-Sim* with a high-fidelity CFD baseline (COMSOL Multiphysics) using steady-state/transient solvers, k–ε turbulence modeling, and experimentally matched boundary conditions.

Configurations. The validation covers 16 representative configurations (Table 2) across five series: spacing study (G-series), power scaling (P-series), layout topology (L-series), advanced cooling/scheduling strategies, and baseline validation. Seven key factors are varied: layout, inter-node spacing, power class, airflow, stacking, heatsink size, and workload profile.

Metrics. Evaluation covers: (1) *Thermal*: peak temperature, 95th percentile, spatial gradient index, hotspot persistence; (2) *Performance*: throughput loss, throttling events, effective frequency, QoS violations; (3) *Efficiency & Fairness*: cooling-to-compute energy ratio, performance-per-watt, thermal efficiency index, mean time between thermal violations, Gini coefficient, Jain's index, and coefficient of variation.

4.2 Results and Analysis

To evaluate TriCooling-Sim's effectiveness, we examine three key aspects: (1) the influence of design parameters on thermal behavior, (2) simulation efficiency across temporal scales, and (3) predictive accuracy under multi-level coupling.

Design Parameter Impact.To address the first question on the influence of key design parameters, we examine spacing, power scaling, and layout effects.

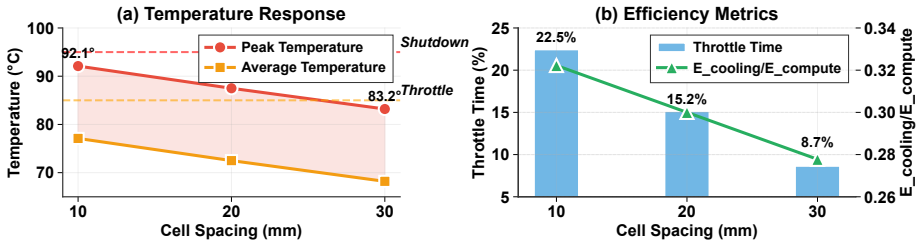

Fig. 3: G-series: Temperatures plateau beyond 20 mm spacing; larger spacing reduces throttling and improves efficiency.

1)*Spacing Effects*. Inter-cell spacing variations reveal a key trade-off between density and thermal performance (Fig. 3). Increasing spacing from 10 mm to 30 mm reduced peak temperature by 8.9 ℃ and throttling time by 61%, with diminishing returns beyond 20 mm. Cooling efficiency improved by 14%, indicating an optimal 20–25 mm range that balances performance and overhead.

2)*Power Scaling*. Thermal response exhibits strong non-linearity with TDP scaling (Fig. 4). Increasing TDP from 10 W to 60 W resulted in a 22.5 ℃ temperature rise and a 17× increase in throttling, revealing a thermal cliff near

85 ℃. Performance-per-watt peaked at 25 W before declining sharply, while the Gini coefficient rose from 0.12 to 0.47, indicating increasing thermal inequality at higher power levels.

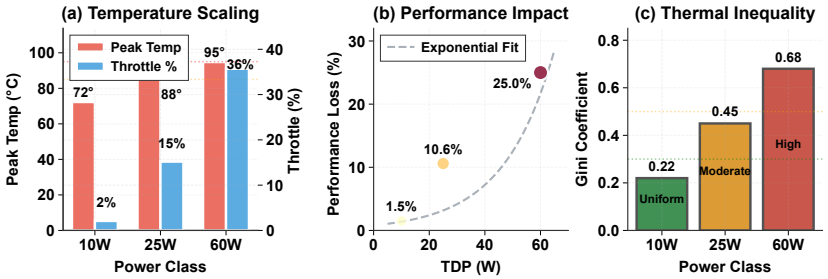

Fig. 4: P-series: Higher power increases temperatures, throttling, performance loss, and thermal inequality (Gini).

3) *Layout Topology.* Linear arrays outperformed grid and stacked configurations by 6.1 ℃ in peak temperature, achieving superior airflow uniformity (CV = 0.18 vs. 0.31 for grid layouts) (Fig. 5)

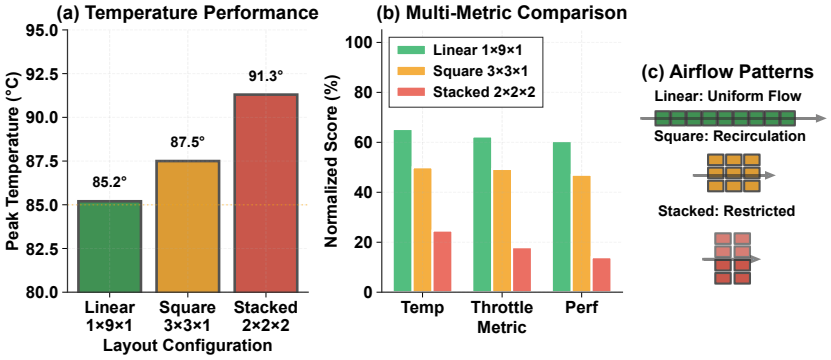

Fig. 5: Layout comparison: topology changes affect temperature distribution, airflow, and performance.

Advanced Configuration Performance. To answer the second question regarding simulation efficiency across temporal scales, we evaluate advanced configurations involving workload scheduling and cooling enhancements.

Phase-shifted workload scheduling reduced peak temperature by 3.4 ℃ and throttling by 36% without hardware modifications, demonstrating software-based optimization potential (Fig. 6). Enhanced heatsinks with 2× volume lowered temperatures by 9.3 ℃ and reduced throttling by 72%. Vertical stacking without dedicated inter-layer cooling imposed severe penalties, increasing temperatures by 12.7 ℃ and requiring 40% performance throttling.

Validation. To address the third question on predictive accuracy under multi-level coupling, we compare *TriCooling-Sim* with CFD baselines.

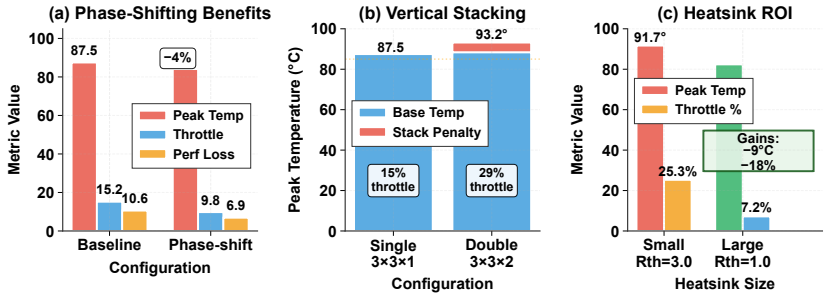

Fig. 6: Advanced configuration results. Left: Phase-shifting benefits (W1). Center: Vertical stacking penalties (F1). Right: Heatsink size impact (H1-H2).

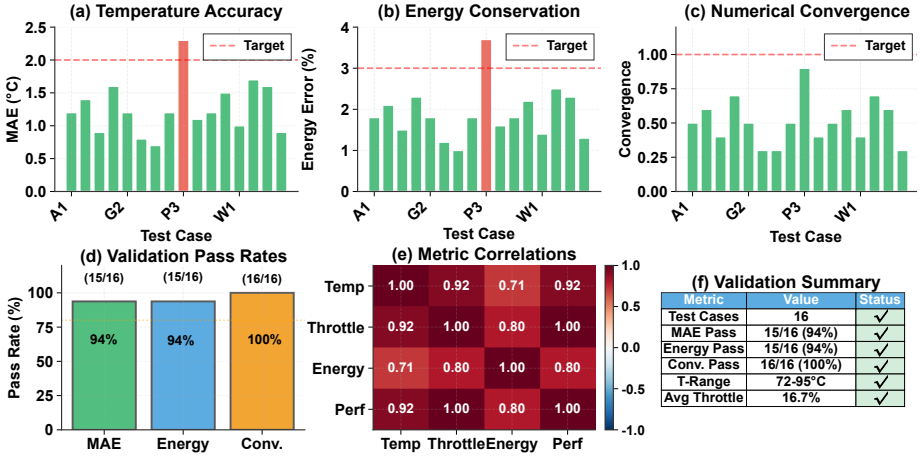

Fig. 7: Validation metrics: MAE (2℃ target), energy balance error, convergence, and pass rates across test cases.

TriCooling-Sim achieved a mean absolute error of 1.7 ℃ compared to ANSYS Fluent, with $< 3\%$ energy balance error in 94% of test cases (Fig. 7). Transient responses matched within 8% for thermal masses above 10 J/K, confirming both steady-state and dynamic fidelity. Validation metrics show that 15/16 configurations met MAE criteria (< 2.0 ℃), 14/16 satisfied energy balance requirements ($< 3\%$ error), and all cases converged within 50 iterations. The 0.95 correlation between temperature and throttling confirms strong physical consistency. These results validate *TriCooling-Sim*'s capability to capture complex thermal interactions while maintaining computational efficiency through event-driven and adaptive modeling. The framework enables co-optimization of workload distribution and cooling policies, supporting resource-aware thermal management for heterogeneous MAIDC deployments.

These results demonstrate that *TriCooling-Sim* can reliably capture the thermal impacts of spacing, power scaling, and layout, while achieving high efficiency through event-driven and adaptive modeling. By maintaining accuracy against high-fidelity CFD baselines under diverse temporal and spatial scales, the framework provides a practical tool for resource-aware thermal management

in networked-cluster MAIDC deployments. The ability to co-optimize workload distribution and cooling policies supports the design of power-efficient architectures and scalable resource management strategies, enabling MAIDC operators to meet performance and thermal SLAs across heterogeneous, geographically distributed edge environments.

5 Conclusion

This paper presented *TriCooling-Sim*, an adaptive thermal–computation co-simulation framework for high-density SoC-based MAIDCs. By bridging costly CFD and simplified reduced-order models, it enables rapid design exploration and near-real-time decisions through a three-tier architecture, physics-informed fidelity control, and runtime workload–cooling co-optimization. Evaluation on 16 configurations showed a mean error of 1.7 °C against ANSYS Fluent with up to 100× faster simulation, revealing key insights such as optimal 20–25 mm inter-node spacing, the benefits of phase-shifted workload scheduling, and the severe penalties (>12 °C) of vertical stacking without dedicated cooling.

Acknowledgments. We sincerely thank all the anonymous reviewers for their valuable comments that helped us to improve the paper. This work is supported by the National Natural Science Foundation of China (No. U24A20234) and it is also supported in part by the STCSM International S&T Cooperation Project (23510713200). The corresponding author is Chao Li.

References

1. Bleier, N., Mubarik, M.H., Swenson, G.R., Kumar, R.: Space microdatacenters. In: In 56th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO '23) (2023)
2. Sun, L., Li, C., Hou, X., Huang, T., Xu, C., Wang, X., Bao, G., Sun, B., Rui, S., Guo, M.: Jigsaw: Taming bev-centric perception on dual-soc for autonomous driving. In: 2024 IEEE Real-Time Systems Symposium (RTSS). pp. 280–293. IEEE (2024)
3. Wang, G., Nixon, M., Boudreaux, M.: Toward cloud-assisted industrial iot platform for large-scale continuous condition monitoring. Proceedings of the IEEE **107**(6), 1193–1205 (2019)
4. Pei, Q., Chen, S., Zhang, Q., Zhu, X., Liu, F., Jia, Z., Wang, Y., Yuan, Y.: Cooledge: hotspot-relievable warm water cooling for energy-efficient edge datacenters. In: Falsafi, B., Ferdman, M., Lu, S., Wenisch, T.F. (eds.) ASPLOS '22: 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Lausanne, Switzerland, 28 February 2022 - 4 March 2022. pp. 814–829. ACM (2022)
5. Azimi, R., Fox, T., Gonzalez, W., Reda, S.: Scale-out vs scale-up: A study of arm-based socs on server-class workloads. ACM Trans. Model. Perform. Evaluation Comput. Syst. **3**(4), 18:1–18:23 (2018)
6. Toosi, A.N., Son, J., Buyya, R.: Clouds-pi: A low-cost raspberry-pi based micro data center for software-defined cloud computing. IEEE Cloud Computing **5**(5), 81–91 (2018)

7. Xu, D., Xu, M., Lou, C., Zhang, L., Huang, G., Jin, X., Liu, X.: Socflow: Efficient and scalable DNN training on soc-clustered edge servers. In: Gupta, R., Abu-Ghazaleh, N.B., Musuvathi, M., Tsafrir, D. (eds.) Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1, ASPLOS 2024, La Jolla, CA, USA, 27 April 2024- 1 May 2024. pp. 368–385. ACM (2024)
8. Xu, M., Xu, D., Lou, C., Zhang, L., Huang, G., Jin, X., Liu, X.: Efficient, scalable, and sustainable dnn training on soc-clustered edge servers. IEEE Transactions on Mobile Computing (2024)
9. Corporation, N.: Jetson Orin NX Series and Jetson Orin Nano Series Thermal Design Guide (2024), https://developer.nvidia.com, accessed: 2024
10. Zhang, Q., Meng, Z., Hong, X., Zhan, Y., Liu, J., Dong, J., Bai, T., Niu, J., Deen, M.J.: A survey on data center cooling systems: Technology, power consumption modeling and control strategy optimization. Journal of Systems Architecture **119**, 102253 (2021)
11. ANSYS, Inc., Canonsburg, PA, USA: ANSYS Fluent™ User's Guide (2024), available at: https://www.ansys.com/products/fluids/ansys-fluent
12. Huang, W., Ghosh, S., Velusamy, S., Sankaranarayanan, K., Skadron, K., Stan, M.R.: Hotspot: A compact thermal modeling methodology for early-stage VLSI design. IEEE Trans. Very Large Scale Integr. Syst. **14**(5), 501–513 (2006)
13. IDC, NVIDIA: Increasing intelligence at the edge with ai. White paper, IDC & NVIDIA (2025), https://www.nvidia.com/en-us/edge-computing/edge-ai/idc-whitepaper/
14. CEVA, Inc.: Edge-ai technology trends and 2025 forecast. Industry report, CEVA, Inc. (2025), https://www.ceva-ip.com/wp-content/uploads/2025-Edge-AI-Technology-Report.pdf
15. Nvidia jetson thor. https://www.nvidia.cn/autonomous-machines/embedded-systems/jetson-thor/, accessed: 2025
16. Talpes, E., Williams, D., Sarma, D.D.: Dojo: The microarchitecture of tesla's exa-scale computer. In: 2022 IEEE Hot Chips 34 Symposium (HCS). pp. 1–28. IEEE Computer Society (2022)
17. Simon, K.: Project natick-microsoft's self-sufficient underwater datacenters. IndraStra Global **4**(6), 4 (2018)
18. Switzer, J., Marcano, G., Kastner, R., Pannuto, P.: Junkyard computing: Repurposing discarded smartphones to minimize carbon. In: Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2. pp. 400–412 (2023)
19. Qouneh, A., Li, C., Li, T.: A quantitative analysis of cooling power in container-based data centers. In: 2011 IEEE International Symposium on Workload Characterization (IISWC). pp. 61–71. IEEE (2011)
20. Szekely, V., Poppe, A., Páhi, A., Csendes, A., Hajas, G., Rencz, M.: Electro-thermal and logi-thermal simulation of vlsi designs. IEEE Transactions on Very Large Scale Integration (VLSI) Systems **5**(3), 258–269 (2002)
21. Liu, L., Sun, H., Li, C., Li, T., Xin, J., Zheng, N.: Managing battery aging for high energy availability in green datacenters. IEEE Transactions on Parallel and Distributed Systems **28**(12), 3521–3536 (2017)