

3D-ICE 4.0: Accurate and efficient thermal modeling for 2.5D/3D heterogeneous chiplet systems

Kai Zhu*, Darong Huang*, Luis Costero^o, David Atienza*

*{kai.zhu, darong.huang, david.atienza}@epfl.ch – *Embedded Systems Laboratory, EPFL, Switzerland*

^olcostero@ucm.es – *Universidad Complutense de Madrid, Spain*

Abstract—The increasing power densities and intricate heat dissipation paths in advanced 2.5D/3D chiplet systems necessitate thermal modeling frameworks that deliver detailed thermal maps with high computational efficiency. Traditional compact thermal models (CTMs) often struggle to scale with the complexity and heterogeneity of modern architectures. This work introduces 3D-ICE 4.0*, designed for heterogeneous chip-based systems. Key innovations include: (i) preservation of material heterogeneity and anisotropy directly from industrial layouts, integrated with OpenMP and SuperLU MT-based parallel solvers for scalable performance, (ii) adaptive vertical layer partitioning to accurately model vertical heat conduction, and (iii) temperature-aware non-uniform grid generation. The results with different benchmarks demonstrate that 3D-ICE 4.0 achieves speedups ranging from $3.61\times$ – $6.46\times$ over state-of-the-art tools, while reducing grid complexity by more than 23.3% without compromising accuracy. Compared to the commercial software COMSOL, 3D-ICE 4.0 effectively captures both lateral and vertical heat flows, validating its precision and robustness. These advances demonstrate that 3D-ICE 4.0 is an efficient solution for thermal modeling in emerging heterogeneous 2.5D/3D integrated systems.

Index Terms—Thermal modeling, 2.5D/3D integration, chiplet systems, adaptive algorithm, non-uniform, parallel acceleration

I. INTRODUCTION

The explosive growth of compute-intensive workloads, from high-performance computing (HPC) to large language models (LLMs), has pushed chip power densities to unprecedented levels. Modern AI and HPC accelerators routinely dissipate hundreds of watts per package [1], and multi-kilowatt servers concentrate substantial heat fluxes that, if not modeled and appropriately managed, can accelerate wear-out or trigger thermal throttling, ultimately degrading throughput [2], [3]. Temperature has therefore become a first-order design constraint on par with performance and power [4], requiring thermal models that can accurately predict full-chip temperature distributions under realistic and time-varying workloads.

Advanced 2.5D/3D heterogeneous integration (e.g., chiplet-based system-in-package or High Bandwidth Memory (HBM) stacked on logic dies) offers a promising way to scale compute and bandwidth beyond monolithic designs. However, these packaging approaches introduce complex heat flow paths [5], lateral and vertical thermal coupling across heterogeneous materials (such as active silicon, interposer, underfill, micro-bumps, and through-silicon via (TSV) arrays), creating multi-scale paths with significant anisotropic characteristics [6]. Hotspots in one die can increase the temperatures in neighboring dies. For example, in a tightly integrated 2.5D Graphics Processing Unit (GPU) system, the heat from a high-power

die can raise the temperature of a neighboring HBM stack enough to induce a four-fold increase in the refresh rate [7], significantly reducing the throughput of the system. These realities require more detailed and accurate thermal models spatially and temporally to capture localized hotspots and thermal crosstalk in such complex architectures [4], [8].

Thermal analysis faces a non-trivial trade-off between accuracy and efficiency. Finite element methods (FEM) can accurately capture detailed heat conduction but are computationally and memory intensive, often requiring prohibitive costs for large-scale designs [9]. Compact thermal models (CTMs) based on equivalent thermal RC network (e.g., Hotspot [10], 3D-ICE [11], PACT [8]) run orders of magnitude faster than FEM and are widely used for thermal modeling and design-space exploration [12]. However, this speed often comes at the cost of fidelity. Current CTMs have three common limitations: 1) reliance on coarse grids and homogenized materials and the difficulty of ingesting detailed, anisotropic geometries from industrial layouts (e.g., GDSII) into RC networks—which can misestimate on-die temperature gradients and hotspot magnitudes; 2) the use of a layer-based model with a division of one layer for each functional component, which can be insufficient to capture accurate cross-layer heat flow; and 3) the use of global uniform grids for simplicity, which either under-resolves hotspot regions or over-resolves other regions, yielding a suboptimal accuracy–efficiency trade-off as die size or layer count increase. Although 3D-ICE 3.1 supports non-uniform grids [3], a thermal-aware grid size selection strategy remains unexplored, which means users need to assign grid sizes to each region manually based on domain knowledge, which can be suboptimal for complex floorplans. Therefore, a fine-grained CTM that preserves detailed geometry and material-aware lateral and vertical conduction in 2.5D/3D heterogeneous systems while remaining computationally efficient is needed.

Based on these challenges and requirements, we present 3D-ICE 4.0, an open-sourced* framework and the key contributions include:

- **Fine-detail layout and anisotropy-aware modeling:** An interface that automatically builds detailed models from industrial layout files, retaining detailed material distributions and anisotropic properties to enable accurate prediction of localized hotspots and cross-die interactions. Meanwhile, OpenMP [13]-based thermal matrix construction and a multi-threaded solver (SuperLU MT [14]) are used for rapid steady-state and transient thermal simulation

* Open-source repository: <https://github.com/esl-epfl/3d-ice>.

TABLE I
COMPARISON OF REPRESENTATIVE COMPACT THERMAL MODELS.

Thermal models	Geometry construction			Grid generation			Solution
	Heterogeneous	Anisotropic	Automatic	Adaptive layer division	Non-uniform grids	Thermal-aware grids	Parallel acceleration
Hotspot 7.0 [10]	✓						
3D-ICE 3.1 [3]					✓		
PACT [8]	✓						✓
MFIT [9]	✓	✓			✓		✓
ARTSim 2.0 [22]	✓				✓		✓
3D-ICE 4.0	✓	✓	✓	✓	✓	✓	✓

for large-scale chip models.

- **Adaptive layer division:** A strategy that subdivides layers with high vertical thermal resistance to better resolve cross-layer heat flow, improving accuracy for multilayer 2.5D/3D chiplet systems.
- **Temperature-aware non-uniform grids:** A local grid refinement method that concentrates resolution in thermal critical regions to capture fine-grained heat flow variations, which can achieve a target error with far fewer grids than uniform meshes, improving simulation efficiency.

II. RELATED WORK

Thermal modeling for 2.5D/3D chiplet systems spans three main families [15]: (i) analytical/Green’s function methods, (ii) numerical methods (e.g., FEM and CTMs), and (iii) Machine Learning (ML) methods. Green’s function approaches precompute the system’s impulse responses and obtain temperature via convolution between Green’s function and power maps, enabling rapid evaluations for fixed stacks [16]. However, changes in geometry, materials, or boundary conditions typically require recomputation of Green’s function. FEM, as implemented in commercial software such as Ansys [17] and COMSOL [18], can offer high precision for complex structures and is often treated as reference solutions, but their runtime and memory costs limit rapid design space exploration and runtime thermal management. CTMs [10], [11] approximate heat flow with equivalent thermal RC networks and are widely used for early design [12] and thermal management [3] due to orders-of-magnitude speedups over FEM [9]. Recent works also explore ML surrogates for fast temperature prediction, such as convolutional neural networks (CNNs) [19], graph neural networks (GNNs) [20], and physics-informed neural networks (PINNs) [21]. ML methods typically require extensive training data generated by FEM or CTMs, and struggle to generalize to unseen architectures [20]. Consequently, CTMs remain a practical and reliable choice for thermal analysis of chiplet systems.

Hotspot [10] pioneered RC-network-based CTMs for early-stage thermal analysis. To improve flexibility and efficiency, 3D-ICE 3.1 [11] and ARTSim 2.0 [22] introduce non-uniform grids. As simulated systems become increasingly complicated, PACT [8] integrates a parallel SPICE solver to accelerate fine-grained thermal simulations. Similarly, MFIT [9] and ARTSim 2.0 [22] use sparse linear algebra BLAS libraries and Python libraries SciPy and NumPy to speed up sparse solves, respectively. Despite these advancements, prior CTMs generally fall

short in: 1) fast modeling of fine-detail heterogeneous layout, 2) accurate modeling of vertical heatflows, and 3) clear policies for non-uniform grids. The comparison of different representative CTMs is summarized in Table I.

III. 3D-ICE 4.0 FRAMEWORK

The typical CTM workflow (e.g., 3D-ICE 3.1) is shown in Fig. 1, and the new or updated stages in 3D-ICE 4.0 are highlighted in orange. After parsing the inputs (a), various components and a multilayer stack are built (b.1 and b.2). The stack is discretized into grids to obtain an RC network (b.3), which is transformed into a sparse linear system for steady-state or transient analysis (c and d). Unlike prior CTMs, 3D-ICE 4.0 supports heterogeneous and anisotropic materials, and directly imports detailed material distributions from industrial layout files (b.1.1). To better capture complex cross-layer heat flow in 2.5D/3D chiplet systems, we introduce an adaptive layer division strategy (b.2.1). In addition, a thermal-aware local grid refinement (b.3.1) is employed to achieve a better trade-off between accuracy and efficiency.

Profiling shows that RC matrix assembly and sparse linear solutions dominate the runtime for CTMs and limit the scalability of detailed large-scale models. We address this bottleneck with OpenMP-based parallelized assembly [13] and a high-performance sparse solver SuperLU MT [14]. We will detail the novelties and functionalities of 3D-ICE 4.0 in the following subsections.

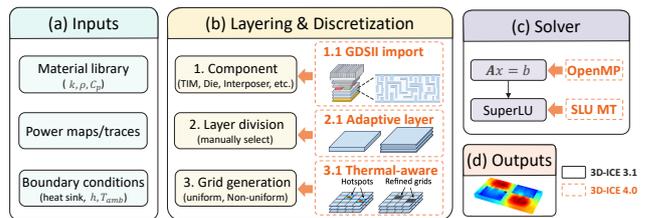


Fig. 1. Workflow of compact thermal modeling framework. Newly added/updated stages in 3D-ICE 4.0 are highlighted in orange.

A. Model generation with detailed layouts

Accurate thermal simulation requires models that preserve the spatial heterogeneity of materials and layouts. In many prior CTMs [10], [11], complex layers are approximated by an effective medium calculated from the volume fractions of the constituent materials. To quantify the impact of this simplification, we model a 4-chiplet system in 3D-ICE 3.1 using homogenized layers, and the predicted hotspot is 390.15

K, as shown in Fig. 2 (a). In contrast, when the detailed distribution of the material is retained, the hotspot increases to 394.22 K, as presented in Fig. 2 (b). This discrepancy underscores the importance of modeling detailed layouts.

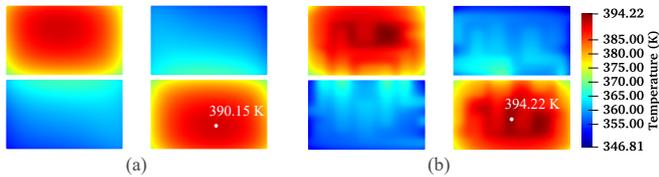


Fig. 2. Thermal map for a 4-chiplet system with (a) abstract homogeneous model, (b) heterogeneous model with detailed material distribution.

The manual construction of an RC network that preserves fine-detail material distributions is laborious and error-prone. We automate this step by directly importing material layouts from industry design files (e.g., GDSII files). Two challenges exist: i) CTMs typically assume rectangular tiles to describe the floorplan, whereas real designs contain irregular geometries; and ii) GDSII layers often include overlapping objects spanning disparate length scales. Our framework addresses these issues by i) incorporating equivalent models for irregular shapes [6], [23]; and ii) applying an adaptive and layout-aware tiling scheme. The extraction pipeline partitions each layer into non-overlapping tiles, computes each tile’s intersection area with the polygons and paths in the GDSII layout, and then generates anisotropic per-tile equivalent models for subsequent thermal simulation.

A quadtree-based tiling strategy (Algorithm 1) automates tile generation from layout geometry. The inputs include the GDS file, target layer, maximum recursion depth (i_{max}), and a stopping condition (i_{con}). All layer primitives (e.g., polygons, paths/FlexPaths) are first polygonized and merged into disjoint shapes via geometric union (line 2), then indexed with an R-tree built on their bounding boxes to accelerate spatial queries. The initial domain is the bounding box of the merged shapes (line 4), and we invoke the recursive SUBDIVIDE function. At each quadtree node, we query the R-tree for candidate shapes that intersect the current region (line 8), compute the overlap ratio ρ , and evaluate the stopping rule (line 10). If the criterion is satisfied, one tile record including position, size, and ρ , is emitted to the tile list; otherwise, the region is subdivided into four quadrants and processed recursively (lines 13-15). The fidelity can be adjusted flexibly by i_{max} and i_{con} . For a complex layout in Fig. 3 (a), the resulting tile partition and associated ρ map are shown in Fig. 3 (b) and (c), which can be used to generate the detailed heterogeneous and anisotropic model for the layer. The accuracy of the built model is validated in Section IV-B.

To accelerate simulations with detailed material distributions, 3D-ICE 4.0 parallelizes RC matrix assembly with OpenMP [13] and uses SuperLU MT [14] for multi-thread preordering, factorization, and solution. As shown in Section IV-A, our framework can achieve a higher parallel efficiency than the SPICE-based solver in PACT [8].

Algorithm 1 Tile generation with quadtree-based subdivision

Inputs: $gds, layer, i_{max}, i_{con}$
Output: $tiles$ (list of (position, size, ρ))

- 1: $S \leftarrow READGDS(gds, layer)$
- 2: $M \leftarrow MERGE(S)$
- 3: $I \leftarrow BUILDINDEX(M)$ \triangleright R-tree over merged polygons
- 4: $R \leftarrow MAKEWINDOW(M)$ \triangleright bounding region
- 5: $tiles \leftarrow SUBDIVIDE(I, R, i, i_{max}, i_{con})$
- 6: **return** $tiles$

.....

- 7: **function** SUBDIVIDE($I, R, i, i_{max}, i_{con}$)
- 8: $C \leftarrow CANDIDATES(I, R)$ \triangleright query R-tree
- 9: $A_{ov} \leftarrow OVERLAPAREA(R, C), \rho \leftarrow A_{ov}/AREA(R)$
- 10: **if** Satisfy (i_{con}) **or** $i \geq i_{max}$ **then**
- 11: $tiles \leftarrow (R, \rho)$
- 12: **else**
- 13: $(R_1, R_2, R_3, R_4) \leftarrow SPLITQUAD(R)$
- 14: **for** $r \in \{R_1, R_2, R_3, R_4\}$ **do**
- 15: $SUBDIVIDE(I, r, i+1, i_{max}, i_{con})$
- 16: **end for**
- 17: **end if**
- 18: **end function**

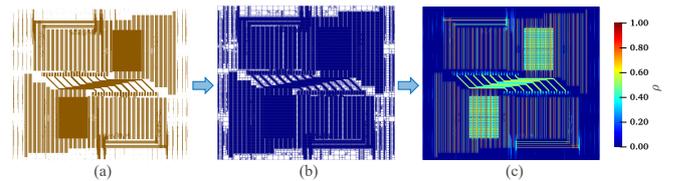


Fig. 3. Extraction of detailed material distribution, (a) GDSII layout of one layer, (b) generated tiles, (c) overlapped ratio for each tile.

B. Adaptive layer division

Vertical heat flow in 2.5D/3D heterogeneous chiplet systems is highly nonuniform and sensitive to material and geometric details [24]. Conventional CTMs often align simulation layers (RC network layers) one-to-one with functional layers. This practice can induce large contrasts in vertical thermal resistances, where layers with high thermal resistance accumulate disproportionate temperature gradients and can not reflect the accurate inter-layer temperature profile [25].

To alleviate this problem, we introduce an adaptive layer division strategy during model construction, based on vertical thermal resistance. For layer j with N_j floorplan elements, the vertical resistance of element i is $R_{j,i}^\perp = h_j / (k_{j,i}^\perp \cdot A_{j,i})$, where h_j is the layer thickness, $k_{j,i}^\perp$ is the equivalent vertical thermal conductivity, and $A_{j,i}$ is the area. The layer’s equivalent vertical resistance is the parallel combination of each element $R_j^\perp = (\sum_{i=1}^{N_j} \frac{1}{R_{j,i}^\perp})^{-1}$. We quantify heterogeneity across an n -layer stack by the variance

$$\text{Var}(R^\perp) = \frac{1}{n} \sum_{j=1}^n (R_j^\perp - \bar{R}^\perp)^2 \quad (1)$$

where \bar{R}^\perp is the mean vertical resistance.

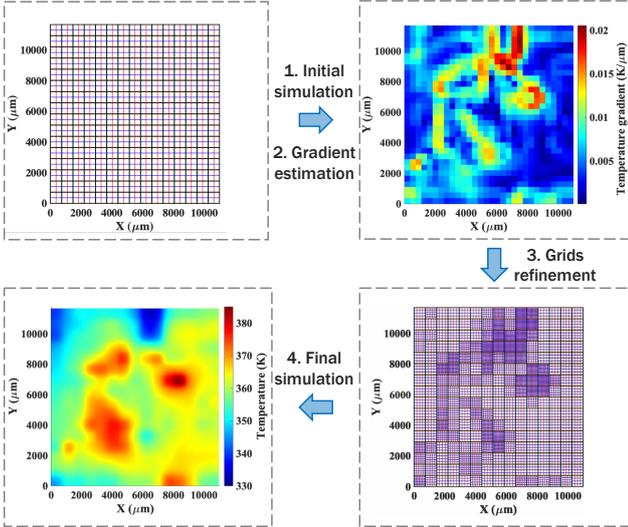


Fig. 4. Workflow of the generation of thermal-aware non-uniform grids and simulation.

Layers that dominate $\text{Var}(R^\perp)$ or exhibit large R_j^\perp are iteratively subdivided into multiple simulation sublayers, reducing and redistributing the per-layer vertical resistance. Iteration stops when the variance falls below a prescribed threshold or reaches a maximum depth of iteration. The refinement is performed before the simulation using the material and geometric parameters, and the associated overhead is negligible. The accuracy gains of the layer division strategy are validated in Section IV-B.

C. Local grid refinement

3D-ICE 3.1 [3] and ARTSim 2.0 [22] support non-uniform grids within one layer and across different layers, offering more flexibility and a better accuracy-efficiency trade-off than uniform grids. Finer grids can be used in specific regions of interest without modifying the grids in other regions to reduce computational overhead and simulation time. However, a systematic policy for choosing the local grid granularity has not been fully explored. Therefore, we propose an automatic refinement strategy guided by local temperature gradients, as shown in Fig. 4. Starting from a coarse mesh, we perform an initial solve and estimate the average temperature gradient G for each floorplan element. Elements exhibiting large gradients (e.g., near hotspots) are selectively refined to better capture steep variations. For a given element, the grid size in the x-direction is chosen as:

$$\text{Gridsize}_x = l_x \cdot \alpha \cdot \left(\frac{G_{base}}{G + \epsilon} \right) \quad (2)$$

where l_x is the length of the floorplan element in the x-direction, G_{base} and α are control parameters for the refinement of the grid and $\epsilon > 0$ avoids division by zero. The calculated Gridsize_x needs to be regulated to $[l_{min}, l_x]$, where l_{min} is the minimum granularity permitted by runtime and memory constraints. The size of the grid in the y-direction can be computed in a similar way. The efficiency gains of this temperature-aware local refinement over uniform grids are demonstrated in Section IV-C.

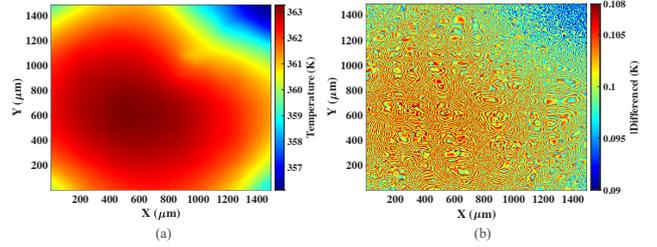


Fig. 5. (a) Temperature map from 3D-ICE 4.0 with a uniform 256×256 grid, (b) the absolute difference map between 3D-ICE 4.0 and PACT.

IV. EXPERIMENTAL RESULTS

In this section, three validation examples are provided to evaluate the accuracy and efficiency improvement of 3D-ICE 4.0. All simulations run on a server with two AMD EPYC 9734 CPUs and 1.0-TB DDR5 memory, under Ubuntu 24.04.

A. Case1: Accuracy and efficiency validation

To validate the accuracy of our framework, we reproduce the example of the state-of-the-art parallel compact thermal model PACT [8], using a uniform 256×256 grid. Fig. 5 (a) shows the steady-state temperature map produced by 3D-ICE 4.0, and Fig. 5 (b) shows the absolute difference with respect to PACT. The small discrepancy demonstrates the accuracy of 3D-ICE 4.0 with respect to PACT, which has been validated against Hotspot [10] and COMSOL [8], and also provides indirect validation of 3D-ICE 4.0 against these simulators.

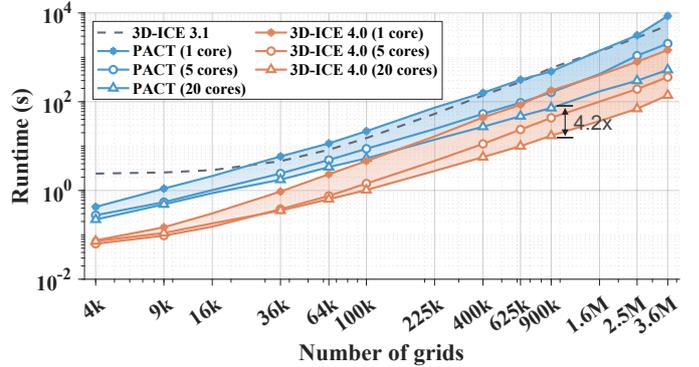


Fig. 6. Runtime for 3D-ICE 3.1, PACT and 3D-ICE 4.0 with 1, 5, and 20 cores versus different problem sizes.

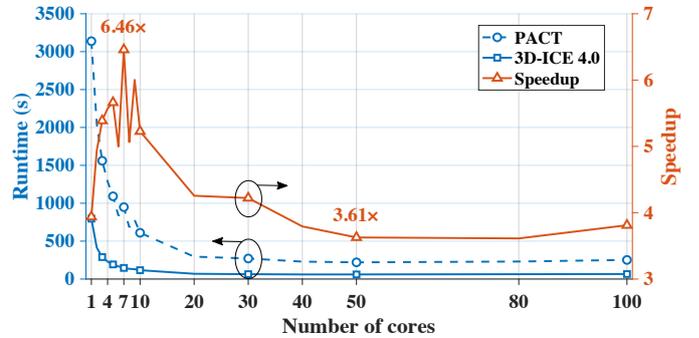


Fig. 7. Runtime for PACT and 3D-ICE 4.0 with different numbers of cores.

To assess parallel efficiency, we sweep the grid resolution and compare the simulation time for serial solver 3D-ICE 3.1,

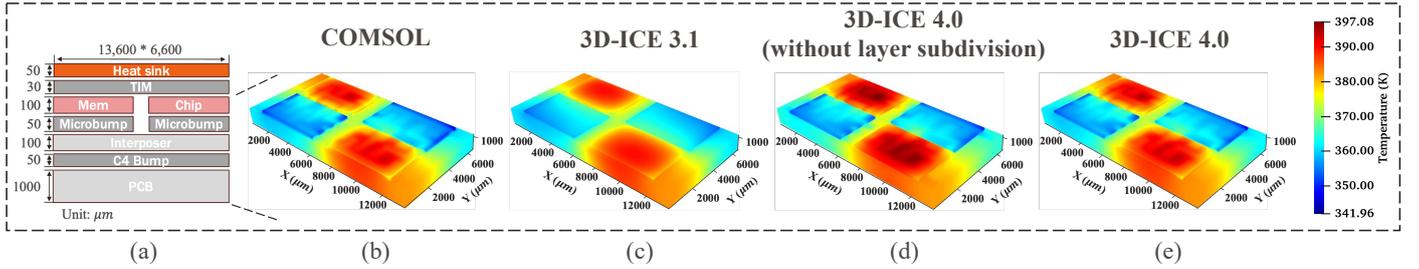


Fig. 8. Simulated geometry and 3D temperature distribution. (a) Geometry and dimension of the chiplet system, temperature map from (b) COMSOL, (c) 3D-ICE 3.1, (d) 3D-ICE 4.0 without layer subdivision, (e) 3D-ICE 4.0 with layer division.

PACT, and 3D-ICE 4.0 using 1, 5, and 20 CPU cores in Fig. 6. As problem size increases, both PACT and 3D-ICE 4.0 exhibit higher speedup ratios, and for any fixed core count, 3D-ICE 4.0 consistently achieves shorter runtimes than PACT across the test, indicating better parallel efficiency and scalability.

To compare the speedup ratio with more different core numbers, we fix the grid count as 2.5 million, then vary the core count and compare 3D-ICE 4.0 with PACT, as shown in Fig. 7. It can be found that the runtime keeps decreasing with increasing cores up to around 30, after which gains saturate. Across all core counts tested, 3D-ICE 4.0 is $3.61\times$ – $6.46\times$ faster than PACT. For time division analysis, when the number of cores is set to 7, PACT requires 949.0 s to finish the simulation, 148.2 s for netlist import and matrix assembly, and 800.8 s for solving. In contrast, 3D-ICE 4.0 finishes in 146.9 s, with only 0.8 s for parsing the inputs and assembling the matrix system, and 146.1 s for solving. In addition, the peak memory usage of our framework (35.28 GB) is 27.3% lower than that of PACT (48.43 GB), which highlights the memory efficiency of our approach.

TABLE II
THERMAL PROPERTIES OF DIFFERENT COMPONENTS

Component	Thermal conductivity [W/(m · K)]	Density [kg/m ³]	Heat capacity [J/(kg · K)]
Heatsink	385	8,900	387
TIM	5	2,500	1,000
Chip, PCB	130	2,300	700
microbump, C4 bump	5.5(//), 113(⊥)	7,380	250
Underfill	1.5	1,400	1,100

B. Case2: Modeling with detailed material distribution and accuracy gains by adaptive layer division

After initial validation, we evaluate how detailed material layouts and adaptive layer division affect accuracy of thermal modeling of a 2.5D chiplet system [26]. The simulated system is shown in Fig. 8 (a), material properties are listed in Table II, and the interposer layout is imported from the GDSII file in [27]. The steady-state temperature map from COMSOL serves as the reference (Fig. 8 (b)). Using 3D-ICE 3.1 with a homogenized model fails to capture the temperature gradients, as shown in Fig. 8 (c). Incorporating detailed material distributions in 3D-ICE 4.0 can capture these temperature gradients, but still deviates from COMSOL due to the limited vertical

resolution, as shown in Fig. 8 (d). After 8 iterations of adaptive layer division, as illustrated in Section III-B, results from 3D-ICE 4.0 align closely with COMSOL, as presented in Fig. 8 (e), underscoring the importance of layer division. For a fair comparison, we match the number of grid unknowns in 3D-ICE 3.1 and 3D-ICE 4.0 to the degrees of freedom (DoFs) of the COMSOL setup.

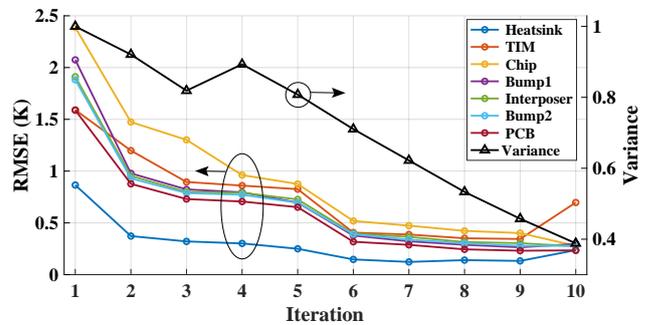


Fig. 9. RMSE for functional layers during layer division iteration.

Across iterations, the layers with the larger vertical thermal resistance are adaptively split into thinner layers with the lower vertical resistance. The normalized variance of vertical resistance across all simulation layers (Fig. 9) decreases as the distribution becomes more uniform, and the root mean square error (RMSE) for each functional layer with respect to COMSOL falls accordingly (Fig. 9). For the chip and interposer, Fig. 10 (a) and (b) compare the results from 3D-ICE 4.0 with COMSOL: the first iteration shows notable errors, especially in high-power regions, whereas after 8 iterations-by which time the TIM and PCB are each subdivided into four sub-layers, and the chip into two-the agreement is markedly improved. It is worth noting that we keep the total number of grids per functional layer constant across iterations, which means the accuracy improvement arises from improved vertical temperature modeling rather than an increased grid count. A temperature slice through the stack’s center along the z -axis is shown in Fig. 10 (c), which further confirms that the vertical temperature profile converges toward the COMSOL reference after iteration.

To assess the accuracy of our framework for transient thermal analysis, we apply Gaussian-modulated sinusoidal power inputs:

$$P(t) = P_0 + P_0 \exp\left(-\frac{(t - t_0)^2}{\tau^2}\right) \sin(\omega t) \quad (3)$$

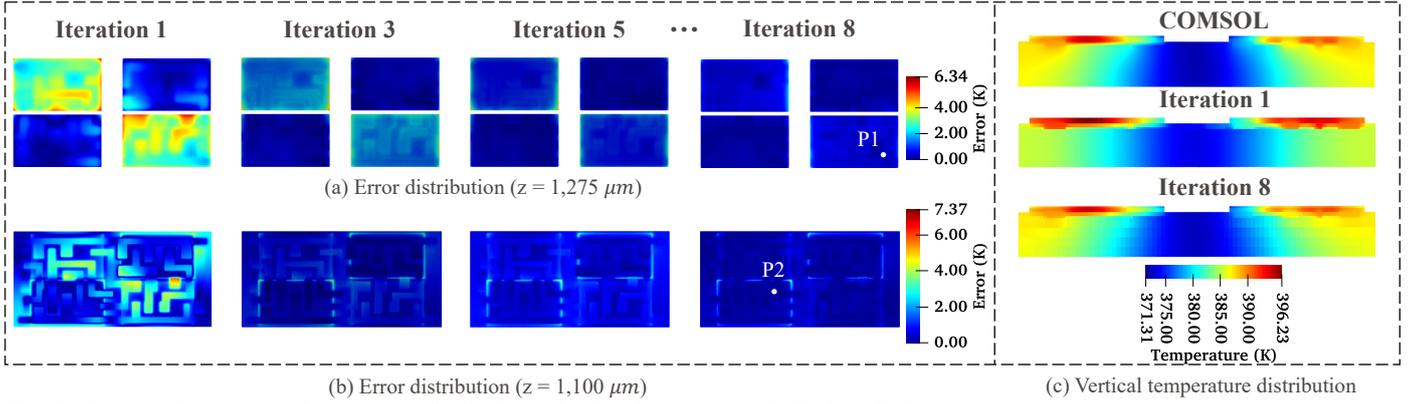


Fig. 10. Cross-sectional error and temperature distribution during iteration. (a) error distribution of chip ($z = 1,275 \mu\text{m}$) and (b) interposer ($z = 1,100 \mu\text{m}$) from 3D-ICE 4.0 compared with COMSOL, (c) vertical temperature slice at the center point along the z -axis.

where $\tau^2 = 0.1$, $\omega = 10\pi$, $t_0 = 0.5$, $P_0 = 25W$ and $50W$ for the memory and logic dies, respectively. We run a transient simulation and record temperatures at two observation points, the locations are shown in Fig. 10. Fig. 11 compares transient temperature traces of 3D-ICE 4.0 without layer division, with layer division, and the COMSOL reference. Without layer division, the temperature traces deviate markedly from COMSOL, while the traces with layer division agree well with COMSOL in both amplitude and timing, underscoring the necessity of layer division and validating that 3D-ICE 4.0 can accurately capture fast, complex thermal dynamics.

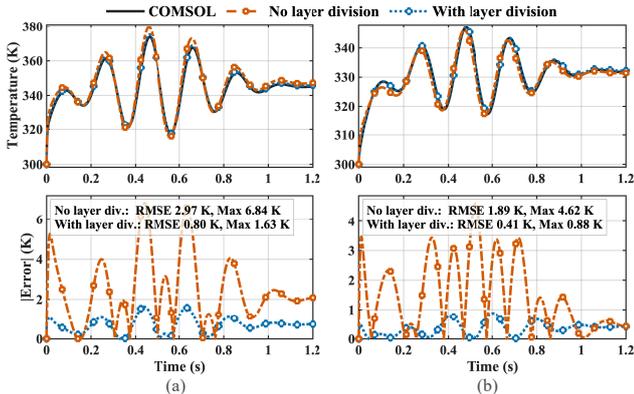


Fig. 11. Transient temperature at two observation points: 3D-ICE 4.0 without/with adaptive layer division versus COMSOL, and absolute error with respect to COMSOL, (a) point 1, (b) point 2.

C. Case3: Efficiency improvement by local refined grids

After the accuracy validation of adaptive layer division, we evaluate the efficiency gains of local refined grids introduced in Section III-C. The test case is a $100 \mu\text{m}$ -thick die with a heat sink attached on top; the die power map is taken from the gate-level simulation in [8]. Following the workflow in Fig. 4, we start from a coarse mesh (960 grids), perform an initial simulation, compute the spatial distribution of temperature gradients, and refine grids according to the local gradients.

We set $\alpha = 1$ in Eq. (2) and sweep G_{base} to control the total grid count, then run a final simulation with the local refined grids. The RMSE versus grid count is shown in Fig. 12. For comparison, we also solve with uniform grids over a similar range of grid counts and plot the error curve. The non-uniform

strategy reaches a target error with fewer grids: to achieve $\text{RMSE} < 0.3\text{K}$, the uniform strategy requires 11,760 grids versus 8,714 for the non-uniform strategy (25.9% fewer). For $\text{RMSE} < 0.25\text{K}$, the requirements are 15,360 and 11,787 cells, respectively (23.3% reduction).

The corresponding absolute error maps for the 15,360 uniform grids and the 11,787 non-uniform grids are also shown in Fig. 12. The peak error for uniform grid simulation can be up to 1 K, and concentrates in regions with high temperature gradients. By using finer grids in these regions, non-uniform grid simulation reduces the peak error to 0.6 K. Therefore, local-refined grids not only improve efficiency (fewer grids for a target RMSE) but also enhance the ability to capture steep temperature gradients.

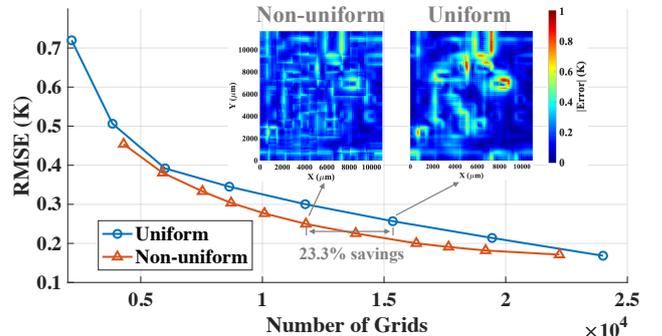


Fig. 12. Error with different numbers of grids for uniform grids and non-uniform grids.

V. CONCLUSION

In this work, we have presented 3D-ICE 4.0, a new compact thermal modeling framework, which preserves material heterogeneity and anisotropy from industrial layouts, introduces adaptive layer partitioning, applies temperature-aware non-uniform grids to reduce complexity, and employs parallel acceleration for scalable thermal simulation. Experiments show that 3D-ICE 4.0 can achieve $3.61\times$ – $6.46\times$ speedups over PACT and closely matches COMSOL in both steady-state and transient simulations. For a target error, the grid strategy can reduce unknowns by 23–26%, improving efficiency without sacrificing accuracy. These results indicate that 3D-ICE 4.0 is promising for faster design-space exploration and more reliable runtime thermal management for emerging 2.5D/3D chiplet systems.

ACKNOWLEDGMENT

This work was supported in part by the Swiss State Secretariat for Education, Research, and Innovation (SERI) through the SwissChips research project.

REFERENCES

- [1] I. Latif, A. C. Newkirk, M. R. Carbone, A. Munir, Y. Lin, J. Koomey, X. Yu, and Z. Dong, "Single-node power demand during ai training: Measurements on an 8-gpu nvidia h100 system," *IEEE Access*, 2025.
- [2] Y. H. Yi, C. Kim, C. Zhou, A. Kteyan, and V. Sukharev, "Studying the impact of temperature gradient on electromigration lifetime using a power grid test structure with on-chip heaters," in *2023 IEEE International Reliability Physics Symposium (IRPS)*. IEEE, 2023, pp. 1–5.
- [3] D. Huang, L. Costero, and D. Atienza, "An Evaluation Framework for Dynamic Thermal Management Strategies in 3D MultiProcessor System-on-Chip Co-Design," *IEEE Transactions on Parallel and Distributed Systems*, 2024.
- [4] L. Jiang, Y. Liu, and M.-C. Cheng, "Fast-accurate full-chip dynamic thermal simulation with fine resolution enabled by a learning method," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 42, no. 8, pp. 2675–2688, 2022.
- [5] T. Kim, J. Lee, J. Kim, E.-C. Lee, H. Hwang, Y. Kim, and D. K. S. Oh, "Thermal modeling and analysis of high bandwidth memory in 2.5 d si-interposer systems," in *2022 21st IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm)*. IEEE, 2022, pp. 1–5.
- [6] C. Wang, Q. Xu, C. Nie, H. Cao, J. Liu, D. Zhang, and Z. Li, "A multiscale anisotropic thermal model of chiplet heterogeneous integration system," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 32, no. 1, pp. 178–189, 2023.
- [7] G. Karfakis, M. Bouzidi, Y. Im, A. Graening, S. K. Sitaraman, and P. Gupta, "Optimizing thermal performance in 2.5 d systems using embedded isolators," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2025.
- [8] Z. Yuan, P. Shukla, S. Chetoui, S. Nemtsov, S. Reda, and A. K. Coskun, "PACT: An extensible parallel thermal simulator for emerging integration and cooling technologies," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 4, pp. 1048–1061, 2021.
- [9] L. Pfromm, A. Kanani, H. Sharma, P. Solanki, E. Tervo, J. Park, J. R. Doppa, P. P. Pande, and U. Y. Ogras, "Mfit: Multi-fidelity thermal modeling for 2.5 d and 3d multi-chiplet architectures," *arXiv preprint arXiv:2410.09188*, 2024.
- [10] J.-H. Han, X. Guo, K. Skadron, and M. R. Stan, "From 2.5 d to 3d chiplet systems: Investigation of thermal implications with hotspot 7.0," in *2022 21st IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm)*. IEEE, 2022, pp. 1–6.
- [11] F. Terraneo, A. Leva, W. Fornaciari, M. Zapater, and D. Atienza, "3d-ice 3.0: efficient nonlinear mpsoe thermal simulation with pluggable heat sink models," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 4, pp. 1062–1075, 2021.
- [12] Q. Wang, X. Li, T. Jia, Y. Lin, R. Wang, and R. Huang, "Atplace2. 5d: Analytical thermal-aware chiplet placement framework for large-scale 2.5 d-ic," in *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, 2024, pp. 1–9.
- [13] E. Ayguadé, N. Copt, A. Duran, J. Hoeflinger, Y. Lin, F. Massaioli, X. Teruel, P. Unnikrishnan, and G. Zhang, "The design of openmp tasks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 20, no. 3, pp. 404–418, 2008.
- [14] X. S. Li, J. Demmel, J. Gilbert, L. Grigori, and M. Shao, "Superlu," in *Encyclopedia of Parallel Computing*. Springer, 2011, pp. 1955–1962.
- [15] D. Atienza, K. Zhu, D. Huang, and L. Costero, "A 20-year retrospective on power and thermal modeling and management," *IEEE Design & Test*, 2025.
- [16] H. Sultan and S. R. Sarangi, "A fast leakage-aware green's-function-based thermal simulator for 3-d chips," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 28, no. 11, pp. 2342–2355, 2020.
- [17] *Ansys Fluent User's Guide*, Ansys, Inc., Canonsburg, PA, USA, 2025, available online: https://ansyshelp.ansys.com/public/account/secured?returnurl=%2FViews%2FSecured%2Fprod_page.html%3Fpn%3DFluent.
- [18] *COMSOL Multiphysics Reference Manual*, COMSOL AB, Stockholm, Sweden, 2024, available online: https://doc.comsol.com/6.3/doc/com.comsol.help.comsol/html_COMSOL_ReferenceManual.html.
- [19] V. A. Chhabria, V. Ahuja, A. Prabhu, N. Patil, P. Jain, and S. S. Sapatnekar, "Thermal and ir drop analysis using convolutional encoder-decoder networks," in *Proceedings of the 26th Asia and South Pacific Design Automation Conference*, 2021, pp. 690–696.
- [20] L. Chen, W. Jin, and S. X.-D. Tan, "Fast thermal analysis for chiplet design based on graph convolution networks," in *2022 27th Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE, 2022, pp. 485–492.
- [21] Z. Liu, Y. Li, J. Hu, X. Yu, S. Shiao, X. Ai, Z. Zeng, and Z. Zhang, "Deepoheat: operator learning-based ultra-fast thermal simulation in 3d-ic design," in *2023 60th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 2023, pp. 1–6.
- [22] Y. Safari, A. Corbier, D. Al Saleh, F. R. Amik, and B. Vaisband, "Thermal simulator for advanced packaging and chiplet-based systems," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2025.
- [23] C. Nie, Q. Xu, and L. Chen, "Equivalent thermal model of through silicon via and bump for advanced packaging of integrated circuits," *Microelectronics Reliability*, vol. 137, p. 114790, 2022.
- [24] Ç. Koroğlu and E. Pop, "High thermal conductivity insulators for thermal management in 3d integrated circuits," *IEEE Electron Device Letters*, vol. 44, no. 3, pp. 496–499, 2023.
- [25] A. Hankin, D. Werner, M. Amiraski, J. Sebot, K. Vaidyanathan, and M. Hempstead, "Hotgauge: A methodology for characterizing advanced hotspots in modern and next generation processors," in *2021 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, 2021, pp. 163–175.
- [26] P. Scheffler, T. Benz, V. Potocnik, T. Fischer, L. Colagrande, N. Wistoff, Y. Zhang, L. Bertaccini, G. Ottavi, M. Eggmann *et al.*, "Occamy: A 432-core dual-chiplet dual-hbm2e 768-dp-gflop/s risc-v system for 8-to-64-bit dense and sparse computing in 12-nm finfet," *IEEE Journal of Solid-State Circuits*, 2025.
- [27] A. Hammoud, V. Shankar, R. Mains, T. Ansell, J. Matres, and M. Saligane, "Openfasoc: An open platform towards analog and mixed-signal automation and acceleration of chip design," in *2023 International Symposium on Devices, Circuits and Systems (ISDCS)*, vol. 1. IEEE, 2023, pp. 01–04.