MedCrave
Step into the World of Research
Material Science & Engineering International Journal

Mini Review
Open Access
CrossMark

# Industry trends, simulation-guided optimization, and hotspot-aware zoned cooling for high-power Artificial Intelligence (AI) chips

## Abstract

As AI accelerators push beyond 700 W per package, thermal management has shifted from a packaging afterthought to a first-order constraint on performance, reliability, and rack-level scalability. This paper presents a simulation-guided thermal design study of a representative AI-chip stack: 700 W dissipated across an 812.25 mm² die, equivalent to 8.12 cm², corresponding to an average heat flux of 86.13 W/cm². The study couples this model with an industry review of cooling strategies ranging from forced air to embedded microfluidics. A five-layer chip stack consisting of Thermal Interface Material (TIM), Integrated Heat Spreader (IHS), and a cold plate was developed in ANSYS® Mechanical and assessed by energy balance, mesh convergence, and analytical sanity checks rather than by experimental validation. The fixed-temperature sink boundary condition was used to isolate package conduction and spreading effects; it does not resolve coolant heating, pressure drop, or internal cold-plate flow. Within this fixed-sink extraction procedure, the same-footprint bottleneck extraction indicates that the cold plate and IHS dominate the measured package-level temperature decrease, while the extracted TIM drop is small and is interpreted cautiously. Systematic geometry sweeps show that changing IHS and cold plate thicknesses reduces peak die temperature by 14.75 °C between the worst and the best passive configurations within the idealized fixed-sink model. Building on the optimized geometry, a hotspot-aware zoned cooling concept, implemented as spatially non-uniform convection on the cold plate surface, reduces die peak temperature by up to 4.01 °C for center-hotspot and corner-hotspot workloads by redistributing cooling capacity toward the active hotspot rather than increasing the overall cooling strength. A cyber-physical control architecture is then proposed, pairing a Graph Neural Network (GNN) and Gated Recurrent Unit (GRU) predictor with a constrained zone allocation controller for future real-time redirection of cooling effort as workload hotspots migrate; this controller is presented as a proposed framework, not as a trained or experimentally validated implementation.

**Keywords:** thermal management, AI accelerators, direct to chip liquid cooling, integrated heat spreader, ANSYS® mechanical, parameter sweep, hotspot-aware zoned cooling, graph neural network, gated recurrent unit, cyber-physical systems.

Volume 10 Issue 2 - 2026

Ved Dwivedi,[1,2] Balraj S. Mani,[3] Nuggehalli M. Ravindra[1,4]
[1]Department of Physics, New Jersey Institute of Technology, USA
[2]John P. Stevens High School, USA
[3]Department of Mechanical and Industrial Engineering, New Jersey Institute of Technology, USA

**Correspondence:** Nuggehalli M. Ravindra, Department of Physics, New Jersey Institute of Technology, Newark, New Jersey, USA

**Received:** May 26, 2026 | **Published:** June 08, 2026

## Nomenclature

| Symbol | Definition | Unit |
|---|---|---|
| A | Area | m², cm², or mm² as specified |
| $A_{die}$ | Die area | cm² or mm², as specified |
| $A_j$, Az | Area of tile or cooling zone | m² |
| $a_{i,t}$ | Compute-activity proxy at tile i and time t | dimensionless or normalized counter |
| B | Cooling-budget constraint | W/K or area-weighted h budget |
| h | Convective heat-transfer coefficient | W/m²·K |
| $\bar{h}$ | Area-weighted mean heat-transfer coefficient | W/m²·K |
| $h_z$ | Heat-transfer coefficient assigned to cooling zone z | W/m²·K |
| k | Thermal conductivity | W/m·K |
| λ, μ | Controller weighting coefficients | dimensionless or normalized cost weights |
| $m_{i,t}$ | Memory-traffic proxy at tile i and time t | dimensionless or normalized counter |
| $\mathcal{N}(i)$ | Set of neighboring tiles adjacent to tile i | dimensionless |
| P | Total heat input | W |
| $P_{i,t}$ | Tile power or normalized power proxy at tile i and time t | W if physical; dimensionless if normalized |
| q" | Average die heat flux, or heat input per unit die area | W/cm² |
| $r_{max}$ | Maximum allowed actuation change per control period | W/m²·K per control period or normalized units |
| T | Temperature | °C |

| Symbol | Definition | Unit |
|---|---|---|
| $T_{i,t}$ | Temperature of tile i at time t | °C |
| $T_{max}$ | Maximum die temperature | °C |
| $\Delta T$ | Temperature difference or temperature drop | °C |
| $u_{min}$, $u_{max}$ | Lower and upper actuation limits | W/m² ·K or normalized actuation units |
| $u_t$ | Cooling-actuation vector at time t | W/m² ·K or normalized actuation units |
| $u_t$* | Optimal cooling-actuation vector at time t | same as $u_t$ |
| xi,t | Feature vector for tile i at time t | mixed/normalized |

| Abbreviation | Description |
|---|---|
| AI | Artificial Intelligence |
| CDU | Coolant Distribution Unit |
| DTC | Direct-to-Chip |
| GNN | Graphical Neural Networks |
| GRU | Gated Recurrent Unit |
| IHS | Integrated Heat Spreader |
| PCM | Phase Change Material |
| RDHx | Rear Door Heat Exchanger |
| TIM | Thermal Interface Material |
| TPMS | Triply Periodic Minimal Surfaces |
| TPU | Tensor Processing Unit (TPU) |

## Introduction

Chip power density in AI accelerators has risen sharply over the past several generations, and thermal management now limits sustained throughput, device reliability, and rack-level scalability.[1,2] Current-generation devices such as the NVIDIA® H100 SXM at 700 W and the AMD® Instinct MI300X at 750 W already strain conventional cooling, and roadmap projections indicate continued upward pressure.[3,4] Cooling can no longer be treated as a system-level add-on; it shapes chip floor planning, package architecture, memory integration, and Data Center deployment.[5,6]

From a control perspective, thermal management is also a runtime allocation problem. Hotspot location shifts with workload phase, available cooling capacity is finite, and inefficient spatial distribution of that capacity raises junction temperature and accelerates reliability degradation.[6] Addressing the problem therefore requires coordinated design across materials, packaging, fluid delivery, and control systems.

This paper separates the work into three evidence levels. First, it surveys the current landscape of AI-chip cooling, including air cooling, DTC (Direct-To-Chip) liquid cooling, immersion cooling, and embedded microfluidics. Second, it presents original steady-state ANSYS® Mechanical simulations of a five-layer, 700 W chip stack, using energy balance, mesh convergence, and analytical sanity checks to verify numerical consistency while recognizing the absence of experimental calibration. Third, it evaluates passive geometry sweeps and a hotspot-aware zoned convection concept under controlled boundary conditions, producing a 14.75 °C peak-temperature spread between the worst and lowest-temperature passive configurations and up to 4.01 °C additional reduction from area-neutral zoned cooling. Finally, the GNN-GRU cyber-physical controller is included as a proposed future framework for adaptive cooling control; it is not claimed to be trained, benchmarked, or experimentally validated in this study.

## Industry trends in AI-chip thermal management

The trajectory of Data Center cooling follows a consistent pattern: each successive technology moves the point of heat removal closer to silicon. Here, Figure 1 sketches the Chip-to-Facility thermal path for liquid-cooled AI accelerators, and Figure 2 summarizes the representative rack-level power densities for major cooling strategies.

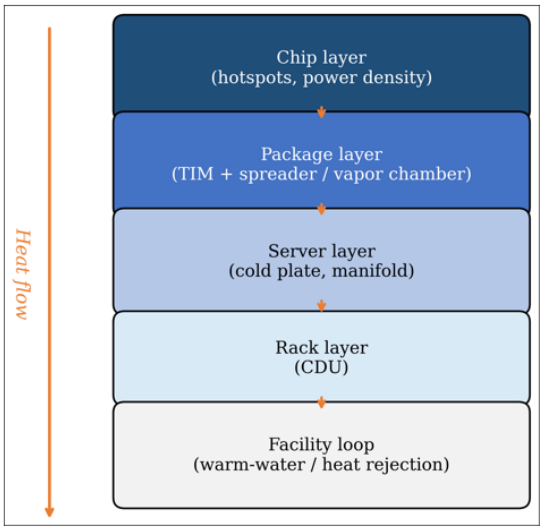


**Figure 1** Chip-to-Facility thermal path for a liquid-cooled AI accelerator rack. Heat generated at the chip layer is transferred through the package, server cold plate and manifold, rack-level Coolant Distribution Unit (CDU), and facility heat-rejection loop. Spreader refers to Integrated Heat Spreader (IHS).

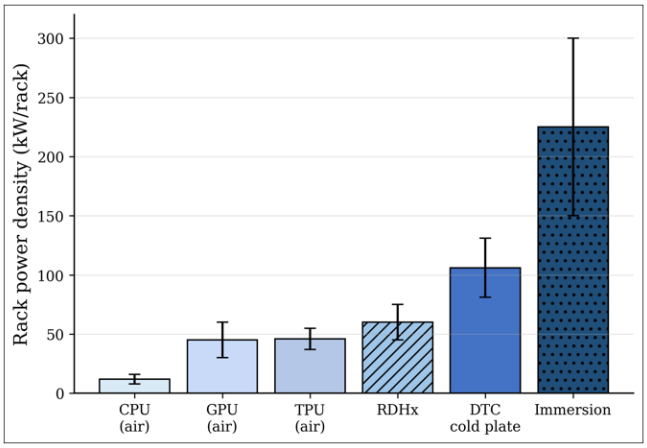


**Figure 2** Representative rack-level thermal density by cooling strategy. Error bars indicate approximate deployment ranges. Graphics Processing Unit (GPU) and Tensor Processing Unit (TPU) racks reflect air-cooled deployments, while immersion cooling reaches the highest rack-scale capacity.

**Citation:** Dwivedi V, Mani BS, Ravindra NM. Industry trends, simulation-guided optimization, and hotspot-aware zoned cooling for high-power Artificial Intelligence (AI) chips. *Material Sci & Eng Int J.* 2026;10(2):69–78. DOI: 10.15406/mseij.2026.10.00270

Air cooling remains the simplest infrastructure option but is fundamentally limited by the low volumetric heat capacity of air. Practical deployments remain below roughly 15 kW/rack for Central Processing Unit (CPU) servers and below roughly 60 kW/rack for aggressively air-cooled GPU or TPU systems.[2]

DTC liquid cooling via manifolded cold plates has emerged as the dominant production approach for the current high-power AI accelerators. The H100 SXM ships with liquid-first thermal assumptions, and the MI300X follows the same paradigm under the Open Accelerator Module (OAM) form factor.[3,4,7] The concept is attributed to Tuckerman and Pease's microchannel heat-sink work.[8] and now supports rack densities above 100 kW/rack in high-density deployments.[5]

Immersion cooling eliminates much of the server-level air path by submerging servers in dielectric fluid and extends rack-scale thermal capacity beyond conventional air and cold plate approaches.[9] This progression is summarized in Figure 3, which shows the movement from air cooling to DTC liquid cooling, immersion cooling, and embedded microfluidics as both cooling capability and infrastructure complexity increase. Figure 4 illustrates the immersion cooling concept. Embedded microfluidics move the coolant path closer still, integrating microchannels within or immediately beneath silicon so that heat can be removed near the active transistor layer.[10,11] As shown in Figure 5, the design features a representative embedded-channel configuration.

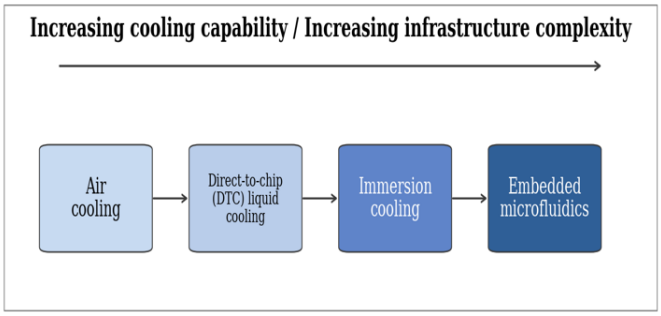


**Figure 3** Industry progression of AI-chip cooling technologies. Each step moves heat removal closer to the junction, while increasing infrastructure complexity and integration difficulty.

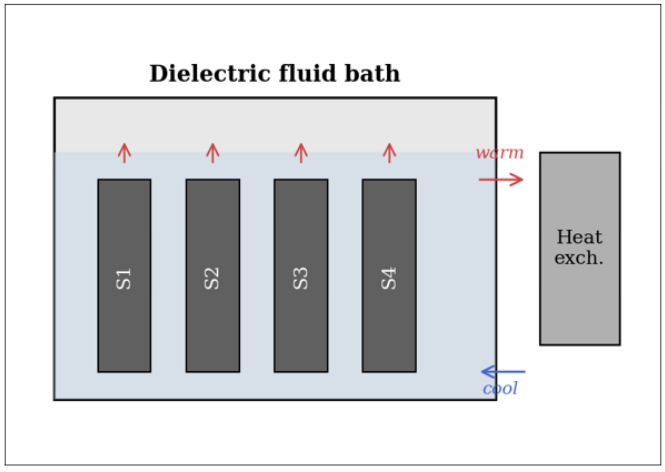


**Figure 4** Single-phase immersion cooling. Servers (S1, S2, S3, S4) are submerged in a dielectric bath, and an external heat exchanger rejects absorbed heat from the warmed fluid.

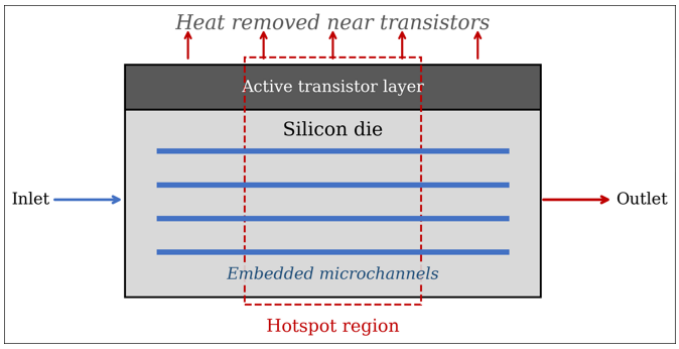


**Figure 5** Embedded microchannel cooling. Channels fabricated within or near the silicon substrate provide direct coolant access near the active transistor layer and preferentially remove heat from high-power regions.

This progression frames the simulation study that follows. Because DTC liquid cooling is the dominant near-term production baseline for high-power AI accelerators, the model in Sections 3 to 5 targets a DTC-style package and cold plate regime. The microfluidic and cyber-physical concepts are treated as possible future extensions rather than validated product designs.

## Model development and validation

### Representative die footprint and stack construction

The modeled die footprint was sized to match the current reticle-class AI logic dies. NVIDIA® reports 814 mm² for the GH100.[7] and 826 mm² for the GA100,[12] while current lithography fields impose a maximum monolithic area of approximately 858 mm².[13] A 28.50 mm × 28.50 mm die, corresponding to 812.25 mm², was adopted as the representative footprint, as shown in Figure 6.

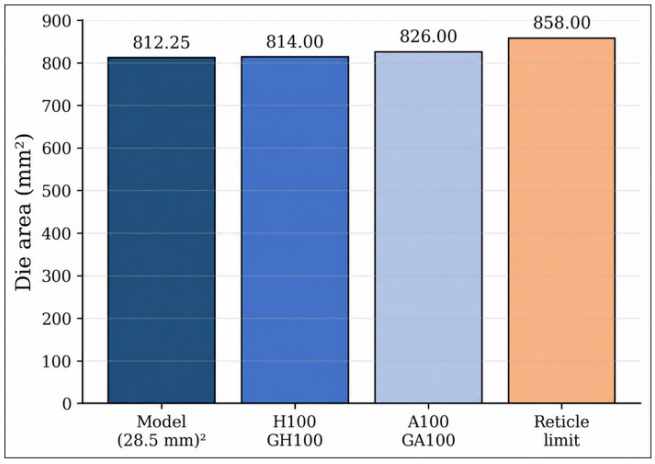


**Figure 6** Die-Area Comparison The modeled footprint of 812.25 mm² is close to current reticle-class accelerator dies and remains below the approximate reticle limit.

For a 700 W package power applied over the 812.25 mm² die, the average die heat flux is calculated by Eq. (1).

$$q'' = \frac{P}{A_{die}} = \frac{700\ \text{W}}{8.12\ \text{cm}^2} = 86.13\ \text{W/cm}^2 \quad (1)$$

where q″ is the average die heat flux (W/cm²), read as "q double prime," and denotes heat input per unit area rather than a derivative, P is total heat input (W), and $A_{die}$ denotes die area (cm²).

**Citation:** Dwivedi V, Mani BS, Ravindra NM. Industry trends, simulation-guided optimization, and hotspot-aware zoned cooling for high-power Artificial Intelligence (AI) chips. *Material Sci & Eng Int J.* 2026;10(2):69–78. DOI: 10.15406/mseij.2026.10.00270

This heat flux is representative of a high-power accelerator package and is sufficiently large to make package-level thermal resistance a first-order design concern.

The simulated stack comprises five solid layers: die, TIM 1, IHS, TIM 2, and cold plate. The die top surface was partitioned into a 4 × 4 array of 16 equal tiles, each 7.13 mm × 7.13 mm, enabling spatially non-uniform power maps. Table 1 lists the baseline geometry and material definitions, and Figure 7 shows the baseline stack and layer temperature drops.

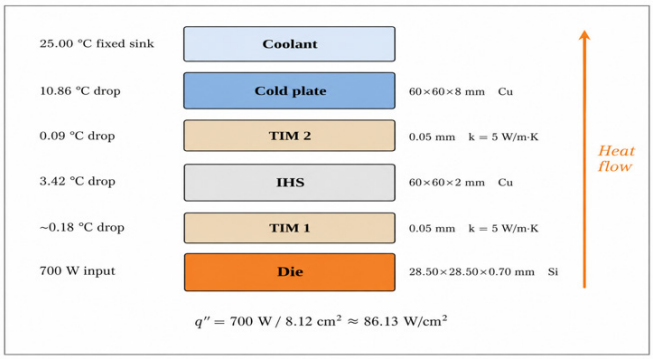


**Figure 7** Five-layer chip stack cross-section with baseline per-layer temperature drops. The cold plate shows the largest through-thickness temperature drop under the fixed-temperature sink model.

**Table 1** Baseline model geometry and materials

| Layer | Geometry | Material |
|---|---|---|
| Die | 28.50 × 28.50 × 0.70 mm | Silicon |
| TIM 1 | 28.50 × 28.50 × 0.05 mm | Generic isotropic TIM, k = 5 W/m·K |
| IHS | 60 × 60 × 2 mm | Copper (C10100) |
| TIM 2 | 60 × 60 × 0.05 mm | Generic isotropic TIM, k = 5 W/m·K |
| Cold plate | 60 × 60 × 8 mm | Copper (C10100) |

## Boundary conditions and validation

Two boundary-condition regimes were used. The passive conduction baseline fixed the cold plate top surface at 25.00 °C and treated all remaining external faces as adiabatic, isolating package-level conduction and spreading from coolant-loop effects. This is an idealized boundary condition and should not be interpreted as a full DTC liquid-loop model. A separate convection regime applied spatially varying heat-transfer coefficients to the cold plate surface to evaluate hotspot-aware zoned cooling in Section 5.

**Implementation details for reproducibility:** The model was solved as a steady-state thermal analysis in ANSYS® Mechanical. The die, TIM 1, IHS, TIM 2, and cold plate were treated as bonded solid bodies with ideal thermal continuity at interfaces; explicit contact resistance, radiation, and coolant-side conjugate fluid flow were not included. Silicon and C10100 copper used the ANSYS® Engineering Data material definitions, while TIM 1 and TIM 2 were modeled as isotropic solids with prescribed thermal conductivity k = 5 W/m·K in the baseline case.

Heat loading was applied as spatially non-uniform die-tile thermal loading over the 4 × 4 tile layout. For the 700 W baseline power map, the four center tiles carried 84 W per tile, the eight middle tiles carried 42 W per tile, and the four outer/corner tiles carried 7 W per tile. The passive conduction runs fixed the cold-plate top surface at 25.00 °C and imposed adiabatic conditions on the remaining external faces.

ANSYS® Mechanical generated a program-controlled three-dimensional thermal-solid mesh, with local refinement applied to the die and TIM regions. The refinement sequence used 897, 1,589, and 4,117 total elements, with local die/TIM element sizing reduced to approximately 1.07 mm in the final refinement. The comparison criterion was the change in die $T_{max}$ between refinements; the final change of approximately 0.07 °C was small relative to the 3.00 °C to 15.00 °C design differences discussed later.

The baseline power map totaled 700 W: 336 W distributed across the center four tiles, 336 W across the mid-eight tiles, and 28 W across the four corner tiles. The reaction at the 25.00 °C boundary was approximately negative 700 W, confirming energy balance with the applied heat input; the sign reflects ANSYS® convention for heat leaving the model.

Mesh convergence was assessed by successive local refinement of the mesh in the die and TIM regions. The peak die temperatures were 54.38 °C with 897 elements, 53.34 °C with 1,589 elements, and 53.41 °C with 4,117 elements. The final refinement changed the peak die temperature by 0.07 °C, confirming that the refined mesh is adequate for comparative parameter sweeps. Figure 8 shows the convergence trend.

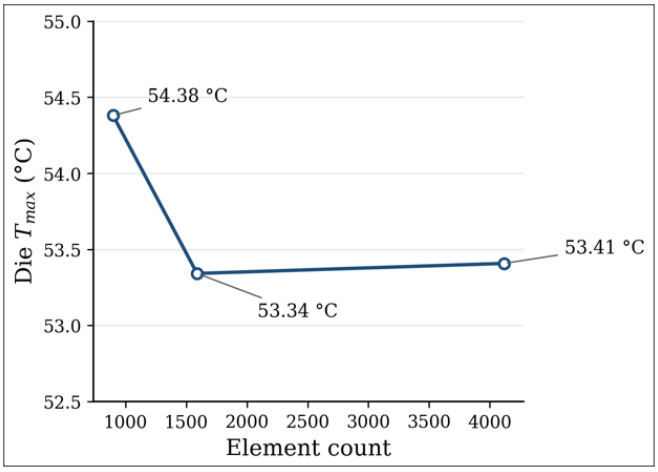


**Figure 8** Temperature convergence with mesh refinement. The peak die temperature stabilizes between meshes with 1,589 and 4,117 elements, with a residual change of approximately 0.07 °C.

**Validation scope:** Energy balance and mesh independence verify numerical consistency of the implemented model, but they do not by themselves validate the physical cold-plate or TIM response against experimental data. The simulations should therefore be interpreted as comparative thermal design screening under the stated assumptions, not as calibrated product-level prediction.

## Baseline results and bottleneck analysis

The refined baseline model with 4,117 elements and the fixed-temperature sink produced a maximum die temperature of 53.41 °C, a minimum die temperature of 28.31 °C, and an area-averaged die temperature of 40.41 °C.

To attribute the temperature rise between individual layers, the die footprint was projected onto the IHS, TIM 2, and cold plate interfaces, and each temperature drop was evaluated over that same projected area. Full-face averaging would mix a 28.50 mm × 28.50 mm die footprint with 60 mm × 60 mm package surfaces. This would combine lateral heat spreading with through-thickness conduction, allowing

cooler regions over the larger package area to dilute the interface averages and produce misleading bottleneck rankings.

Figure 9 shows the matched-footprint temperature-drop results. In this extraction, the approximate TIM 1 drop is 0.18 °C, TIM 2 drop is 0.09 °C, IHS drop is 3.42 °C, and cold plate drop is 10.86 °C. The cold plate and IHS therefore dominate the measured package-level temperature drop within the implemented fixed-sink extraction procedure. Because the model uses ideal bonded interfaces and an idealized fixed-temperature sink, this bottleneck ranking is interpreted as a result of the present simulation setup rather than as a general statement that TIM resistance is always negligible in high-power packages.

**Analytical sanity check:** For comparison, a one-dimensional Fourier estimate for a 0.05 mm TIM layer carrying 700 W through k = 5 W/m·K over the 812.25 mm² die area gives $\Delta T \approx P \cdot t/(k \cdot A) \approx 8.6$ °C. This estimate is substantially larger than the extracted ANSYS® TIM drop; so the TIM-related outputs are reported cautiously and should be re-examined with explicit contact/interface definitions or experimental data before being used as product-level TIM conclusions.

## Parameter sweeps and passive optimization

### TIM sensitivity

TIM sensitivity refers here to the change in die temperature caused by changing TIM thermal conductivity or TIM thickness while holding geometry, power map, and boundary conditions fixed. TIM conductivity was swept over 2, 10, and 20 W/m·K, spanning lower-conductivity polymer or ceramic-filled TIMs through aggressive high-conductivity composite candidates. TIM 1 thickness was also varied from 0.02 mm to 0.10 mm with k fixed at 5 W/m·K.

Across the conductivity sweep, the maximum die temperature changed by only 0.01 °C. Across the TIM 1 thickness sweep, the maximum die temperature changed by less than 0.01 °C. These millidegree-scale changes are far smaller than the IHS and cold plate contributions in Figure 9 within the implemented fixed-sink simulation. Thus, the present parametric sweep shows low sensitivity of this specific model to the tested TIM changes, but it should not be interpreted as proof that TIM thermal resistance is negligible in a real package. This conclusion should not be generalized to embedded microfluidic, contact-resistance-dominated, or very-low-resistance cooling regimes, where the remaining thermal budget may be small enough for TIM properties and interface resistance to become important.

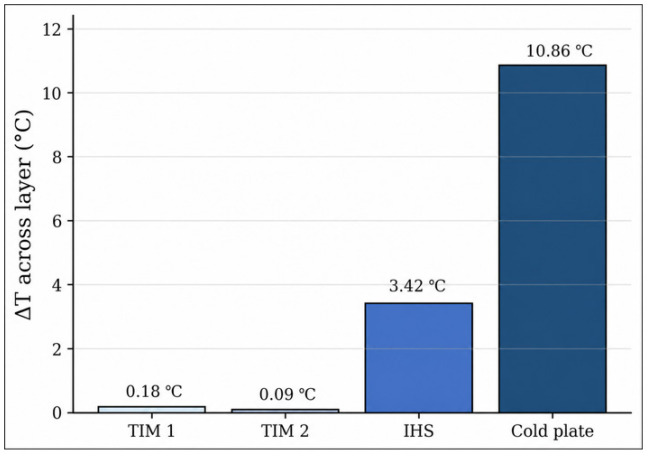


**Figure 9** Matched-footprint temperature drops across package layers. The cold plate and IHS dominate the extracted package-level temperature drop in the current fixed-sink ANSYS® model, while the extracted TIM layers contribute less than 0.30 °C combined. These values are model-extraction results under idealized bonded-contact and fixed-temperature assumptions, not a general validation of negligible TIM resistance.

### IHS thickness sweep

Reducing IHS thickness from 3 mm to 1 mm decreased peak die temperature from 54.71 °C to 52.09 °C, as shown in Figure 10. Under the fixed-temperature sink condition, the IHS acts primarily as a vertical conduction path from the die footprint toward an idealized isothermal boundary; shortening that path reduces through-thickness temperature drop.

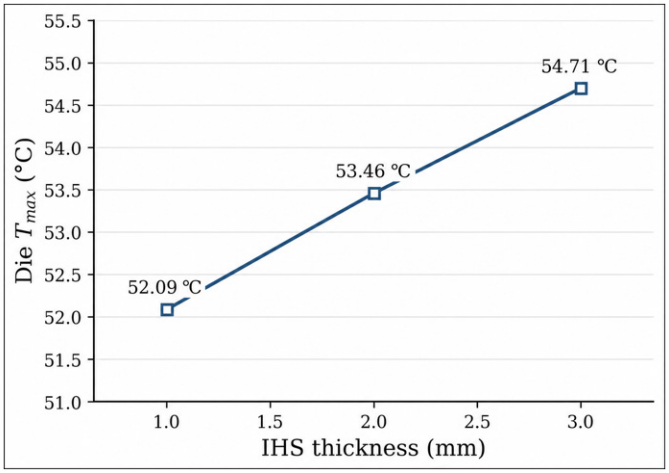


**Figure 10** IHS thickness sweep. Peak die temperature decreases as IHS thickness is reduced from 3 mm to 1 mm under the fixed-temperature sink boundary condition.

This trend should be interpreted only within the fixed-temperature sink model. In a convection-limited physical cold plate, the IHS also provides lateral spreading that can reduce local heat flux at the cooler interface. A thicker IHS may therefore be beneficial when the dominant resistance shifts from bulk conduction to the IHS-to-cooler boundary. The present result shows what occurs under the idealized fixed-temperature sink, not a universal rule that thinner heat spreaders are always superior.

A manufactured DTC cold plate would require a full thermal-resistance network, conjugate heat-transfer simulation, or experimental benchmark before this IHS-thickness trend could be translated into a design recommendation.

### Cold plate thickness sweep

Changing cold plate thickness produced the largest passive effect. As shown in **Figure 11**, reducing cold plate thickness from 12 mm to 4 mm decreased peak die temperature from 57.69 °C to 46.19 °C, and the center-four-tile average decreased from 53.47 °C to 43.19 °C. This response is consistent with the bottleneck analysis: the cold plate carried the largest measured temperature drop in the baseline model.

This cold-plate thickness trend is also specific to the fixed-temperature top-surface boundary. In a physical cold plate, reducing thickness would change channel geometry, coolant-side surface area, pressure drop, flow distribution, and lateral spreading; conjugate heat-transfer modeling is needed before translating this trend into a hardware design rule.

**Citation:** Dwivedi V, Mani BS, Ravindra NM. Industry trends, simulation-guided optimization, and hotspot-aware zoned cooling for high-power Artificial Intelligence (AI) chips. *Material Sci & Eng Int J.* 2026;10(2):69–78. DOI: 10.15406/mseij.2026.10.00270

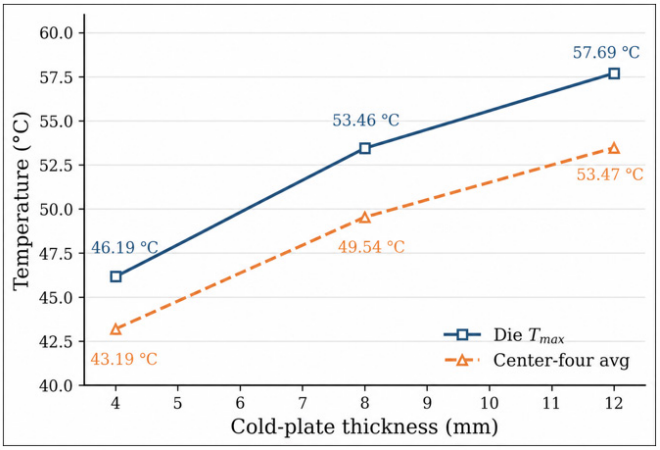


**Figure 11** Cold plate thickness sweep. Both the die peak temperature and center-four-tile average temperature decrease substantially as cold plate thickness is reduced under the fixed-temperature sink model.

### Optimized passive baseline

Combining the lowest-temperature values from the fixed-sink IHS and cold plate sweeps, 1 mm IHS and 4 mm cold plate, produced a peak die temperature of 43.75 °C and a center-hotspot average of 40.99 °C. This represents the best/optimized passive combination of parameters. 3 mm IHS and 12 mm cold plate, reached 58.50 °C, representing the worst passive combination. The resulting 14.75 °C spread demonstrates that passive geometry materially affects the thermal level in the modeled package, but only within the fixed-temperature conduction-screening framework used here. **Table 2** consolidates the fixed-temperature passive simulations.

The optimized passive geometry was adopted as the baseline for the convection-based hotspot-aware zoned cooling study.

**Table 2** Fixed-temperature conduction baseline and passive geometry sweep results

| Configuration | IHS thickness (mm) | Cold plate thickness (mm) | Boundary condition | Die $T_{max}$ (°C) | Hotspot average (°C) |
|---|---|---|---|---|---|
| Refined baseline | 2.00 | 8.00 | Fixed T = 25.00 °C | 53.41 | 40.41* |
| IHS sweep: thin | 1.00 | 8.00 | Fixed T = 25.00 °C | 52.09 | — |
| IHS sweep: thick | 3.00 | 8.00 | Fixed T = 25.00 °C | 54.71 | — |
| Cold plate sweep: thin | 2.00 | 4.00 | Fixed T = 25.00 °C | 46.19 | 43.19 |
| Cold plate sweep: thick | 2.00 | 12.00 | Fixed T = 25.00 °C | 57.69 | 53.47 |
| Optimized passive | 1.00 | 4.00 | Fixed T = 25.00 °C | 43.75 | 40.99 |
| Worst passive | 3.00 | 12.00 | Fixed T = 25.00 °C | 58.50 | 54.23 |

* For the refined baseline, the value is the full-die area average rather than the center-hotspot average. All configurations use 700 W total input and a 28.50 mm × 28.50 mm die.

## Hotspot-aware zoned cooling

### Zoned convection concept

Passive geometry optimization reduces the overall temperature level but does not address the spatial non-uniformity of real workloads. To exploit this non-uniformity, a zoned cooling concept was evaluated in which the convection coefficient on the cold plate surface varies spatially to match the underlying hotspot pattern. This study treats zoning as a thermodynamic redistribution problem, not as a fully resolved fluid-flow design.

For the center-hotspot workload, the cold plate top surface was partitioned into zones aligned with the 4 × 4 die tile layout: the center four tiles used the convective heat-transfer coefficient h = 9,000 W/m²·K, the mid-eight tiles used h = 5,000 W/m²·K, the outer four tiles used h = 1,000 W/m²·K, and the remaining cold plate area used h = 5,000 W/m²·K. The uniform reference value was h = 5,000 W/m²·K. These values were selected to test spatial redistribution while preserving the same area-weighted mean cooling budget over the die-aligned zones. The reference value provides a controlled comparison; it is not intended to represent the upper limit of production DTC liquid cooling loops, where local coefficients of 10,000 to 50,000 W/m²·K may be achievable depending on channel geometry and flow rate.[8,10]

Figure 12 visualizes the spatial redistribution strategy used in the zoned cooling study. Rather than increasing the overall cooling budget, the approach reallocates local cooling intensity toward predicted hotspot regions while reducing cooling effort in lower-power regions. In the center-hotspot case, the largest convection coefficients are concentrated near the middle of the die, whereas the corner-hotspot case shifts the strongest cooling region toward the active corner. The darker regions therefore represent preferential thermal extraction zones aligned with the dominant power-generation regions of the workload. This allows the study to isolate whether spatially targeted cooling alone can reduce hotspot temperature under a fixed area-averaged cooling constraint.

The zone dimensions were chosen to match the 4 × 4 thermal-tile discretization, so each die-aligned zone corresponds to an approximately 7.13 mm × 7.13 mm region. The 25.00 °C ambient or coolant-reference temperature was retained for controlled comparison across uniform and zoned cases. In this study, h is a boundary-condition surrogate for local cooling intensity; it is not computed from channel Reynolds number, flow rate, manifold pressure drop, or detailed coolant temperature rise.

### Area-neutral cooling budget

Any claimed improvement from zoning is meaningful only if the total cooling capacity is held constant. The improvement must arise from spatial redistribution, not from adding net cooling effort. Because all 16 die-aligned tiles share equal area, the area-weighted mean heat-transfer coefficient is given by Eq. (2).

**Citation:** Dwivedi V, Mani BS, Ravindra NM. Industry trends, simulation-guided optimization, and hotspot-aware zoned cooling for high-power Artificial Intelligence (AI) chips. *Material Sci & Eng Int J.* 2026;10(2):69–78. DOI: 10.15406/mseij.2026.10.00270

$$h = \frac{\sum_z h_z \cdot A_z}{\sum_z A_z} \quad (2)$$

$$= \frac{4(9000) + 8(5000) + 4(1000)}{16}$$

$$= 5000 W / m^2 \cdot K$$

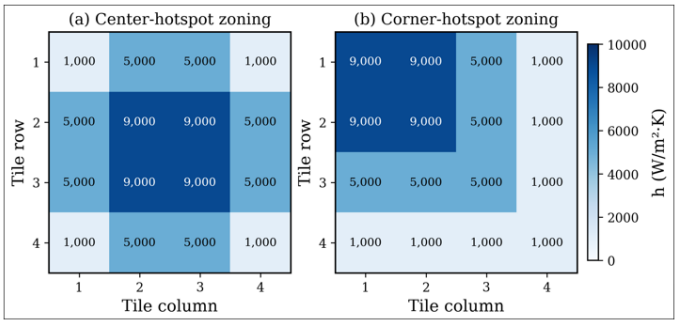


**Figure 12** Convection-coefficient maps for (a) center-hotspot and (b) corner-hotspot zoning. Darker zones receive a larger local heat-transfer coefficient h. The area-weighted mean heat-transfer coefficient over the 16 die-aligned zones remains 5,000 W/m² ·K in both cases.

where $\bar{h}$ is the area-weighted mean heat-transfer coefficient (W/m²·K), $h_z$ is the coefficient assigned to zone z, and $A_z$ is the corresponding zone area (m²). For equal-area tiles, $A_z$ cancels from numerator and denominator.

Thus, the zoned and uniform cases have the same area-averaged cooling budget over the die footprint. The simulation comparison is area-neutral: zoning reallocates cooling intensity without increasing the mean heat-transfer coefficient.

## Zoning results

Under the center-hotspot power map, uniform convection produced a peak die temperature of 109.13 °C and a center-hotspot average of 102.52 °C. With zoned convection, these values decreased to 105.22 °C and 99.49 °C, corresponding to reductions of 3.91 °C and 3.03 °C, respectively.

To test whether the effect depends on the hotspot location, the same total 700 W input was shifted to a corner-hotspot layout, and the high-h zone was repositioned accordingly. For this corner-hotspot case, zoning decreased the peak die temperature from 107.13 °C to 103.12 °C and decreased hotspot average temperature from 99.44 °C to 95.75 °C. These reductions were 4.01 °C and 3.69 °C, respectively. Figure 13 and Table 3 summarize the comparison.

**Table 3** Convection-based hotspot-aware zoning results

| Workload | Cooling case | IHS thickness (mm) | Cold plate thickness (mm) | Die $T_{max}$ (°C) | Hotspot average (°C) |
|---|---|---|---|---|---|
| Center-hotspot | Uniform h = 5,000 W/m² ·K | 1.00 | 4.00 | 109.13 | 102.52 |
| Center-hotspot | Zoned h = 9,000/5,000/1,000 W/m² ·K | 1.00 | 4.00 | 105.22 | 99.49 |
| Corner-hotspot | Uniform h = 5,000 W/m² ·K | 1.00 | 4.00 | 107.13 | 99.44 |
| Corner-hotspot | Zoned, corner-targeted | 1.00 | 4.00 | 103.12 | 95.75 |

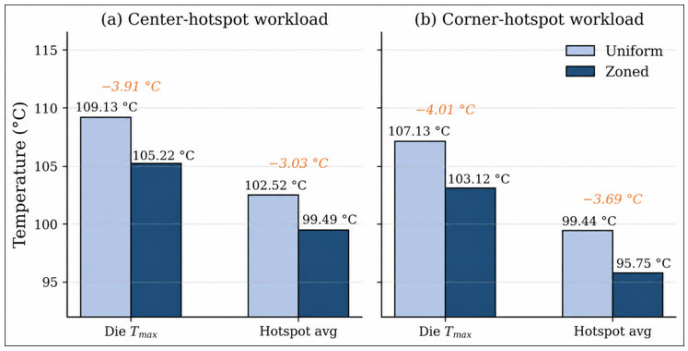


**Figure 13** Uniform vs zoned convection for (a) center-hotspot and (b) corner-hotspot workloads. The area-averaged h is identical in all cases, so the temperature reductions arise from spatial redistribution of cooling effort.

All convection cases use 700 W total input, the optimized passive geometry, and ambient temperature of 25.00 °C.

The consistency in the improvement across center-hotspot and corner-hotspot layouts indicates that spatially targeted cooling can reduce peak and local hotspot temperatures by redistributing cooling capacity toward the active hotspot rather than increasing the overall cooling strength. Physical realization could take the form of manifolded cold plates with variable flow distribution, Micro-Electro-Mechanical-Systems (MEMS)-scale flow-control valves, or actively managed microchannel networks.

**Practical feasibility requires additional modeling:** In a manifolded cold plate, local heat-transfer coefficients cannot usually be changed independently without coupled effects on pressure drop, pump power, channel flow maldistribution, and neighboring zones. Variable-flow manifolds, MEMS-scale valves, or actively managed microchannels could approximate the imposed zoning pattern, but a product-level design would need valve-authority estimates, pump-performance curves, coolant-temperature rise, and transient controller latency before the present boundary-condition map could be implemented.

## Proposed cyber-physical hotspot prediction and control framework

The zoned cooling results motivate a software-driven control layer above the thermal hardware. The simulations show that hotspot location affects die temperature and that spatial redistribution recovers approximately 3.00 °C to 4.00 °C without increasing the area-averaged heat-transfer coefficient. A controller that observes thermal telemetry and reallocates cooling zones could therefore adapt to changing workloads. The architecture in this section is a proposed future implementation path only; no GNN-GRU model was trained, no controller was benchmarked, and no hardware-in-the-loop validation was performed in the present work.

**Citation:** Dwivedi V, Mani BS, Ravindra NM. Industry trends, simulation-guided optimization, and hotspot-aware zoned cooling for high-power Artificial Intelligence (AI) chips. *Material Sci & Eng Int J.* 2026;10(2):69–78. DOI: 10.15406/mseij.2026.10.00270

## Graph representation of the die

The 4 × 4 tile array maps naturally onto an undirected grid graph G = (V, E), where V contains 16 tile nodes and E connects nearest-neighbor tiles. This representation encodes the locality of solid-state heat diffusion: the temperature of a tile depends strongly on its own power dissipation and the state of adjacent tiles. At each control period t, tile i carries the feature vector in Eq. (3).

The equations in this section should therefore be read as a mathematical control formulation rather than as a validated algorithm. They define the information that a future controller could use and the constraints it should obey, but they do not supply trained weights, prediction-error statistics, real-time latency measurements, or closed-loop stability results.

$$x_{i,t} = [p_{i,t},\ T_{i,t},\ \bar{T}\mathcal{N}(i)_t,\ a_{i,t},\ m_{i,t},\ \Delta T_{i,t}] \quad (3)$$

where $x_{i,t}$ is the tile feature vector, $p_{i,t}$ is tile power in watts (W) or a normalized power proxy, $T_{i,t}$ is local tile temperature (°C), $\bar{T}\mathcal{N}(i)_t$ is mean neighbor temperature (°C), $a_{i,t}$ is a compute-activity proxy (dimensionless or normalized counter), $m_{i,t}$ is a memory-traffic proxy (dimensionless or normalized counter), and $\Delta T_{i,t}$ is the recent temperature change (°C).

In practical terms, Eq. (3) defines the information available to the controller for each tile. Each tile is described not only by its own temperature and power, but also by the nearby thermal context and recent temperature change, so the model can distinguish a stable warm region from a newly forming hotspot.

## Spatiotemporal predictor

A GNN message-passing stage can encode lateral thermal coupling between adjacent tiles. A compact form of the update is shown in Eq. (4).

$$z_{i,t}^{(\ell+1)} = \sigma\left( W_s z_{i,t}^{(\ell)} + \sum_{j \in N(i)} W_s z_{i,t}^{(\ell)} + b \right) \quad (4)$$

where $z_{i,t}^{(\ell)}$ and $z_{i,t}^{(\ell+1)}$ are dimensionless hidden representations of tile i at GNN layers $\ell$ and $\ell$+1, respectively; $W_s$ and $W_n$ are learned self and neighbor weight matrices (dimensionless for normalized hidden states); b is a learned bias vector (dimensionless); σ is a dimensionless activation function; and $\mathcal{N}(i)$ is the dimensionless set of neighboring tiles adjacent to tile i.

In physical terms, Eq. (4) lets each tile update its internal representation using both its own state and the states of neighboring tiles. This mirrors lateral heat spreading: heat generated in one tile affects adjacent regions, so nearby tiles must be considered together rather than independently.

The GNN output can then be passed to a GRU, which models thermal persistence over consecutive control periods. The forecast equations are summarized in Eqs. (5a) and (5b).

$$s_{i,t+1} = GRU(z_{i,t}, s_{i,t}) \quad (5a)$$

$$\hat{T}_{i,t+1} = W_o s_{i,t+1} + b_o \quad (5b)$$

where $s_{i,t}$ and $s_{i,t+1}$ are dimensionless recurrent hidden states at times t and t+1; $z_{i,t}$ is the dimensionless GNN output for tile i; $\hat{T}_{i,t+1}$ is the predicted next-step tile temperature (°C); $W_o$ is the learned output weight matrix that maps the hidden state to temperature units (°C per normalized hidden-state unit); and $b_o$ is the learned output bias (°C). A GRU is preferred over a Long Short-Term Memory (LSTM) model here because it has fewer gates and parameters, which is useful when the training data are limited or generated from expensive physics simulations.

In practical terms, Eq. (5a) stores how the thermal state evolves over time, while Eq. (5b) converts that stored state into a predicted next-step temperature. This allows the controller to act on where the hotspot is likely to be next, not only on the temperature measured at the current instant.

## Constrained zone allocation controller

Predicted temperature maps drive a supervisory controller that allocates cooling effort $u_t$ across Z cold plate zones. A representative constrained optimization problem, which can be implemented as a quadratic programming (QP)-style supervisory allocation problem, is given by Eq. (6), with physical and budget constraints in Eq (7).

$$u_t^* = \arg\min_u \left[ \max_i \hat{T}_{i,t+1}(u) + \lambda \| u - u_{t-1} \|_2 + \mu C(u) \right] \quad (6)$$

where $u_t^*$ is the optimal zone allocation vector at time t (W/m²·K or normalized actuation units); u is a candidate actuation vector with the same units as $u_t$; $u_{t-1}$ is the previous-period actuation vector (W/m²·K or normalized actuation units); $\hat{T}_{i,t+1}(u)$ is the predicted tile temperature under candidate actuation u (°C); λ is the movement-penalty weight and μ is the actuation-cost weight (dimensionless or normalized so the objective terms are comparable); and C(u) is a pump, valve, or flow-control cost model expressed in normalized cost units.

In practical terms, Eq. (6) asks the controller to choose a cooling allocation that lowers the predicted hottest tile while avoiding unnecessary actuator motion and excessive flow-control cost. The objective therefore balances thermal reduction against control stability and hardware effort.

$$u_{\min} \le u_t \le u_{\max}; \sum_z A_z u_{z,t} = B; u_t - u_{t-1} \le r_{\max} \quad (7)$$

where $u_{min}$ and $u_{max}$ are lower and upper actuator limits (W/m²·K or normalized actuation units); $u_t$ is the actuation vector at time t (W/m²·K or normalized actuation units); $u_{z,t}$ is the actuation assigned to zone z at time t (W/m²·K or normalized actuation units); $A_z$ is the area of zone z (m²); B is the area-weighted cooling budget (W/K or normalized budget units); and $r_{max}$ is the maximum allowed actuation change per control period (W/m²·K per control period or normalized units per control period).

In practical terms, Eq. (7) prevents unrealistic control actions. The first constraint keeps each actuator within its allowable range, the second keeps the total cooling budget fixed, and the third limits how abruptly the cooling allocation can change between control periods.

The equality constraint in Eq. (7) is the control-theoretic analogue of the area-neutral condition verified in Eq. (2). It prevents the controller from lowering temperature by simply increasing the total cooling budget. Instead, the controller must shift cooling effort toward predicted hotspots while respecting actuator limits and cost penalties.

Figure 14 summarizes the proposed closed-loop cyber-physical cooling architecture. Hardware telemetry from the chip package continuously provides spatial temperature, power, and actuator-state information to the prediction layer. The GNN component captures lateral thermal coupling between neighboring die regions, while the

GRU captures temporal thermal evolution as workloads change over time. These predictions are then passed to the constrained optimizer, which computes a physically allowable cooling allocation for the cold plate zones. The resulting actuator response modifies local cooling intensity, and the updated thermal state is measured again through telemetry, completing the feedback loop. Conceptually, this architecture extends the static zoned-cooling simulations into a future adaptive thermal-management framework capable of responding dynamically to migrating AI-chip hotspots.

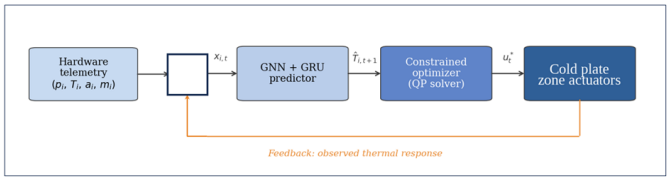


**Figure 14** Proposed cyber-physical control architecture. Hardware telemetry provides the measured system state at time step t, including tile power, temperature, actuator state, and material or operating parameters. The GNN + GRU predictor estimates the next-step temperature field, and the constrained optimizer computes the zone allocation vector for the cold plate actuators. After actuation, the updated thermal response is measured as the next telemetry input, closing the feedback loop. The feedback path is conceptual in this study and does not represent an experimentally identified transfer function.

## Algorithmic control loop

Algorithm 1 summarizes the proposed hotspot-aware cooling control loop. The algorithm is intended as a design formulation rather than a validated controller.

Algorithm 1. Hotspot-aware cooling control.

Require: Tile telemetry $x_{i,t}$, cooling budget B, actuator limits, trained predictor parameters, and controller weights.

Ensure: Zone allocation vector $u_t^*$.

1. Initialize recurrent states $s_{i,0}$ and initial zone allocation $u_0$.

2. At each control period t, assemble feature vectors $x_{i,t}$ for all die tiles.

3. Apply GNN message passing to compute $z_{i,t}$ for all tiles using Eq. (4).

4. Update GRU states and predict $\hat{T}_{i,t+1}$ using Eqs. (5a) and (5b).

5. Solve the constrained optimization problem in Eqs. (6), (7).

6. Apply $u_t^*$ to cold plate zone actuators.

7. Observe the new thermal response and update telemetry for the next control period.

Algorithm 1 is consequently a design specification for future implementation. A complete validation would require a training dataset spanning realistic workload power maps, comparison against simpler thermal predictors, robustness testing under workload distribution shift, and experimental or hardware-in-the-loop confirmation that the actuator commands can be applied within physical pump and valve limits.

# Discussion

The results support a design hierarchy for DTC-cooled AI packages operating in the 700 W class. Under the fixed-temperature boundary condition, the cold plate is the dominant package-level bottleneck, the IHS is secondary, and the TIM layers are comparatively small contributors. Passive geometry optimization lowers the absolute temperature level, while hotspot-aware zoned convection provides an additional 3.00 °C to 4.00 °C reduction by redistributing cooling effort toward high-power regions. This sequence - bottleneck identification, passive geometry optimization, and targeted active zoning - provides a structured template for package-level thermal co-design.

Several limitations require emphasis. The fixed-temperature sink used for the passive sweeps is an idealized boundary condition. It isolates conductive spreading effects, but it does not model coolant heating, flow maldistribution, pressure drop, or cold plate internal channels. Consequently, the finding that thinner IHS and thinner cold plate perform better is valid only under this fixed-temperature sink model. In a real DTC system, the lateral spreading benefit of a thicker IHS or cold plate may offset additional through-thickness conduction resistance when the boundary is convection-limited.

The zoned cooling simulations also simplify the physical hardware. The model applies spatially varying convection coefficients at the cold plate surface rather than resolving conjugate fluid flow through channels and manifolds. The result should therefore be read as a thermodynamic proof of concept: if cooling intensity can be redistributed spatially under a fixed mean budget, peak temperature can be reduced. A product-level design would require flow-channel geometry, pressure-drop accounting, pump-power analysis, manufacturability constraints, and experimental validation on a physical test vehicle.

The predictor-controller pipeline in Section 6 is likewise a proposed architecture, not a demonstrated closed-loop system. Its latency, prediction accuracy, training-data requirements, and robustness under workload distribution shift remain open questions. Training data could be generated by sweeping ANSYS® or conjugate heat-transfer simulations over representative power maps, then fine-tuning on hardware telemetry from on-die thermal sensors. Because the graph contains only 16 nodes, inference could plausibly run on a small, embedded controller, but that claim requires implementation and timing measurements.

A complementary passive direction involves replacing planar cold plate geometries with triply periodic minimal surface (TPMS) lattices such as gyroid, IWP, or primitive structures. Ansari and Duwig reported that a full-scale gyroid heat sink can reduce thermal resistance and temperature non-uniformity relative to conventional pin-fin designs at matched porosity, though with higher pumping-power requirements.[14] The hotspot-aware zoning concept developed here could, in principle, be overlaid onto a TPMS substrate by modulating inlet distribution across subregions of the lattice, combining three-dimensional mixing with spatially targeted flow control.

Phase-Change Materials (PCMs) represent another option for transient thermal buffering. Gado investigated composite PCM-TPMS heat sinks and found that metallic TPMS scaffolds filled with eicosane improve base-temperature control and damp pulsed thermal amplitudes compared with pure PCM systems.[15] A PCM-TPMS layer between the IHS and cold plate could absorb short burst power spikes during inference workloads, but its usefulness would depend on recovery time, cycling reliability, and saturation behavior.

At still smaller length scales, the continuum Fourier heat-conduction model used here would eventually become insufficient. If embedded microfluidic cooling or nanoscale thermal engineering approaches are pushed toward length scales comparable to phonon mean free paths, ballistic phonon transport, electron-phonon coupling, and local thermal non-equilibrium would require more detailed transport models.[16]

**Citation:** Dwivedi V, Mani BS, Ravindra NM. Industry trends, simulation-guided optimization, and hotspot-aware zoned cooling for high-power Artificial Intelligence (AI) chips. *Material Sci & Eng Int J.* 2026;10(2):69–78. DOI: 10.15406/mseij.2026.10.00270

To maintain scope on high-power AI cooling, the discussion is limited to package, cold-plate, and Data-Center relevant mechanisms. The most direct extension of the present study is a conjugate DTC cold-plate model that couples the solid stack to manifold geometry, coolant flow rate, pressure drop, pump power, and non-uniform inlet distribution. That model would allow the zoned-cooling concept to be evaluated as a practical hardware-control problem rather than as an imposed boundary-condition map.

## Conclusions

This study combined an industry survey with a simulation-guided thermal design analysis for a representative 700 W, 86.13 W/cm² AI-chip stack. Three principal findings emerged. First, under the idealized fixed-temperature DTC-style boundary condition, the cold plate and IHS account for most of the extracted package-level temperature decrease, while the extracted TIM drops are small; however, the TIM interpretation should be treated cautiously because a one-dimensional analytical TIM estimate gives a larger temperature drop than the ANSYS® extraction. Second, changing IHS and cold plate thickness between the worst and best passive configurations reduces peak die temperature by 14.75 °C within the fixed-sink screening model. Third, hotspot-aware zoned cooling reduces peak die temperature by up to 4.01 °C across center-hotspot and corner-hotspot workloads by redistributing cooling capacity toward the active hotspot rather than increasing the overall cooling strength.

Broadly, the work frames AI-chip thermal management as a cyber-physical co-design problem in which package geometry, spatially adaptive cooling, workload telemetry, and constrained optimization operate together. These findings should be interpreted as simulation-guided comparative trends, not final hardware design rules. Immediate next steps include conjugate fluid-flow modeling inside the cold plate, pressure-drop and pump-power analysis, experimental validation of zoned cooling on a physical test vehicle, and training and evaluation of the GNN-GRU predictor and zone allocation controller.

## Data Availability

The ANSYS® simulation inputs, boundary-condition specifications, and tabulated results underlying this study were generated as part of the project by Ved Dwivedi using ANSYS® Mechanical resources available through John P. Stevens High School. They are not treated as an external paper or numbered reference. The data are available from the corresponding author upon reasonable request. Inquiries should be directed to the first author at John P. Stevens High School or to the faculty advisors at the New Jersey Institute of Technology.

## Acknowledgments

The authors acknowledge the support of John P. Stevens High School in making the ANSYS® simulation program available for this project. The helpful comments of the Reviewers and the Editors are acknowledged with sincere thanks. Authors are thankful to the Ravindra Family for funding this research study and sponsoring the publication in the Journal.

## Use of AI and machine learning

In preparing this manuscript, the authors used AI-assisted tools to help identify and retrieve relevant references and to support portions of the writing and editing process. All technical content, simulation results, analysis, and conclusions are the responsibility of the authors. No AI tool was used to generate or alter simulation data.

## References

1. International Energy Agency. *Data Centres and Data Transmission Networks*. Paris, France: International Energy Agency; 2023.
2. ASHRAE TC 9.9. *Liquid Cooling Guidelines for Datacom Equipment Centers*. Atlanta, GA: ASHRAE; 2022.
3. NVIDIA. H100 GPU product specifications. Accessed June 8, 2026.
4. AMD. AMD Instinct MI300X accelerators. Accessed June 8, 2026. https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html
5. Uptime Institute. *The reality of liquid cooling*. Uptime Institute Research Report; 2023.
6. Sabry MR, Zohdy MA, Fan D, et al. Thermal modeling, analysis, and management in multicore ICs: a survey. *IEEE Trans Components Packaging Manuf Technol*. 2018;8(6):1019–1036.
7. NVIDIA. NVIDIA Hopper architecture in–depth. NVIDIA Technical Blog. Published 2022.
8. Tuckerman DB, Pease RFW. High–performance heat sinking for VLSI. *IEEE Electron Device Lett*. 1981;2(5):126–129.
9. Zimmermann A, Trefzer C, Gerstenberger M, et al. Single–phase immersion cooling for high–performance computing. *Appl Therm Eng*. 2018;145:196–206.
10. Kandlikar SG, Garimella S, Li D, et al. Heat transfer in microchannels—a critical review. *J Heat Transfer*. 2005;127(8):883–902.
11. Zhang Y, Wang X, Liu J, et al. Embedded microchannel cooling for high–power–density integrated circuits. *IEEE Trans Components Packaging Manuf Technol*. 2020;10(3):455–467.
12. NVIDIA. NVIDIA Ampere architecture in–depth. NVIDIA Technical Blog. Published 2020. Accessed June 8, 2026.
13. WikiChip. Mask/Reticle. Published 2024.
14. Ansari D, Duwig C. A gyroid TPMS heat sink for electronic cooling. *Energy Convers Manag*. 2024;319:118918.
15. Gado MG. Thermal management and heat transfer enhancement of electronic devices using integrative phase change material (PCM) and triply periodic minimal surface (TPMS) heat sinks. *Appl Therm Eng*. 2025;258:124504.
16. Romeo C, Baldi A, Askes SHC. Engineering light–driven thermal landscapes at the nanoscale. *APL Mater*. 2025;13:080601.