Research Article
Open Access
CrossMark

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

## Abstract

High-power artificial-intelligence accelerators require package and cooling designs that address both total heat load and spatial hotspots. A 700 W, 812.25 mm² (86.18 W/cm²) five-layer die–TIM1–IHS–TIM2–cold-plate stack was evaluated in ANSYS Mechanical using idealized fixed-temperature and prescribed-convection boundaries. An initial analytical discrepancy was traced to a 0.21384 mm automatic contact tolerance that exceeded the 0.05 mm TIM layers and created nonphysical Die–IHS and IHS–cold-plate bypass contacts. Suppressing Contact Regions 5 and 3 restored the intended serial path, with four through-thickness divisions in each TIM. The corrected fine-mesh fixed-sink baseline produced $T_{min}$ = 32.221 °C, $T_{avg}$ = 54.664 °C, $T_{max}$ = 78.998 °C, and a −700 W reaction. Increasing both TIM conductivities from 2 to 20 W/(m·K) reduced $T_{max}$ from 113.050 to 60.161 °C, while increasing TIM1 thickness from 0.02 to 0.10 mm raised $T_{max}$ from 70.287 to 95.944 °C. A combined tested configuration using $k$ = 20 W/(m·K), a 3 mm IHS, and a 4 mm cold plate produced $T_{max}$ = 55.059 °C, 30.3% below the corrected baseline. At the same prescribed mean $h$ = 5,000 W/(m²·K), hotspot-aware zoning reduced $T_{max}$ by 3.610 °C for the baseline package and 2.440 °C for the combined package while reducing temperature spread by 7.97% and 8.44%. The fine-to-refined-fine mesh-sensitivity difference in $T_{max}$ was 0.006 °C, although the full grid sequence was non-monotonic. The results constitute numerical screening and analytical benchmarking, not experimental validation; coolant channels, pressure drop, coolant heating, transient behavior, structural stress, and controller implementation were not modeled, and the GNN–GRU framework remains proposed.

**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 |
| $Aj$, $Az$ | Area of tile or cooling zone | m² |
| $ai,t$ | 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) |
| $h$-bar | 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) |
| $\lambda$, $\mu$ | Controller weighting coefficients | dimensionless or normalized cost weights |
| $mi,t$ | Memory-traffic proxy at tile $i$ and time $t$ | dimensionless or normalized counter |
| $N(i)$ | Set of neighboring tiles adjacent to tile $i$ | dimensionless |
| $P$ | Total heat input | W |
| $pi,t$ | 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 |
| $Ti,t$ | Temperature of tile $i$ at time $t$ | °C |
| $Tmax$ | 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 |

Submit Manuscript | http://medcraveonline.com



©2026 Dwivedi et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and build upon your work non-commercially.

Copyright: ©2026 Dwivedi et al.



| Symbol | Definition | Unit |
|---|---|---|
| $u_t$ | Cooling-actuation vector at time $t$ | W/(m² ·K) or normalized actuation units |
| $u_t^*$ | Optimal cooling-actuation vector at time $t$ | same as $u_t$ |
| $x_{i,t}$ | Feature vector for tile $i$ at time $t$ | mixed/normalized |
| $t$ | Layer thickness | m or mm, as specified |
| $T_{min}$ | Minimum die temperature | °C |
| $T_{avg}$ | Full-die area-average temperature | °C |
| Temperature spread | $T_{max}$ - $T_{min}$ | °C |

| Abbreviation | Description |
|---|---|
| AI | Artificial Intelligence |
| CDU | Coolant Distribution Unit |
| DTC | Direct-to-Chip |
| GNN | Graph Neural Network |
| 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 current AI-chip cooling approaches, including air cooling, direct-to-chip (DTC) liquid cooling, immersion cooling, and embedded microfluidics. Second, it presents corrected steady-state ANSYS Mechanical simulations of a five-layer, 700 W chip stack. An analytical inconsistency prompted an audit of the original contact topology, which identified and removed nonphysical contacts that bypassed both TIM layers. Third, corrected fixed-sink screening shows strong TIM-conductivity sensitivity and a 30.3% $T_{max}$ reduction for the combined tested material-and-geometry configuration. Separate prescribed-convection cases show an additional zoned reduction of 2.440–3.610 °C, depending on package configuration. The GNN–GRU cyber-physical controller is retained only as a proposed future framework; it was not trained, benchmarked, or experimentally validated.

## 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.

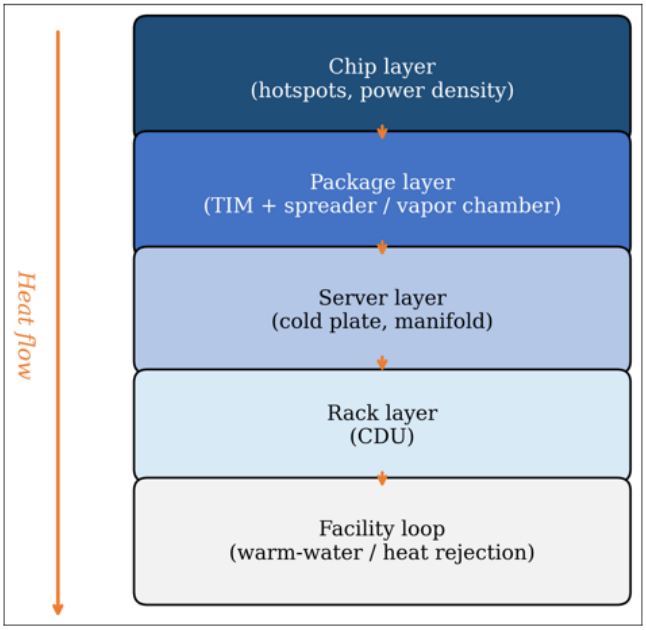


**Figure 1** Chip-to-facility thermal path for a liquid-cooled AI accelerator rack. Heat generated at the chip layer is transferred through the package, server cold plate and manifold, rack-level Coolant Distribution Unit (CDU), and facility heat-rejection loop.

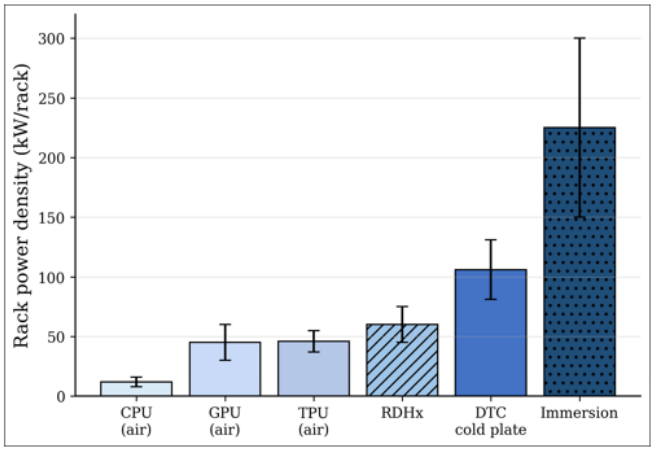


**Figure 2** Representative rack-level thermal density by cooling strategy. Error bars indicate approximate deployment ranges. Graphics Processing Unit (GPU) and Tensor Processing Unit (TPU) racks reflect air-cooled deployments, while immersion cooling reaches the highest rack-scale capacity.

Air cooling remains the simplest infrastructure option but is fundamentally limited by the low volumetric heat capacity of air.

**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):69–79. DOI: 10.15406/mseij.2026.10.00270

Copyright: ©2026 Dwivedi et al.



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.

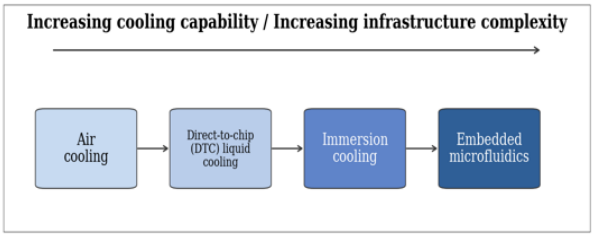


**Figure 3** Industry progression of AI-chip cooling technologies. Each step moves heat removal closer to the junction, while increasing infrastructure complexity and integration difficulty.

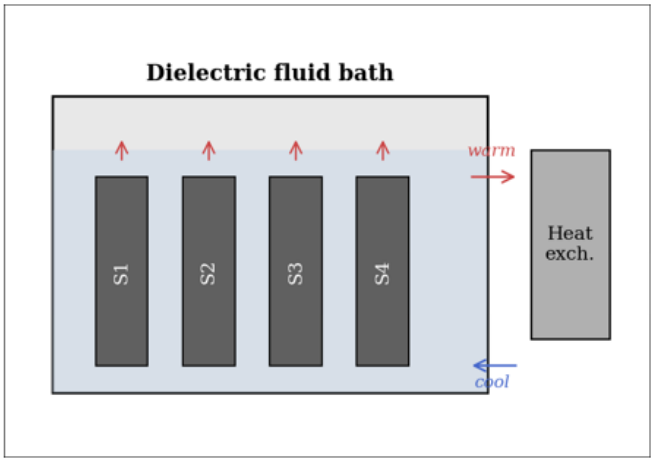


**Figure 4** Single-phase immersion cooling. Servers (S1, S2, S3, S4) are submerged in a dielectric bath, and an external heat exchanger rejects absorbed heat from the warmed fluid.

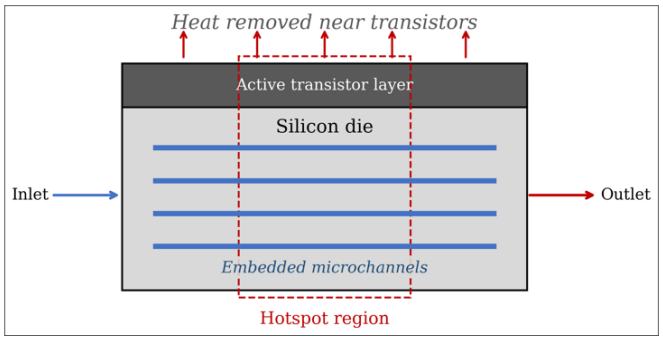


**Figure 5** Embedded microchannel cooling. Channels fabricated within or near the silicon substrate provide direct coolant access near the active transistor layer and preferentially remove heat from high-power regions.

This progression frames the simulation study that follows. Sections 3-5 examine a DTC-style solid-package screening model with either an idealized fixed-temperature sink or prescribed convection at the cold-plate top surface; they do not model a coolant loop, channels, coolant heating, or pressure drop. Embedded microfluidics and cyber-physical control remain future extensions rather than validated product designs.

## Model development, numerical verification, and analytical benchmarking

### 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.

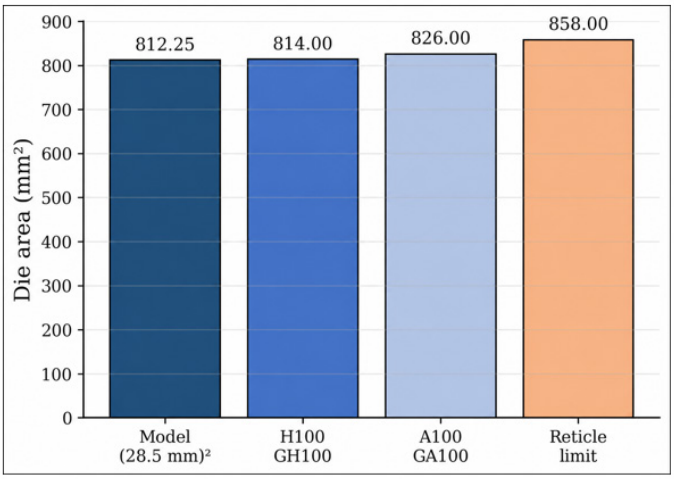


**Figure 6** Die-area comparison. The modeled footprint of 812.25 mm² is close to current reticle-class accelerator dies and remains below the approximate reticle limit.

For a 700 W package power applied over the 812.25 mm² die (8.1225 cm²), the average die heat flux is 86.18 W/cm², as calculated by Eq. (1).

$$q'' = \frac{P}{A_{die}} = \frac{700W}{8.1225cm^2} = 86.18W/cm \qquad (1)$$

**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):69–79. DOI: 10.15406/mseij.2026.10.00270

Copyright: ©2026 Dwivedi et al.



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).

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, TIM1, IHS, TIM2, and cold plate. The die top surface was partitioned into a 4 × 4 array of 16 equal tiles, enabling spatially nonuniform power maps. Table 1 summarizes the baseline geometry and available material definitions, and Figure 7 identifies the original and corrected contact topologies.

**Table 1** Baseline model geometry and materials

| Layer | Dimensions | Material | Thermal conductivity used | Property dependence/source |
|---|---|---|---|---|
| Die | 28.50 × 28.50 × 0.70 mm | Silicon, pure | 154.3 W/(m·K) | Constant; ANSYS Engineering Data reference field at 23 °C |
| TIM1 | 28.50 × 28.50 × 0.05 mm | Generic_TIM | 5 W/(m·K) | Constant prescribed isotropic bulk conductivity |
| IHS | 60 × 60 × 2 mm | Copper, pure, C10100, hard | 396.7 W/(m·K) | Constant; ANSYS Engineering Data reference field at 23 °C |
| TIM2 | 60 × 60 × 0.05 mm | Generic_TIM | 5 W/(m·K) | Constant prescribed isotropic bulk conductivity |
| Cold plate | 60 × 60 × 8 mm | Copper, pure, C10100, hard | 396.7 W/(m·K) | Constant; ANSYS Engineering Data reference field at 23 °C |

*Note: Explicit thermal contact resistance was not modeled. All intended adjacent interfaces were bonded after suppressing the nonphysical nonadjacent auto-contacts.*

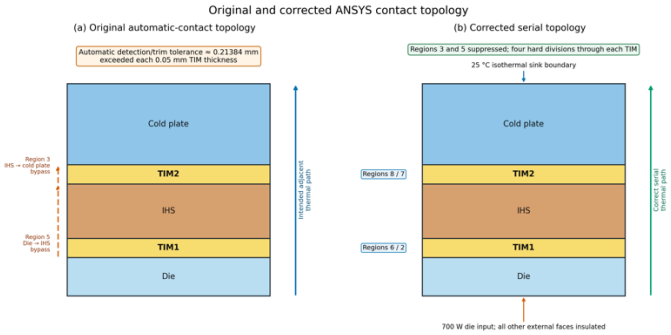


**Figure 7** Original and corrected ANSYS contact topology. The automatic contact tolerance exceeded the TIM thickness and created nonphysical nonadjacent bonded contacts that bypassed TIM1 and TIM2. Suppressing the Die–IHS and IHS–cold-plate bypasses restored the intended serial heat path.

## Model implementation and contact correction

The model was solved as a three-dimensional steady-state thermal analysis in ANSYS Mechanical 2025 R2 Student with Mechanical APDL as the solver target. Silicon and C10100 copper were assigned constant thermal conductivities of 154.3 and 396.7 W/(m·K), respectively, and both Generic_TIM layers used constant isotropic bulk conductivity of 5 W/(m·K) at baseline. Density and specific heat were not defined because steady-state conduction contains no transient thermal-storage term.

The 700 W die power map comprised four hotspot tiles at 84 W each (336 W), eight adjacent tiles at 42 W each (336 W), and four far tiles at 7 W each (28 W). For fixed-sink screening, the cold-plate top surface was prescribed at 25 °C and all remaining external faces were insulated. This is an idealized isothermal sink for conduction screening, not a coolant, DTC loop, or conjugate heat-transfer model.

For the prescribed-convection cases, the fixed-temperature condition was suppressed, the reference ambient temperature was 25 °C, and the cold-plate top used either uniform $h$ = 5,000 W/(m²·K) or the zoned map defined in Section 5. The prescribed coefficients were not derived from Reynolds number, channel geometry, flow rate, coolant heating, or pressure drop.

A contact audit was performed after the originally extracted TIM1 drop conflicted with a one-dimensional Fourier estimate. ANSYS automatic contact generation used a detection/trim tolerance of approximately 0.21384 mm, exceeding each 0.05 mm TIM. Contact Region 5 therefore bonded Die directly to IHS and bypassed TIM1, while Contact Region 3 bonded IHS directly to the cold plate and bypassed TIM2. Both invalid regions were suppressed. Active bonded Contact Regions 6, 2, 8, and 7 restore the intended serial path Die→TIM1→IHS→TIM2→ColdPlate. Initial-contact diagnostics reported all retained interfaces closed. Although the Contact Tool displayed a 0.0500 mm geometric-search gap for Region 7, CAD face coordinates independently confirmed that the TIM2 and cold-plate interface faces were coincident at Y = 2.8 mm, giving an operational gap of zero.

TIM1 and TIM2 were each assigned four hard through-thickness divisions. Four corrected meshes were evaluated, spanning 31,670–92,429 nodes and 11,995–35,668 solid elements; global element size decreased from 8.5538 to 3.9601 mm and local body size from 2.1386 to 0.9901 mm. Jacobian Ratio (Gauss Points) was used as the mesh-quality metric, with positive values and no inverted elements required. The refined-fine mesh had minimum 0.91056, average 0.99911, maximum 1.0, and standard deviation of 0.0070925.

The mesh used SOLID291 and SOLID279 thermal solid elements, CONTA174/TARGE170 bonded contact elements, and SURF152 surface-load elements. One load step was solved; the displayed endpoint of 1 s is a load-step coordinate rather than physical transient time. Automatic time stepping, heat convergence, temperature convergence, line search, and the Mechanical solver type remained Program Controlled. Solver output identified a distributed sparse matrix direct solver using four distributed-memory processes. The refined-fine case contained 91,392 equations and completed one substep and one cumulative iteration with zero solver errors. Corrected mesh-sensitivity and mesh-quality results are summarized in Table 2. The results of corrected mesh-sensitivity study, after removal of nonphysical bypass contacts, are summarized in Figure 8.

**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):69–79. DOI: 10.15406/mseij.2026.10.00270

Copyright: ©2026 Dwivedi et al.



**Table 2** Corrected mesh-sensitivity and mesh-quality results

| Mesh | Nodes | Elements | Global (mm) | Local (mm) | Jac. min | Jac. avg | $T_{min}$ | $T_{avg}$ | $T_{max}$ |
|---|---|---|---|---|---|---|---|---|---|
| Coarse | 31,670 | 11,995 | 8.5538 | 2.1386 | 0.29043 | 0.98223 | 32.624 | 53.933 | 81.359 |
| Medium | 56,202 | 21,399 | 6.0480 | 1.5122 | 0.21370 | 0.98978 | 32.483 | 54.505 | 82.694 |
| Fine | 85,753 | 32,989 | 4.2769 | 1.0693 | 0.56531 | 0.99910 | 32.221 | 54.664 | 78.998 |
| Refined-fine | 92,429 | 35,668 | 3.9601 | 0.9901 | 0.91056 | 0.99911 | 32.851 | 54.760 | 79.004 |

*Note: Temperatures are in °C. Each mesh used four hard divisions through TIM1 and TIM2 and returned a −700 W reaction. The sequence was non-monotonic; therefore, conventional Richardson extrapolation and GCI were not reported. Fine-to-refined-fine mesh-sensitivity changes were 0.006 °C (0.0076%) in $T_{max}$, 0.096 °C in Tavg, and 0.630 °C in $T_{min}$.*

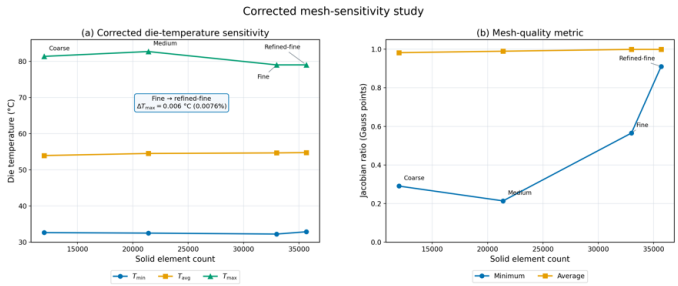


**Figure 8** Corrected mesh-sensitivity study after removal of nonphysical bypass contacts. The four-grid sequence is non-monotonic; the direct fine-to-refined-fine mesh-sensitivity change in die *Tmax* is 0.006 °C, and Jacobian-ratio statistics are reported for each mesh.

Every final corrected steady-state case returned a −700 W reaction at the heat-removal boundary for the 700 W applied load; the negative sign is the ANSYS convention for heat leaving the model. This global balance is necessary but does not establish the layer-by-layer path or experimental validity. Radiation, explicit thermal contact resistance, coolant CFD, internal channel geometry, pressure drop, coolant heating, and transient response were not modeled.

## Corrected baseline, heat-path verification, and analytical TIM benchmark

With corrected contacts, the fine-mesh fixed-sink baseline produced die $T_{min}$ = 32.221 °C, Tavg = 54.664 °C, $T_{max}$ = 78.998 °C, and temperature spread $T_{max} - T_{min}$ = 46.777 °C. The heat-removal-boundary reaction was −700 W.

The former near-zero TIM drops and associated bottleneck ranking were invalid because the nonadjacent Die–IHS and IHS–ColdPlate bonds allowed heat to bypass the 0.05 mm TIM bodies. Suppressing these bypasses changes the baseline temperature and reverses the earlier conclusion that the tested TIM conductivity was negligible.

For an ideal one-dimensional TIM1 path, $\Delta T$ = Pt/(kA), where P = 700 W, A = 812.25 mm², t is TIM1 thickness, and k is bulk conductivity. Table 3 reports the conductivity and thickness benchmarks. Comparable ANSYS face drops were not extracted for the conductivity sweep and are therefore not reported. For TIM1 thicknesses of 0.02, 0.05, and 0.10 mm, subtracting the ANSYS-displayed average temperature of the TIM1 top face from that of the bottom face yielded temperature differences of 3.115, 6.204, and 12.515 °C, respectively. These differences were used only to confirm the analytical thickness trend and were not treated as formal validation metrics because the displayed face averages were not obtained using a rigorously area-integrated paired-face calculation. The remaining gap is attributable to the comparison metric rather than an energy-balance discrepancy: the Fourier calculation is an area-averaged one-dimensional estimate based on the full 700 W heat flow, whereas the ANSYS quantities are Mechanical display averages from a spatially nonuniform three-dimensional temperature field and were not computed as rigorously area-weighted paired-face averages. The two quantities are therefore not mathematically identical, so the displayed ANSYS values are used only for trend checking.

**Table 3** One-dimensional TIM1 analytical benchmarks and qualified displayed thickness-sweep face-average differences

| Sweep | *k* (W/(m·K)) | *t* (mm) | Analytical $\Delta T$ (°C) | Displayed ANSYS $\Delta T$ (°C) | Qualification |
|---|---|---|---|---|---|
| Conductivity | 2 | 0.05 | 21.545 | — | No comparable face-drop extraction |
| Conductivity | 5 | 0.05 | 8.618 | — | No comparable production face-drop extraction |
| Conductivity | 10 | 0.05 | 4.309 | — | No comparable face-drop extraction |
| Conductivity | 20 | 0.05 | 2.155 | — | No comparable face-drop extraction |
| Thickness | 5 | 0.02 | 3.447 | 3.115 | Displayed full-face averages; trend only |
| Thickness | 5 | 0.05 | 8.618 | 6.204 | Refined-fine reference; trend only |
| Thickness | 5 | 0.10 | 17.236 | 12.515 | Displayed full-face averages; trend only |

The −700 W heat-removal-boundary reaction for every 700 W case provides global steady-state energy-balance verification. Direct fine-to-refined-fine mesh sensitivity of 0.006 °C in $T_{max}$ and 0.096 °C in $T_{avg}$ quantifies the final-grid response, while the Fourier calculation supplies an independent analytical benchmark. These checks constitute numerical verification and analytical benchmarking, not experimental validation.

# Corrected parameter sensitivity and package optimization

## TIM conductivity sensitivity

Both TIM layers were assigned the same isotropic bulk conductivity and swept through *k* = 2, 5, 10, and 20 W/(m·K) while geometry, power map, and fixed-sink boundary were held constant.

**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):69–79. DOI: 10.15406/mseij.2026.10.00270

The strong sensitivity is physically consistent with restoration of the serial heat path and replaces the invalid millidegree trend obtained with bypass contacts. The result applies to the tested isotropic bulk conductivities and ideal bonded interfaces; it does not quantify explicit contact resistance, anisotropy, pump power, or product-level TIM behavior. Figure 9 shows the results of the corrected TIM-conductivity sensitivity and one-dimensional TIM1 benchmark.

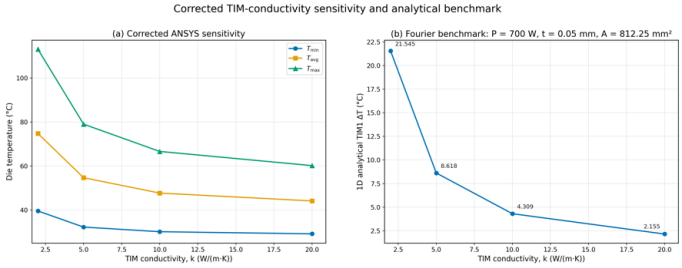


**Figure 9** Corrected TIM-conductivity sensitivity and one-dimensional TIM1 benchmark. Restoring the serial heat path produces strong sensitivity to TIM conductivity. The analytical panel shows Fourier expectations and is not an experimental validation.

## TIM thickness sensitivity

Corrected TIM1-thickness cases at $k$ = 5 W/(m·K) used the restored serial contact path, four hard through-thickness divisions, and refined-fine sizing. At 0.02, 0.05, and 0.10 mm, $T_{max}$ was 70.287, 79.004, and 95.944 °C, respectively. Increasing thickness from 0.02 to 0.10 mm raised $T_{max}$ by 25.657 °C; reducing the 0.05 mm refined-fine reference to 0.02 mm lowered $T_{max}$ by 8.717 °C. The 0.05 mm value of 79.004 °C is the refined-fine reference and is distinct from the 78.998 °C fine-mesh baseline used for fixed-temperature cross-configuration comparisons.

The displayed TIM1 full-face-average differences increased monotonically from 3.115 to 12.515 °C and followed the analytical 3.447–17.236 °C trend but remained lower than the one-dimensional estimates. Because these displayed averages were not extracted as a rigorously area-integrated paired-face metric, they are used only as a consistency trend rather than formal validation.

The results of corrected TIM1-thickness sensitivity at $k$ = 5 W/(m·K) and *i*ncreasing TIM1 thickness from 0.02 to 0.10 mm are summarized in Table 4 and Figure 10.

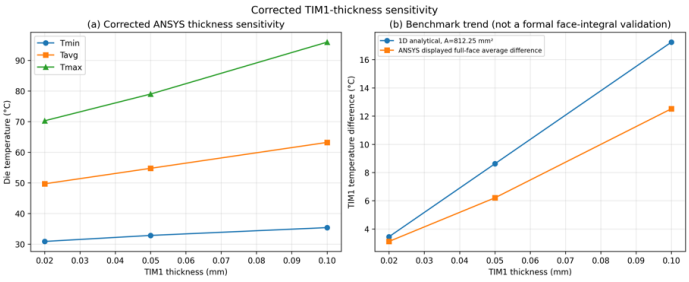


**Figure 10** Corrected TIM1-thickness sensitivity at $k$ = 5 W/(m·K). Increasing TIM1 thickness from 0.02 to 0.10 mm raised die $T_{max}$ from 70.287 to 95.944 °C. Displayed TIM1 face-average differences follow the analytical trend but are not a formal area-integrated validation metric.

**Table 4** Corrected TIM1-thickness sensitivity at $k$ = 5 W/(m·K)

| t (mm) | Stack (mm) | Nodes | Elements | $T_{min}$ | $T_{avg}$ | $T_{max}$ | Spread | Reaction | Display$\Delta T$ | 1D $\Delta T$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.02 | 10.77 | 86,311 | 34,380 | 30.886 | 49.702 | 70.287 | 39.401 | −700 | 3.115 | 3.447 |
| 0.05 | 10.80 | 92,429 | 35,668 | 32.851 | 54.760 | 79.004 | 46.153 | −700 | 6.204 | 8.618 |
| 0.10 | 10.85 | 92,429 | 35,668 | 35.403 | 63.215 | 95.944 | 60.541 | −700 | 12.515 | 17.236 |

## IHS thickness sensitivity

Corrected $T_{max}$ decreased from 80.160 °C at 1 mm to 78.998 °C at the 2 mm fine-mesh baseline and 78.389 °C at 3 mm. Thus, 3 mm was 0.609 °C lower than 2 mm and 1.771 °C lower than 1 mm, making 3 mm the tested choice when the stated objective is minimizing peak temperature. The corresponding results are shown in Figure 11.

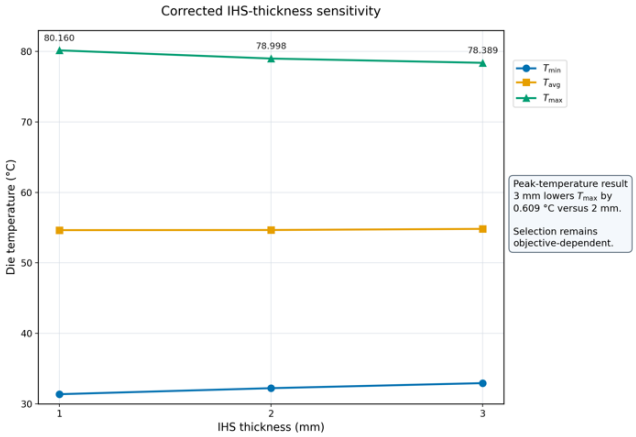


**Figure 11** Corrected IHS-thickness sensitivity. A 3 mm IHS minimizes $T_{max}$ within the tested set, while $T_{avg}$ changes by only 0.177 °C across the 1–3 mm sweep.

$T_{avg}$ was nearly unchanged across the sweep: 54.647, 54.664, and 54.824 °C at 1, 2, and 3 mm, respectively. The 3 mm case therefore had a 0.160 °C higher $T_{avg}$ than 2 mm even though its $T_{max}$ was 0.609 °C lower. This peak difference exceeds the direct fine-to-refined-fine $T_{max}$ mesh-sensitivity difference of 0.006 °C, supporting 3 mm as the lowest tested peak-temperature setting while still indicating a modest, boundary- and objective-dependent effect.

This objective-dependent fixed-sink trend is not a universal IHS-thickness rule. A manufactured DTC cold plate requires conjugate heat transfer or experiments to capture spreading, interface, and coolant-side effects.

## Cold-plate thickness sensitivity

Under the prescribed 25 °C top surface, the corrected fine-mesh 4, 8, and 12 mm cold-plate cases produced $T_{max}$ = 73.009, 78.998, and 82.793 °C, respectively; corresponding $T_{avg}$ values were 50.442, 54.664, and 57.739 °C. The 4 mm case was 9.784 °C lower in $T_{max}$ than the 12 mm case. These results are shown in Figure 12.

**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):69–79. DOI: 10.15406/mseij.2026.10.00270

Copyright: ©2026 Dwivedi et al.



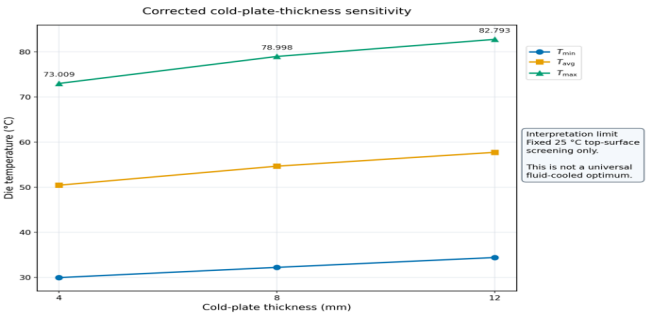


**Figure 12** Corrected cold-plate-thickness sensitivity under a prescribed 25 °C isothermal top boundary. The 4 mm case has the lowest $T_{max}$ within the tested fixed-sink set; this trend is not a universal physical cold-plate optimum.

This trend is specific to the fixed-sink boundary because reducing thickness shortens the conduction distance to the imposed isothermal surface. It is not a universal physical cold-plate design rule: internal channels, coolant heating, lateral spreading, pressure drop, flow distribution, and pump power are absent from the screening model.

### Combined material-and-geometry optimization

The combined tested material-and-geometry configuration used $k$ = 20 W/(m·K) for both TIMs, a 3 mm IHS, and a 4 mm cold plate, for a total stack height of 7.8 mm. It produced $T_{min}$ = 27.636 °C, $T_{avg}$ = 40.727 °C, $T_{max}$ = 55.059 °C, spread = 27.423 °C, and reaction = −700 W. Relative to the corrected fine-mesh baseline, $T_{max}$ fell by 23.939 °C (30.3%), $T_{avg}$ by 13.937 °C (25.5%), and spread by 19.354 °C (41.4%). This is the lowest $T_{max}$ among the tested discrete settings, not a global optimum. The corrected fixed-temperature parameter-sensitivity and combined-configuration results are summarized in Table 5.

The corrected baseline used for all Table 5 comparisons is the fine-mesh result, $T_{max}$ = 78.998 °C; the refined-fine result, $T_{max}$ = 79.004 °C, is reported only in the mesh-sensitivity study and as the 0.05 mm TIM1-thickness reference in Table 4.

**Table 5** Corrected fixed-temperature parameter-sensitivity and combined-configuration results

| Case | TIM $k$ | IHS | Cold plate | $T_{min}$ | $T_{avg}$ | $T_{max}$ | Spread | Reaction |
|---|---|---|---|---|---|---|---|---|
| Corrected baseline | 5 | 2 | 8 | 32.221 | 54.664 | 78.998 | 46.777 | −700 |
| TIM $k$ = 2 | 2 | 2 | 8 | 39.543 | 74.763 | 113.050 | 73.507 | −700 |
| TIM $k$ = 10 | 10 | 2 | 8 | 30.122 | 47.677 | 66.606 | 36.484 | −700 |
| TIM $k$ = 20 | 20 | 2 | 8 | 29.151 | 44.110 | 60.161 | 31.010 | −700 |
| IHS = 1 mm | 5 | 1 | 8 | 31.362 | 54.647 | 80.160 | 48.798 | −700 |
| IHS = 3 mm | 5 | 3 | 8 | 32.934 | 54.824 | 78.389 | 45.455 | −700 |
| Cold plate = 4 mm | 5 | 2 | 4 | 29.952 | 50.442 | 73.009 | 43.057 | −700 |
| Cold plate = 12 mm | 5 | 2 | 12 | 34.407 | 57.739 | 82.793 | 48.386 | −700 |
| Combined tested configuration | 20 | 3 | 4 | 27.636 | 40.727 | 55.059 | 27.423 | −700 |

*Note: TIM k is in W/(m·K); IHS and cold-plate thicknesses are in mm; temperatures are in °C; reaction is in W. All cases use 700 W total input, corrected contact topology, and a 25 °C fixed cold-plate top surface.*

The improvement combines a material change and two geometry changes. The combined tested configuration was subsequently evaluated under prescribed uniform and zoned convection alongside the corrected baseline package.

## 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.

The corrected convection simulations use a top-right corner-hotspot power map. Four hotspot-aligned die-facing zones use $h$ = 9,000 W/(m²·K), eight adjacent zones use $h$ = 5,000 W/(m²·K), four far zones use $h$ = 1,000 W/(m²·K), and the remaining cold-plate top area uses $h$ = 5,000 W/(m²·K). The uniform reference uses $h$ = 5,000 W/(m²·K) over the full cold-plate top.

Figure 13 shows the implemented top-right corner map paired with the corresponding 336/336/28 W power grouping. The former center-hotspot map was not rerun with corrected contacts and is not used to support any performance claim.

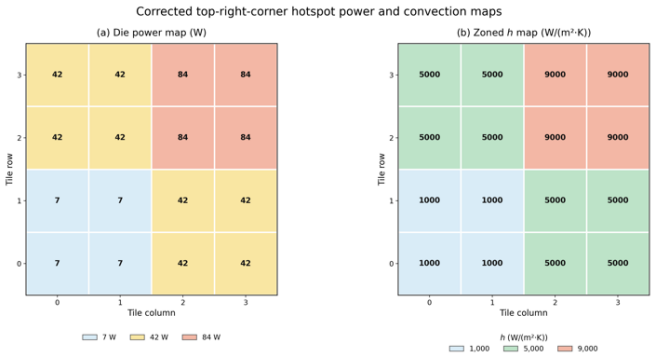


**Figure 13** Corrected top-right corner power and convection-coefficient maps. Four hotspot zones use $h$ = 9,000 W/(m²·K), eight adjacent zones use 5,000 W/(m²·K), and four far zones use 1,000 W/(m²·K); the remaining cold-plate top uses 5,000 W/(m²·K).

Each die-aligned zone corresponds to one tile of the 4 × 4 power map. The fixed-temperature condition was suppressed, and the prescribed-convection reference temperature was 25 °C. The coefficient $h$ is a boundary-condition surrogate for local cooling

**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):69–79. DOI: 10.15406/mseij.2026.10.00270

Copyright: ©2026 Dwivedi et al. 

intensity and was not computed from channel Reynolds number, flow rate, manifold pressure drops, coolant heating, or pump power.

## Area-neutral cooling budget

A zoning comparison is meaningful only when its prescribed mean coefficient is held constant. For the 16 equal-area die-aligned zones, Eq. gives

$$\left[4(9{,}000) + 8(5{,}000) + 4(1{,}000)\right]/16 = 5{,}000\ W/\left(m^2 \cdot K\right).$$

$$h = \frac{\sum_z h_z . A_z}{\sum_z A_z} \qquad (2)$$

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

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

In Eq. (2), $\bar{h}$ is the area-weighted mean heat-transfer coefficient, hz is the coefficient assigned to zone z, and Az is the corresponding zone area. Equal tile areas cancel in the arithmetic above.

The remaining cold-plate top area is also assigned $h = 5{,}000\ W/\left(m^2 \cdot K\right)$; so the entire modeled top surface has the same prescribed mean coefficient as the uniform case. Equal mean h does not establish equal flow rate, pressure drop, pump power, or hydraulic feasibility.

## Zoning results

For the corrected baseline package, uniform convection produced $T_{min}$ = 60.387 °C, $T_{avg}$ = 99.013 °C, $T_{max}$ = 133.610 °C, and spread = 73.223 °C. Zoning produced $T_{min}$ = 62.613 °C, $T_{avg}$ = 98.342 °C, $T_{max}$ = 130.000 °C, and spread = 67.387 °C. Thus, zoning reduced $T_{max}$ by 3.610 °C (2.70%), $T_{avg}$ by 0.671 °C, and spread by 5.836 °C (7.97%).

For the combined tested package, uniform convection produced $T_{min}$ = 59.854 °C, $T_{avg}$ = 85.900 °C, $T_{max}$ = 108.050 °C, and spread = 48.196 °C. Zoning produced $T_{min}$ = 61.480 °C, $T_{avg}$ = 85.455 °C, $T_{max}$ = 105.610 °C, and spread = 44.130 °C. The corresponding reductions were 2.440 °C (2.26%) in $T_{max}$, 0.445 °C in $T_{avg}$, and 4.066 °C (8.44%) in spread. These results are summarized in Table 6 and Figure 14.

**Table 6** Corrected prescribed-convection results for baseline and combined tested packages

| Package | Cooling map | TIM *k* | IHS | Cold plate | $T_{min}$ | $T_{avg}$ | $T_{max}$ | Spread | Reaction | $\Delta T_{max}$ vs uniform |
|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | Uniform 5,000 | 5 | 2 | 8 | 60.387 | 99.013 | 133.610 | 73.223 | −700 | 0.000 |
| Baseline | Zoned 9,000/5,000/1,000 | 5 | 2 | 8 | 62.613 | 98.342 | 130.000 | 67.387 | −700 | −3.610 |
| Combined tested | Uniform 5,000 | 20 | 3 | 4 | 59.854 | 85.900 | 108.050 | 48.196 | −700 | 0.000 |
| Combined tested | Zoned 9,000/5,000/1,000 | 20 | 3 | 4 | 61.480 | 85.455 | 105.610 | 44.130 | −700 | −2.440 |

*Note: TIM k is in W/(m·K); IHS and cold-plate thicknesses are in mm; temperatures are in °C; reaction is in W. All cases use 700 W total input, 25 °C ambient, the corrected contact topology, a top-right corner-hotspot map, and the fixed-temperature boundary suppressed.*

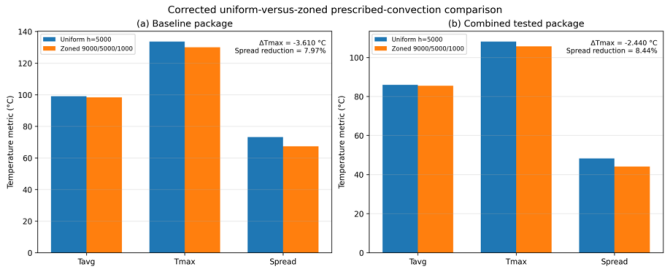


**Figure 14** Corrected uniform-versus-zoned comparison for (a) the baseline package and (b) the combined tested package. Bars show $T_{avg}$, $T_{max}$, and spread at identical area-weighted mean *h*; this does not imply identical hydraulic power.

Zoning raised $T_{min}$ while lowering $T_{max}$ and spread in both package configurations. This is a redistribution effect: cooling effort is shifted away from lower-power regions toward the hotspot, compressing the temperature field rather than cooling every location equally.

Practical feasibility requires conjugate flow and hydraulic analysis. Spatially prescribed h values cannot be assumed independently controllable, and equal mean h is not equivalent to equal pump power or pressure drop. Channel geometry, valve authority, flow maldistribution, coolant rise, pump curves, and actuator latency must be resolved before hardware implementation.

## Interaction of package optimization and zoning

Baseline uniform→combined tested uniform lowered $T_{max}$ by 25.560 °C (19.13%); baseline zoned→combined tested zoned lowered $T_{max}$ by 24.390 °C (18.76%); and baseline uniform→combined tested zoned lowered $T_{max}$ by 28.000 °C (20.96%). Package/material optimization therefore sets the larger absolute thermal reduction, while zoning adds targeted peak suppression and consistently reduces nonuniformity.

# Proposed cyber-physical hotspot prediction and control framework

The corrected prescribed-convection results motivate a possible software-driven control layer above future thermal hardware: static zoning reduced $T_{max}$ by 2.440–3.610 °C without changing the area-weighted mean h. These cases do not demonstrate hotspot migration or controller performance. The architecture below is a proposed framework only; no training dataset, predictor-error metric, inference latency, closed-loop stability result, or hardware actuation test exists in the present study.

**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):69–79. 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. $x_{i,t} = [P_{i,t}, T_{i,t}, TN(i)_t, a_{i,t}, m_{i,t}, \Delta T_{i,t}]^T$ (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), $TN(i)_t$ 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).

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

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 $_{N}(i)$ 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).

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

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

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*; $T_{i,t+1}$ 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).

$$u_t^* = \arg\min_u \left[ \max_i \hat{T}_{i,t+1}(u) + \lambda \| u - u_{t-1} \|_2^2 + \mu C(u) \right] \quad (6)$$

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); $\lambda$ is the movement-penalty weight and $\mu$ 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.

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

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 15 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

Copyright: ©2026 Dwivedi et al.


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.

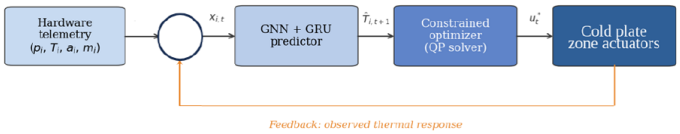


**Figure 15** Proposed conceptual cyber-physical control architecture; not trained, implemented, or experimentally identified. Hardware telemetry supplies tile power, tile temperature, workload counters, and actuator state at time *t*. The GNN–GRU predictor estimates the next-step temperature field, and the constrained optimizer computes the zone allocation vector for future cold-plate actuators.

### 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, future trained predictor parameters (not available in the present study), 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 original auto-contact topology bypassed both TIM layers and invalidated the former bottleneck ranking. After correction, TIM conductivity is a major tested design variable. Cold-plate thickness remains influential under the prescribed fixed-sink boundary, whereas IHS thickness has a smaller, objective-dependent effect. The combined material-and-geometry configuration provides the largest absolute temperature reduction, and zoned prescribed convection supplies a smaller additional reduction while consistently decreasing temperature spread.

The fixed-temperature sink is an idealized conduction-screening boundary. The prescribed-$h$ zoning model likewise imposes local coefficients rather than resolving coolant flow. Neither model includes internal channel geometry, conjugate heat transfer, coolant heating, pressure drop, pump power, flow maldistribution, contact resistance, TIM anisotropy, transient response, structural stress, manufacturing constraints, hardware telemetry, or actuator dynamics. The corrected preference for a 3 mm IHS follows the stated $T_{max}$ objective within the tested set, and the 4 mm cold-plate result follows the shorter conduction path to the imposed isothermal surface. Neither is a universal hardware design rule.

The zoned simulations prescribe spatially varying convection coefficients instead of resolving conjugate fluid flow. Zoning raises $T_{min}$ because cooling capacity is removed from lower-power regions and redirected toward the hotspot; the simultaneous decrease in $T_{max}$ and spread demonstrates compression of the temperature field rather than a contradiction.

Equal area-weighted mean $h$ is a controlled thermal-boundary comparison, not a hydraulic-energy comparison. Different spatial coefficient maps may require different mass-flow distribution, pressure drop, valve authority, and pump power. Those quantities must be modeled before claiming equal cooling cost.

The predictor-controller pipeline in Section 6 is a proposed architecture, not a demonstrated closed-loop system. Its training data, prediction accuracy, latency, stability, distribution-shift robustness, and actuation feasibility remain untested. The present static zoning cases provide motivation only, not controller validation.

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]

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.

**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):69–79. DOI: 10.15406/mseij.2026.10.00270

## Conclusions

This study combined an industry survey with corrected comparative thermal screening of a representative 700 W, 86.18 W/cm² AI-chip stack. The central verification finding was that a 0.21384 mm automatic contact tolerance created nonphysical Die–IHS and IHS–cold-plate bypasses; suppressing Regions 5 and 3 restored the intended serial path and changed the thermal sensitivities and design ranking. With corrected contacts, TIM conductivity was strongly influential, and increasing TIM1 thickness from 0.02 to 0.10 mm raised $T_{max}$ by 25.657 °C. The IHS effect was modest and objective-dependent: 3 mm gave the lowest tested $T_{max}$, 0.609 °C below 2 mm. The 4 mm cold-plate result applies only to the prescribed 25 °C fixed-sink boundary.

The combined tested configuration—$k$ = 20 W/(m·K), a 3 mm IHS, and a 4 mm cold plate—reduced $T_{max}$ by 23.939 °C (30.3%) relative to the corrected fine-mesh baseline. At equal prescribed mean $h$, zoning provided an additional 3.610 °C reduction for the baseline package and 2.440 °C for the combined package while reducing temperature spread. Energy balance, direct mesh sensitivity, and Fourier calculations provide numerical verification and analytical benchmarking, not experimental validation. Future work must prioritize conjugate coolant-flow and hydraulic modeling, transient and structural analyses, experimental validation, and only then training and validating the proposed GNN–GRU controller.

## Data Availability

Corrected ANSYS project archives, contact-topology records, boundary-condition specifications, and tabulated outputs underlying this study are available from the corresponding author upon reasonable request. The validated archive manifest may be supplied as supplementary material. 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.