ETLA-3D: Equivalent Thin Layer Aggregation based Thermal FEM for Hybrid Bonding F2F 3D ICs

Chenghan Wang[1], Zhen Zhuang[1*], Kai Zhu[2], Darong Huang[2], Luis Costero[3],
Rongmei Chen [4*], David Atienza[2], Tsung-Yi Ho[1]
[1] The Chinese University of Hong Kong [2] École Polytechnique Fédérale de Lausanne (EPFL)
[3] Universidad Complutense de Madrid [4] Peking University
∗ Corresponding Authors: zzhuang@link.cuhk.edu.hk, crm@pku.edu.cn

***Abstract*—In 3D face-to-face (F2F) hybrid bonding ICs, sub-micrometer thin layers lead to an extreme aspect ratio between the lateral dimensions and the vertical thickness. This poses major challenges for finite element method (FEM) thermal simulation. To address this, we introduce ETLA-3D, a thermal FEM methodology based on equivalent thin-layer aggregation, designed specifically for hybrid bonding F2F 3D ICs. The method consolidates the physical properties of thin layers into their neighboring layers by introducing new integral terms into the FEM weak form, greatly reducing the complexity of meshing, the simulation degrees of freedom (DoFs) and the computational cost, while preserving accuracy. Experimental results show that ETLA-3D achieves up to 695.8× faster runtime compared to the commercial FEM tool (COMSOL Multiphysics), with a maximum absolute error of less than 1.1 °C. By combining high accuracy with exceptional efficiency, ETLA-3D establishes a reliable and efficient FEM framework to model the thermal behavior of F2F 3D ICs.**

***Index Terms*—hybrid bonding, F2F 3D ICs, thermal simulation, finite element method, thin layer, model order reduction (MOR)**

I. Introduction

Today, the scaling of transistors has reached a significant bottleneck. To achieve More-than-Moore (MtM), 3D integration has emerged as one of the most promising solutions [1]. Existing 3D integration approaches include through-silicon vias (TSVs), microbumping, and hybrid bonding [2]. Among these, hybrid bonding technology directly connects the back-end-of-line (BEOL) layers of two stacked tiers using sub-µm Cu-Cu metal pads in a face-to-face (F2F) style. Due to its smaller pad size (merely hundreds of nanometers [2]), hybrid bonding can achieve a finer pitch and therefore higher integration densities compared to TSV-based and microbumping 3D ICs [3].

Compared with conventional 2D/2.5D ICs, 3D integration introduces more urgent thermal challenges [4]. Addressing these thermal challenges requires accurate and efficient thermal simulation. Existing thermal simulation approaches can be roughly categorized into three types: machine learning (ML)-based methods, stochastic methods, and numerical methods. Among these, FEM is the mainstream choice in commercial simulation tools, such as ANSYS (acquired by Synopsys) [5] and COMSOL Multiphysics [6], due to its flexibility and accuracy.

However, the unique structural characteristics of hybrid bonding 3D ICs pose significant challenges for FEM thermal simulation. A key feature of hybrid bonding is its sub-µm thick-

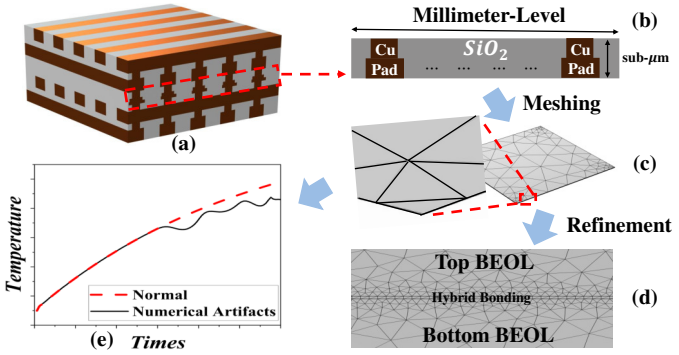

Fig. 1: (a) The BEOL and bonding structure. (b) The extreme aspect ratio of the bonding layer. (c) Coarse-grained mesh of the bonding layer without refinement, which can cause simulation issues. (d) Fine mesh after refinement, leading to huge simulation DoFs and computation cost. (e) Thin unhealthy FEM cells lead to simulation accuracy degradation.

ness. When these thin layers span millimeter-level horizontal dimensions, they create an extreme aspect ratio (exceeding 2000). Such an extreme aspect ratio renders FEM thermal simulation inefficient for four main reasons. First, maintaining mesh quality and ensuring simulation accuracy require iterative mesh refinement in these thin regions, as an extreme element aspect ratio can lead to mesh quality degradation and so-called 'unhealthy' cells [7], [8], causing undesired numerical artifacts. Second, the necessary conformal mesh topology at the interfaces between different geometric domains forces the creation of unnecessarily small cells in adjacent regions. For example, the shared topological nodes between the thin bonding layer and the BEOL layers result in fine meshing within BEOL regions. Third, fine meshing ultimately leads to an enormous number of Degrees of Freedoms (DoFs), which translates directly into a large coefficient matrix and therefore low efficiency for FEM solvers. Such computational inefficiency, where thin layers require extremely fine FEM meshes, has been reflected in various studies [3], [9], [10], [11], [12]. Finally, thin layers introduce significant numerical stiffness for some transient solvers, leading to extremely small time-step sizes and thus overlong runtimes [13]. These limitations motivate the need for a thermal FEM framework that retains the simulation accuracy while mitigating the computational burden.

The contact homogenization technique has been adopted in physical modeling, where thin layers are often simplified as an impedance interface condition [14], [15]. However, this simpli-

978-3-9826741-1-7/DATE26/© 2026 EDAA

fication typically focuses on resistance instead of the inherent heat capacitance and potential heat sources within these thin layers. Furthermore, a similar homogenization approach has not yet been customized for hybrid bonding F2F 3D ICs, taking into account their distinctive structural characteristics, and the quantification of accuracy loss remains unaddressed.

To address these challenges, we propose ETLA-3D, an equivalent thin layer aggregation-based thermal FEM methodology for hybrid bonding F2F 3D ICs. The key idea is to avoid explicit geometric modeling of thin layers while retaining their physical influence on heat transfer. As a result, ETLA-3D significantly reduces the meshing complexity, simulation DoFs and computation cost without compromising simulation accuracy. The key contributions of this work are as follows:

- We analyze and point out the numerical challenges caused by thin layers within hybrid bonding F2F 3D ICs, including meshing issues, high computation costs, and numerical stiffness in transient simulation.
- We propose **ETLA-3D**, an equivalent thin layer aggregation-based thermal FEM methodology. To the best of our knowledge, ETLA-3D is the first thermal FEM implementation of hybrid bonding F2F 3D ICs simulation using an entire open-source FEM toolchain, capable of both steady-state and transient thermal simulation.
- We observed that vertical heat flux is dominant within thin layers. Therefore, passive thin layers are incorporated into the FEM weak form as a Robin-like surface integral term over the interface cells of its two vertically adjacent layers, thus avoiding fine-grained bulk meshing.
- Unlike passive thin layers, active layers also serve as bulk heat sources. ETLA-3D transforms them into an equivalent surface forms integrated into the FEM weak form as a Neumann surface integral term over the interface cells.
- Extensive experimental results demonstrate that ETLA-3D is an efficient second-level simulation tool, achieving up to 695.8× faster runtime compared to COMSOL Multiphysics, while keeping simulation errors < 1.1°C.

II. Related Works

Existing thermal simulation approaches can be roughly categorized into three types: numerical methods, stochastic methods, and machine learning (ML)-based methods.

Numerical methods, which encompass the finite element method (FEM) [16], the finite difference method (FDM) [17], the finite volume method (FVM) [18], and equivalent compact models of the RC network [13], [19], [20], are based on classic physical principles. These methods offer superior robustness and generality. Among them, FEM stands out as one of the most widely used techniques, widely utilized for both early-stage estimation [16], [21] and post-layout sign-off simulation [22], [23]. However, numerical methods are inherently mesh-based and are challenged by the aforementioned issues arising from thin layers, such as meshing complexity, huge simulation DoFs, significant computation costs, and numerical stiffness.

Stochastic methods, such as random walk (RW)-based thermal models [24], [25], are efficient for predicting temperature values of single points but struggle with complex boundary conditions (BCs), especially Robin BCs. They are also inefficient when a full temperature distribution is required.

ML-based methods, including convolutional neural networks (CNNs) [26], graph neural networks (GNNs) [27], physics-informed neural networks (PINNs) [28], [29], and operator learning (OL) [30], [31], have attracted a great deal of attention due to their potential for acceleration. However, data-driven ML models are highly dependent on high-quality ground-truth data, which are typically generated by numerical methods. As discussed previously, current numerical methods struggle with issues introduced by thin layers, which can lead to low-quality training data and prohibitively slow data generation. Consequently, an efficient numerical tool is necessary for ML models, which also motivates the development of ETLA-3D.

III. Preliminaries

A. Governing Equations

In hybrid bonding F2F 3D ICs, the governing PDE of time-dependent heat transfer is [32]:

$$c\frac{\partial u}{\partial t} + \nabla \cdot (-k\nabla u) = f \; in \; \Omega \tag{1}$$

Here, Ω is the computation domain, c is the volumetric heat capacity (J/m^3K), k is the thermal conductivity (W/mK), f is the power density, or heat source term (W/m^3), and $u(x, y, z, t)$ is the space- and time-dependent temperature function, which is the unknown function of the PDE.

The governing PDE requires boundary conditions for a unique solution. Typical BCs of heat transfer include: Dirichlet, Neumann, and Robin conditions. An adiabatic Neumann boundary condition is applied to the computation boundary Γ_N, where no air convection occurs:

$$\frac{\partial u}{\partial \mathbf{n}} = 0 \; on \; \Gamma_N \tag{2}$$

A convective Robin boundary condition is applied to the computation domain Γ_R, where air convection removes heat from the system. This phenomenon can be represented as:

$$\frac{\partial u}{\partial \mathbf{n}} = -h(u - u_{amb}) \tag{3}$$

where h is the heat transfer coefficient (W/m^2K) and u_{amb} is the ambient air temperature. For steady-state thermal simulation, where the temperature function $u(x, y, z)$ is time-independent, the temporal derivative term in (1) becomes zero. In this case, (1) simplifies to a Poisson equation.

B. FEM Weak Form

Given its accuracy and flexibility, FEM is adopted to solve (1)-(3) in this paper. To effectively apply FEM, we must first convert their strong form into the corresponding weak form.

First, (1) is multiplied by a test function v and then integrated over the entire computation domain Ω, yielding:

$$\int_\Omega c\frac{\partial u}{\partial t} v \, d\Omega + \int_\Omega \nabla \cdot (-k\nabla u) v \, d\Omega = \int_\Omega f v \, d\Omega \tag{4}$$

To lower the derivative order, the integration by parts and divergence theorem can be used to replace the Laplacian term in (4), resulting in:

$$\int_\Omega c\frac{\partial u}{\partial t}v\,d\Omega+\int_\Omega k\nabla u\nabla v\,d\Omega+\int_\Gamma -k\nabla uv\cdot\mathbf{n}\,dS=\int_\Omega fv\,d\Omega \tag{5}$$

where dS is the surface micro-unit. Considering the Neumann and Robin boundary condition, (5) can be transformed into:

$$\begin{aligned}&\int_\Omega c\frac{\partial u}{\partial t}v\,d\Omega+\int_\Omega k\nabla u\nabla v\,d\Omega+\int_{\Gamma_R} huv\,dS\\&=\int_\Omega fv\,d\Omega+\int_{\Gamma_R} hu_{amb}v\,dS\end{aligned} \tag{6}$$

The thin layers of hybrid bonding F2F 3D ICs lead to high stiffness in transient thermal simulation[13]. Therefore, we employ the implicit Euler method for temporal discretization, as it is both A- and L-stable, which means it is unconditionally stable and can prevent undesired numerical artifacts [33].

$$\int_\Omega c\frac{\partial u}{\partial t}v\,d\Omega\approx\int_\Omega c\frac{u_{n+1}-u_n}{\Delta t}v\,d\Omega \tag{7}$$

where Δt is the time step of the transient thermal simulation. Finally, the weak form of the heat PDE, equipped with Neumann and Robin boundary conditions, can be expressed as:

$$\begin{aligned}&\int_\Omega c\frac{u_{n+1}}{\Delta t}v\,d\Omega+\int_\Omega k\nabla u_{n+1}\nabla v\,d\Omega+\int_{\Gamma_R} hu_{n+1}v\,dS\\&=\int_\Omega f_{n+1}v\,d\Omega+\int_{\Gamma_R} hu_{amb}v\,dS+\int_\Omega c\frac{u_n}{\Delta t}v\,d\Omega\end{aligned} \tag{8}$$

IV. ETLA-3D Methodology

This section details the technical methodology of ETLA-3D. First, we present the overall framework in Section A. Subsequently, Section B gives the core assumption of ETLA-3D. The aggregation of passive and active thin layers is presented in Subsections C and D, respectively. Finally, Subsection E gives two distinct aggregation strategies from conservative to aggressive, allowing users to balance accuracy and efficiency.

A. Overall Framework

Fig. 2 illustrates the schematic of a flipped hybrid bonding F2F Memory-on-Logic (MoL) 3D IC [34]. The bottom package is not included because only 5% heat is dissipated through the bottom heat path [35]. The die stack consists of seven distinct layers [36]: logic bulk (thinned due to embedded TSVs), logic FEOL, logic BEOL, hybrid bonding, memory BEOL, memory FEOL and memory bulk. Among these, thin layers include the sub-µm hybrid bonding layer (passive) and two FEOL layers (active). These thin layers create significant challenges, including complex meshing and high numerical stiffness. Therefore, ETLA-3D is designed to avoid directly modeling these thin layers, while still preserving their physical effect on heat transfer.

As depicted in Fig. 3, our approach merges these thin layers into their neighboring, thicker layers. ETLA-3D aggregates passive and active layers using different equivalent models, and offers two distinct aggregation strategies as Mode 1 and Mode 2 to balance between accuracy and efficiency.

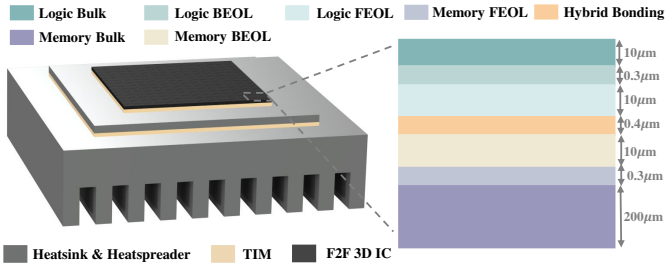

Fig. 2: The schematic of a flipped hybrid bonding F2F Memory-on-Logic (MoL) 3D IC. The F2F 3D IC includes seven distinct layers: two bulk layers, two FEOL layers, two BEOL layers, and one bonding layer. Among these, two FEOL layers and bonding layer are thin layers below 1 µm.

B. The Core Assumption

The two modes of ETLA-3D are built on a core assumption: within these thin layers, vertical heat transfer is the dominant part of heat conduction, while horizontal heat conduction can be considered negligible. This core assumption is valid when 1) layers are extremely thin, 2) power value gradients are not significant.

The reason for 1) is that the extremely small thickness, which means extremely small XZ and YZ cross-sections, leads to massive horizontal thermal resistance. As a result, heat flows almost exclusively in the vertical direction, making the horizontal component of heat transfer negligible.

The reason for 2) is that a great gradient of power maps directly leads to a great in-plane temperature gradient within thin layers, thus offering a driving force for horizontal heat transfer.

The validity of this assumption is further supported by the experimental results presented in the next section, where insignificant horizontal heat fluxes are found.

C. Aggregation of Passive Thin Layers

Passive thin layers, such as the sub-µm hybrid bonding layers, perform two critical physical functions in heat transfer: 1) a thermal resistor between two BEOL layers, 2) a thermal capacitor that stores and discharges heat. Therefore, the aggregation of these passive thin layers must preserve the two functions to avoid information loss.

First, the thermal resistance can be replaced by an equivalent thermal contact surface resistance, $R_{eq}(Km^2/W)$, applied to the interface between the logic and the BEOL layers of the memory. The equivalent surface resistance is calculated as follows:

$$R_{eq}=\frac{d_{HB}}{k_{HB}} \tag{9}$$

where d_{HB} and k_{HB} are the thickness of the hybrid bonding layer and the thermal conductivity, respectively.

Then, the thermal capacitance is divided equally and transferred to the two adjacent BEOL layers. Therefore, the equivalent volumetric heat capacity of each BEOL layer, $c_{BEOL,eq}$, is calculated as follows:

$$c_{BEOL,eq}=\frac{C_{BEOL,eq}}{V_{BEOL}}=c_{BEOL}+\frac{d_{HB}c_{HB}}{2d_{BEOL}} \tag{10}$$

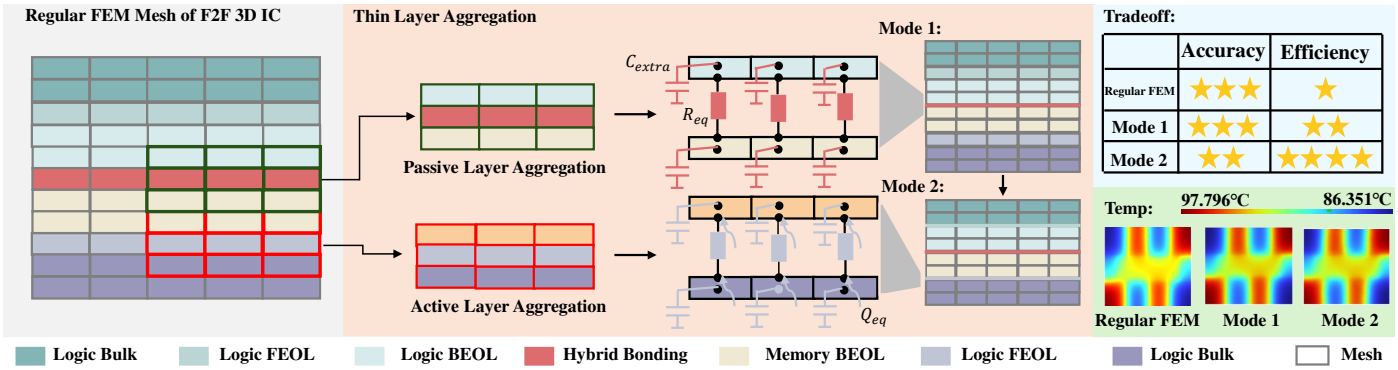

Fig. 3: ETLA-3D can merge both passive and active thin layers into adjacent by replacing them with their interface models connecting their two neighboring layers. ETLA-3D offers two distinct aggregation strategies: Mode 1 (conservative) and Mode 2 (aggressive). Mode 1 offers high accuracy with moderate efficiency, while Mode 2 offers superior efficiency at the cost of slight simulation errors.

where V_{BEOL} is the physical volume of each BEOL layer. Therefore, the equivalent model of Mode 1 is depicted in Fig. 3.

To implement the equivalent model in FEM, the weak form of each BEOL layer (represented by B) should include an extra Robin-like surface integral term, modified from (8) to:

$$\int_{\Omega_B} c_{B,eq}\frac{u_{n+1}}{\Delta t}v\,d\Omega + \int_{\Omega_B} k_B\nabla u_{n+1}\nabla v\,d\Omega + \int_{\Gamma_I}\frac{u_{n+1}}{R_{eq}}v\,dS = \int_{\Gamma_I}\frac{u_{n+1}(-)}{R_{eq}}v\,dS + \int_{\Omega_B} c_{B,eq}\frac{u_n}{\Delta t}v\,d\Omega \tag{11}$$

where Γ_I is the interface of the two BEOL layers when the hybrid bonding layer is removed in the geometry domain, $u_{n+1}(-)$ is the temperature of the other side of the interface. For example, if the integral domain is the logic BEOL layer, $u_{n+1}(-)$ would be the temperature of the bottom surface of the memory BEOL layer, and vice versa. As implied in (11), R_{eq} is applied using a Robin-like integral term with the heat transfer coefficient h replaced by the reciprocal of R_{eq}.

In this way, passive layers can be removed in the geometric domain while their physical influences are preserved as a Robin-like surface integral term over the interface cells. Therefore, fine-grained bulk meshing of passive thin layers can be avoided, effectively reducing the computational cost.

D. Aggregation of Active Thin Layers

Unlike passive thin layers, active layers, such as FEOL layers, serve a third physical function: heat sources. Therefore, the power maps of the active layers are equally divided into two equivalent surface heat fluxes, Q_{eq} (W/m^2), flowing into their adjacent layers:

$$Q_{eq} = \frac{P_{FEOL}d_{FEOL}}{2} \tag{12}$$

Here, P_{FEOL} is the original volumetric power density (W/m^3) of power maps. Their equivalence model is illustrated in Fig. 3. Its equivalent surface contact thermal resistance, R_{eq}, is calculated using the same formula as (9), with the subscripts replaced by FEOL. Similarly, the thermal capacitance of each adjacent layer, $C_{BEOL,eq}$ and $C_{bulk,eq}$, must be increased by half of the thermal capacitance of the FEOL layer, and thus the equivalent volumetric heat capacitance, $c_{BEOL,eq}$ and $c_{bulk,eq}$ are obtained by dividing the thermal capacitance with the physical volume.

To clarify the corresponding FEM weak form, the logic bulk (LB) layer is taken as an example. Considering R_{eq}, $c_{bulk,eq}$, and Q_{eq}, its FEM weak form should incorporate an extra Neumann surface integral term, modified from (8) to:

$$\begin{aligned}&\int_{\Omega_{LB}} c_{LB,eq}\frac{u_{n+1}}{\Delta t}v\,d\Omega + \int_{\Omega_{LB}} k\nabla u_{n+1}\nabla v\,d\Omega \\ &+\int_{\Gamma_I}\frac{u_{n+1}}{R_{eq}}v\,dS = \int_{\Gamma_I}\frac{u_{n+1}(-)}{R_{eq}}v\,dS \\ &+\int_{\Omega_{LB}} c_{LB,eq}\frac{u_n}{\Delta t}v\,d\Omega + \int_{\Gamma_I} Q_{eq}v\,dS\end{aligned} \tag{13}$$

where the script LB represents the logic bulk layer, $u_{n+1}(-)$ is the temperature of the bottom surface of the logic BEOL layer, Γ_I is the interface between the logic bulk layer and the logic BEOL layer when the logic FEOL is removed from the geometry domain. The weak form of the memory bulk FEM is the same formula as in (13) with the subscript replaced by MB. As for the two BEOL layers, an extra R_{eq} of the FEOl layer, and Q_{eq} should be added into (11).

Through transforming bulk heat sources into their equivalent surface forms integrated into the FEM weak form, the active layers can also be aggregated, further reducing mesh density and thus computation cost.

E. The Two-Mode Design of ETLA-3D

ETLA-3D offers two distinct strategies for aggregating these sub-µm thin layers, varying in their scope from a conservative to a more aggressive approach. Conservatively, Mode 1 only merges the passive hybrid bonding layer to its neighboring layers, as shown in Fig. 3. This strategy primarily aims to prevent direct bulk meshing of hybrid bonding layers, thereby reducing meshing complexity, simulation DoFs and cost to a moderate extent, while introducing minimal simulation errors.

More aggressively, Mode 2 extends this by aggregating not only the hybrid bonding layer but also the logic and memory FEOL layers. This broader aggregation significantly reduces mesh complexity, simulation DoFs, and computational cost. However, while Mode 2 introduces relatively larger simulation errors than Mode 1, they remain within acceptable limits.

As shown in Fig. 3, the dual-mode design of ETLA-3D allows users to balance simulation accuracy and computation efficiency as needed.

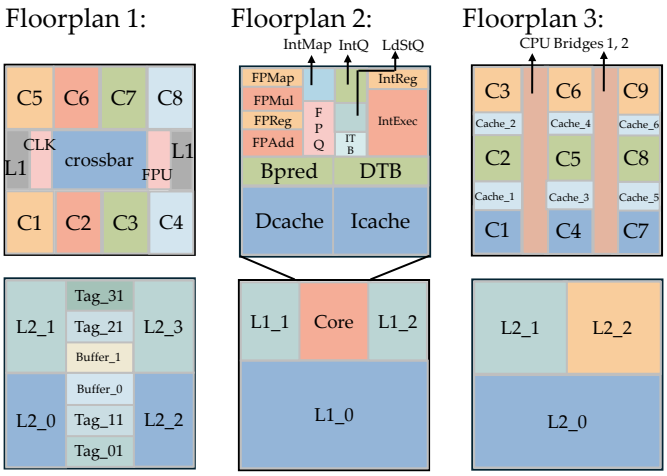

Fig. 4: Three MoL floorplans: Logic floorplan 1 is the SPARC processor, logic floorplan 2 is the EV6 processor with the core magnified, and logic floorplan 3 is a simplified multi-core processor. Floorplans 2 and 3 share the same 3-cache memory (right bottom).

V. Experiments

A. FEM Toolchain Implementation

ETLA-3D is implemented leveraging an entire open-source FEM toolchain: Gmsh [37] is utilized for FEM pre-processing, including geometry modeling and mesh generation. With the weak form of (8), (11), and (13), FEniCSx performs FEM thermal simulation [38], and Paraview [39] is employed for FEM post-processing. Experiments are conducted in a Linux system with an Intel Ultra7 265K CPU and a 64GB RAM.

B. Experimental Setup

Fig. 2 illustrates the schematic of the hybrid bonding F2F MoL 3D IC. The die stack's horizontal dimensions are set to 500 µm (mobile-level). The heat spreader is twice the size of the die stack and the heat sink is twice the size of the heat spreader. The thickness of TIM, heat spreader, and heat sink are 100, 300, and 1000 µm, respectively. The top surface of the heat sink is defined as Γ_R, with $h = 1000 W/m^2K$.

The two FEOL layers are designated as heat sources. As shown in Fig. 4, three distinct floorplan couples are employed, which have been widely utilized in [19], [20], [31], [40].

Four baselines are employed to evaluate the performance of ETLA-3D: COMSOL Multiphysics [6], considered as the ground truth, the original FEniCSx without thin layer equivalence [38], HotSpot 7.0 [19], and 3D-ICE 3.1 [20]. The grid resolution of HotSpot is 128×128. Non-uniform grid is adopted in 3D-ICE. While HotSpot and 3D-ICE utilize direct sparse solvers, they are also employed in COMSOL, the original FEniCSx, and ETLA-3D to ensure fairness.

C. Results of Steady-State Simulation

Table I presents the performance comparison in steady-state thermal simulation. Simulation precision is evaluated using maximum absolute errors (MaxAE, °C) and mean absolute errors (MAE, °C). Simulation efficiency is assessed by degrees of freedom (DoFs) and simulation runtime.

Experimental results indicate that the MaxAE and MAEs of Mode 1 are almost the same as the original FEniCSx, suggesting that Mode 1 only introduces negligible errors ($\leq 10^{-2}$°C).

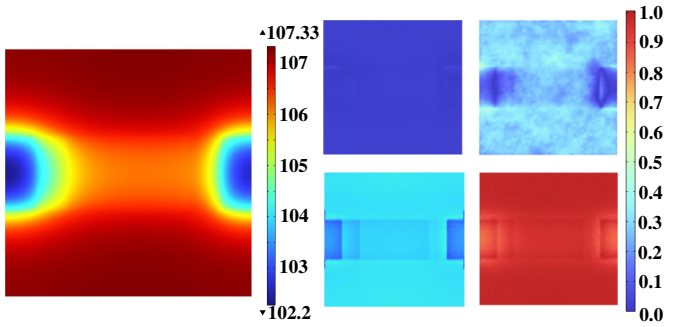

Fig. 5: The accuracy performance of ETLA-3D, HotSpot 7.0 and 3D-ICE 3.1. Left: The temperature map of the logic FEOL layer (by COMSOL Multiphysics as the golden reference). Right: The error map of Mode 1 (top left), Mode 2 (top right), 3D-ICE 3.1 (bottom left), and HotSpot 7.0 (bottom right).

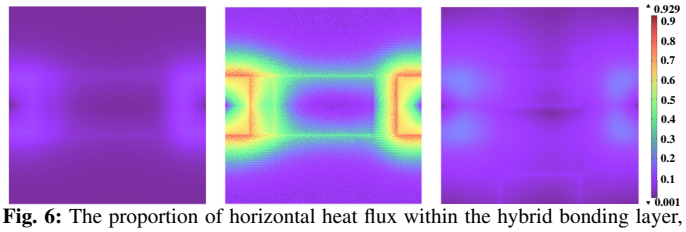

Fig. 6: The proportion of horizontal heat flux within the hybrid bonding layer, logic FEOL layer, memory FEOL layer (from left to right).

For efficiency, the DoFs are reduced from 1,427,259 to 804,288 (a 43.6% reduction), resulting in a nearly 3× speedup over COMSOL. Mode 2 exhibits superior efficiency: the DoFs are reduced from 1,427,259 to mere 29,591 (a 97.9% reduction), yielding a nearly 80× speedup. Crucially, the MaxAE of Mode 2 maintained below 1.1°C. The error maps are shown in Fig. 5. Compared with HotSpot and 3D-ICE, ETLA-3D demonstrates superior performance in both MAE and runtime.

Subsection IV.B states ETLA-3D is valid when the vertical heat transfer is dominant. To further investigate this, the proportion of horizontal (in-plane) heat flux within the hybrid bonding layer, logic FEOL layer, and memory FEOL layer are illustrated in Fig. 6, where the horizontal part is found to be insignificant within hybrid bonding layer (2.16% on average) and memory FEOL layer (1.37% on average). While horizontal part is generally insignificant in most regions of the logic FEOL layer, its average proportion is 21.6%. Crucially, there are still specific regions within the logic FEOL layer where horizontal heat transfer becomes dominant, particularly at the interfaces between functional blocks. This phenomenon arises because significant power value gradients at these interfaces lead to massive local thermal gradients, providing a driving force for horizontal heat transfer.

D. Results of Transient Simulation

Table II presents the performance comparison in transient simulation. The transient simulation was conducted for a duration of 3 seconds, with a constant time interval Δt = 0.03s. MaxAE is defined as the maximum value among the 100 recorded absolute errors, and MAE is their mean value.

Compared with COMSOL, Mode 1 achieves an nearly 20× speedup. Notably, Mode 2 achieves second-level runtimes and an approximate 700× speedup at most, while maintaining MaxAE below 0.5°C. However, HotSpot failed to complete the

TABLE I: The performance comparison in steady-state simulation. COMSOL Multiphysics is considered as the golden results. The DoFs for COMSOL, FEniCSx, HotSpot 7.0, 3D-ICE 3.1 (non-uniform mode), Mode 1, and Mode 2 are: 1401201, 1427259, 163852, 38000/35600/34400, 804288, 29591, respectively.

Floorplans	Powermaps	COMSOL[6]	FEniCSx [38]			HotSpot 7.0 [19]			3D-ICE 3.1 [20]			Mode 1 of ETLA-3D			Mode 2 of ETLA-3D		
		Runtime(s)	MaxAE	MAE	Runtime(s)	MaxAE	MAE	Runtime(s)	MaxAE	MAE	Runtime(s)	MaxAE	MAE	Runtime(s)	MaxAE	MAE	Runtime(s)
SPARC	1	63 (1.0×)	**0.05**	0.02	31.50 (2.0×)	1.01	0.97	9.11 (6.9×)	0.40	0.34	6.85 (9.2×)	0.06	**0.02**	20.06 (3.1×)	0.51	0.27	**0.77 (81.8×)**
	2	59 (1.0×)	0.16	0.05	32.34 (1.8×)	0.99	0.80	9.06 (6.5×)	0.53	0.28	6.71 (8.8×)	**0.16**	**0.05**	20.66 (2.9×)	1.09	0.25	**0.74 (79.7×)**
	3	61 (1.0×)	0.19	0.04	33.94 (1.8×)	0.74	0.59	9.86 (6.2×)	0.47	0.21	6.73 (9.1×)	**0.19**	**0.05**	20.37 (3.0×)	0.94	0.23	**0.81 (75.3×)**
EV6	1	61 (1.0×)	0.05	0.03	31.42 (1.9×)	0.80	0.61	8.64 (7.1×)	0.88	0.24	5.42 (11.3×)	**0.05**	**0.03**	19.55 (3.1×)	0.63	0.07	**0.74 (82.4×)**
	2	58 (1.0×)	0.05	0.03	34.10 (1.7×)	0.84	0.69	8.26 (7.0×)	0.63	0.25	5.39 (10.8×)	**0.05**	**0.03**	19.55 (3.0×)	0.63	0.13	**0.74 (78.4×)**
	3	62 (1.0×)	0.05	0.03	33.42 (1.9×)	0.84	0.69	8.15 (7.6×)	0.90	0.25	5.48 (11.3×)	**0.05**	**0.03**	20.02 (3.1×)	0.48	0.04	**0.76 (81.6×)**
Multi-core	1	61 (1.0×)	0.13	0.05	32.13 (1.9×)	1.01	0.94	6.67 (9.1×)	0.51	0.36	5.16 (11.8×)	**0.13**	**0.05**	19.72 (3.1×)	0.83	0.49	**0.83 (73.5×)**
Ratio		-	1.0	1.0	-	9.8	19.0	-	6.2	6.9	-	**1.0**	**1.0**	-	7.3	5.3	-

TABLE II: The performance comparison in transient simulation. Transient thermal simulation was conducted for a duration of 3 seconds. The constant time interval Δt = 0.03s. The DoFs of COMSOL, FEniCSx, HotSpot 7.0, 3D-ICE 3.1 (non-uniform mode), Mode 1, and Mode 2 are the same as Table I.

Floorplans	Powermaps	COMSOL [6]	FEniCSx [38]			HotSpot 7.0 [19]			3D-ICE 3.1 [20]			Mode 1 of ETLA-3D			Mode 2 of ETLA-3D		
		Runtime(s)	MaxAE	MAE	Runtime(s)	MaxAE	MAE	Runtime(s)	MaxAE	MAE	Runtime(s)	MaxAE	MAE	Runtime(s)	MaxAE	MAE	Runtime(s)
SPARC	1	1542 (1.0×)	0.15	0.01	348.50 (4.4×)	N/A	N/A	>10,000	0.36	0.32	14.44 (106.8×)	**0.15**	**0.01**	72.34 (21.3×)	0.28	0.14	**2.34 (659.0×)**
	2	1496 (1.0×)	0.17	0.02	362.59 (4.1×)	N/A	N/A	>10,000	0.37	0.33	14.39 (104.0×)	**0.17**	**0.02**	76.23 (19.6×)	0.28	0.06	**2.15 (695.8×)**
	3	1523 (1.0×)	0.13	0.02	357.44 (4.3×)	N/A	N/A	>10,000	0.34	0.30	14.53 (104.8×)	**0.13**	**0.02**	74.74 (20.4×)	0.23	0.12	**2.25 (676.9×)**
EV6	1	1524 (1.0×)	0.13	0.01	365.82 (4.2×)	N/A	N/A	>10,000	0.30	0.28	12.14 (125.5×)	**0.13**	**0.01**	73.87 (20.6×)	0.29	0.17	**2.37 (643.0×)**
	2	1503 (1.0×)	0.14	0.02	373.74 (4.0×)	N/A	N/A	>10,000	0.35	0.32	12.26 (122.6×)	**0.14**	**0.02**	73.80 (20.4×)	0.45	0.34	**2.23 (674.0×)**
	3	1505 (1.0×)	0.12	0.01	384.10 (3.9×)	N/A	N/A	>10,000	0.37	0.33	12.17 (123.7×)	**0.12**	**0.01**	73.17 (20.6×)	0.23	0.09	**2.29 (657.2×)**
Multi-core	1	1502 (1.0×)	0.16	0.04	376.13 (4.0×)	N/A	N/A	>10,000	0.36	0.31	11.77 (127.6×)	**0.16**	**0.04**	69.72 (21.5 ×)	0.33	0.27	**2.20 (682.7×)**
Ratio		-	1.0	1.0	-	-	-	-	2.9	17.2	-	**1.0**	**1.0**	-	2.1	9.2	-

transient simulation within 10,000 seconds. This phenomenon is corroborated by findings in [13], where HotSpot required over 100,000 seconds to complete transient simulation of a 100-nm 2-layer chip stack. This extremely low efficiency is directly attributed to the presence of sub-µm thin layers. However, the underlying mechanism remains unclear in [13].

In fact, such inefficiency can be explained by numerical stiffness. HotSpot employs the explicit 4th-order Runge-Kutta (RK-4) method as the transient solver. However, such explicit transient solvers must adhere to the Courant-Friedrichs-Lewy (CFL) condition, which stipulates that their time step length must be sufficiently small to guarantee numerical convergence [41]. In the context of transient thermal simulation, the CFL condition can be expressed as:

$$\Delta t(\frac{\alpha}{\Delta x^2}+\frac{\alpha}{\Delta y^2}+\frac{\alpha}{\Delta z^2}) \approx \Delta t\frac{\alpha}{\Delta z^2} \leq 1 \qquad (14)$$

where α is thermal diffusivity (m^2/s), Δx, Δy, and Δz represent the cell width, height, and thickness, respectively. Sub-µm thin layers result in an extremely small Δz. To ensure numerical convergence, its maximum allowable step length is:

$$\Delta t_{max} \approx \frac{\Delta z^2}{\alpha_{silicon}} = \frac{C_{silicon}\Delta z^2}{k_{silicon}} \approx 1.05 \times 10^{-9} s \qquad (15)$$

Therefore, leveraging RK-4, HotSpot is compelled to advance with a nanosecond-level step length, resulting in its low efficiency. This issue can be avoided by the proposed aggregation technique, or implicit transient solvers, such as the backward Euler method used in other simulators [20].

E. Investigation onto Thickness Sensitivity

As previously stated, ETLA-3D is valid when the vertical heat conduction is dominant, which is true as long as layers are sufficiently thin. Therefore, we conduct an investigation of thickness sensitivity. As shown in Fig. 7, errors grows as the thickness increases. This is because thicker layers promote horizontal heat transfer and consequently undermine the validity of ETLA-3D. The MaxAE of Mode 2 reaches 1.01°C (HotSpot) when h is nearly 4.4 µm (h_1). The MAE reaches 0.97°C when h is nearly 7.5 µm (h_2). Consequently, Mode 2 is recommended when FEOL layer is thinner than h_1, at least h_2 to guarantee accuracy. However, FEOL layers are sub-µm thick in advanced nodes [35], far from the two thresholds h_1 and h_2. Therefore, ETLA-3D can be regarded valid for hybrid bonding F2F 3D ICs in advanced nodes without compromising simulation accuracy.

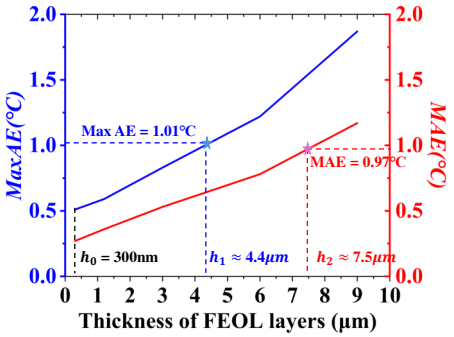

Fig. 7: Investigation on thickness sensitivity: The evolution of MaxAE and MAE as the FEOL thickness increases. The test case is SPARC 1 in Table I.

VI. Conclusion

In this work, we have highlighted the challenges posed by sub-micrometer thin layers in hybrid bonding F2F 3D ICs. To address these, we introduced ETLA-3D, an equivalent thin layer aggregation-based thermal FEM methodology. ETLA-3D offers two operating modes, allowing users to balance accuracy and efficiency as needed. Experimental results demonstrate that ETLA-3D achieves up to 695.8× speedup compared to COMSOL Multiphysics, while maintaining simulation errors below 1.1°C. Looking ahead, we plan to extend ETLA-3D to handle more complex BEOL layer structures.

Acknowledgment

The research work described in this paper was conducted in the JC STEM Lab of Intelligent Design Automation funded by The Hong Kong Jockey Club Charities Trust, and ACCESS-AI Chip Center for Emerging Smart Systems supported by the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government.

This work is jointly supported by the Research Grants Council of Hong Kong SAR (No. CUHK14211324), and the Swiss State Secretariat for Education, Research, and Innovation (SERI) through the SwissChips research project.

REFERENCES

[1] S. Liu, J. Jiang, Z. He, Z. Wang, Y. Lin, B. Yu, and M. Wong, "Routing-aware legal hybrid bonding terminal assignment for 3d face-to-face stacked ics," in *Proceedings of the 2024 International Symposium on Physical Design (ISPD)*, 2024, pp. 75–82.

[2] J. Kim, L. Zhu, H. M. Torun, M. Swaminathan, and S. K. Lim, "A ppa study for heterogeneous 3-d ic options: Monolithic, hybrid bonding, and microbumping," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 32, no. 3, pp. 401–412, 2024.

[3] J. H. Lau, "Recent advances and trends in cu–cu hybrid bonding," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 13, no. 3, pp. 399–425, 2023.

[4] S. Venkateswarlu, S. Mishra, H. Oprins, B. Vermeersch, M. Brunion, J.-H. Han, M. R. Stan, D. Biswas, P. Weckx, and F. Catthoor, "Impact of 3-d integration on thermal performance of risc-v mempool multicore soc," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 31, no. 12, pp. 1896–1904, 2023.

[5] Ansys. [Online]. Available: https://www.ansys.com

[6] COMSOL Multiphysics. [Online]. Available: https://www.comsol.com

[7] P. Chen, N. Niu, D. Zhang, W. Wang, D. Xie, Z. Jin, W. Xing, and L. He, "Neuralmesh: neural network for fem mesh generation in 2.5d/3d chiplet thermal simulation," in *2025 ACM/IEEE Design Automation Conference (DAC)*, 2025.

[8] D. Vartziotis, J. Wipper, and M. Papadrakakis, "Improving mesh quality and finite element solution accuracy by getme smoothing in solving the poisson equation," *Finite elements in analysis and design*, 2013.

[9] H. Oprins, V. Cherman, T. Webers, A. Salahouelhadj, S.-W. Kim, L. Peng, G. Van der Plas, and E. Beyne, "Thermal characterization of the inter-die thermal resistance of hybrid cu/dielectric wafer-to-wafer bonding," in *2016 15th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, 2016, pp. 1333–1339.

[10] V. Cherman, S. Van Huylenbroeck, M. Lofrano, X. Chang, H. Oprins, M. Gonzalez, G. Van der Plas, G. Beyer, K. J. Rebibis, and E. Beyne, "Thermal, mechanical and reliability assessment of hybrid bonded wafers, bonded at 2.5/textu m pitch," in *2020 IEEE 70th Electronic Components and Technology Conference (ECTC)*, 2020, pp. 548–553.

[11] H. Oprins, V. Cherman, T. Webers, S.-W. Kim, J. de Vos, G. Van der Plas, and E. Beyne, "3d wafer-to-wafer bonding thermal resistance comparison: Hybrid cu/dielectric bonding versus dielectric via-last bonding," in *2020 19th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, 2020, pp. 219–228.

[12] A. Pic, R. Prffito, S. Gallois-Garreignot, J.-P. Colonna, V. Fiori, and P.-O. Chapuis, "3d hybrid bonding assembly studied by scanning thermal microscopy, resistive thermometry and finite element modelling," in *2018 19th International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems (EuroSimE)*, 2018, pp. 1–6.

[13] Z. Yuan, P. Shukla, S. Chetoui, S. Nemtzow, 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, 2022.

[14] K. Schmidt and A. Chernov, "Robust transmission conditions of high order for thin conducting sheets in two dimensions," *IEEE transactions on magnetics*, vol. 50, no. 2, pp. 41–44, 2014.

[15] I. Woyna, E. Gjonaj, and T. Weiland, "Broadband surface impedance boundary conditions for higher order time domain discontinuous galerkin method," *COMPEL: The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, 2014.

[16] S. Ladenheim, Y.-C. Chen, M. Mihajlović, and V. F. Pavlidis, "The mta: An advanced and versatile thermal simulator for integrated systems," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 12, pp. 3123–3136, 2018.

[17] T.-Y. Wang and C. Chen, "Thermal-adi - a linear-time chip-level dynamic thermal-simulation algorithm based on alternating-direction-implicit (adi) method," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 11, no. 4, pp. 691–700, 2003.

[18] 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.

[19] J.-H. Han, X. Guo, K. Skadron, and M. R. Stan, "From 2.5d 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)*, 2022, pp. 1–6.

[20] 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*, vol. 35, no. 11, pp. 2161–2176, 2024.

[21] D.-W. Wang, B.-W. Zhang, L.-T. Wang, P. Zhang, and W.-S. Zhao, "A proposal of fast thermal simulation method for 2.5-d advanced packaging to enable efficient thermal-aware placement optimization," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, pp. 1–1, 2025.

[22] "Ansys red hawk." [Online]. Available: https://www.ansys.com/content/dam/amp/2022/july/webpage-requests/semiconductors/redhawk-sc-datasheet-2022.pdf

[23] "Cadence celsius solver," 2021. [Online]. Available: https://www.artedas.fr/documentsPDF/Produits/cadence-celsius-datasheet.pdf

[24] Z. Dong, L. Yang, C. Ding, C. Yan, Z. Bi, S.-G. Wang, D. Zhou, and X. Zeng, "ppirw: An efficient and accurate precalculation path integral random walk solver for steady-state thermal simulation with robin boundary conditions," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 44, no. 4, pp. 1489–1502, 2025.

[25] Y. Liang, W. Yu, and H. Qian, "A hybrid random walk algorithm for 3-d thermal analysis of integrated circuits," in *2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2014, pp. 849–854.

[26] 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 *2021 26th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2021, pp. 690–696.

[27] 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)*, 2022.

[28] L. Chen, J. Lu, W. Jin, and S. X.-D. Tan, "Fast full-chip parametric thermal analysis based on enhanced physics enforced neural networks," in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, 2023, pp. 1–8.

[29] C. Zhou, M. Tang, and L. Chen, "Asrr-pinn: adaptive sub-regional random resampling-based pinn for thermal analysis of 3d-ics," in *2025 ACM/IEEE Design Automation Conference (DAC)*, 2025.

[30] Z. Liu, Y. Li, J. Hu, X. Yu, S. Shiau, 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)*, 2023, pp. 1–6.

[31] Z. Huang, H. Wang, W. Yang, M. Tang, D. Xie, T. Lin, W. Xing, and L. He, "Self-attention to operator learning-based 3d-ic thermal simulation," in *2025 ACM/IEEE Design Automation Conference (DAC)*, 2025.

[32] F. L. Incopera, *Fundamental of Heat and Mass Transfer*. Hoboken, New Jersay: John Wiley & Sons, 2006.

[33] G. W. Ernst Hairer, *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Berlin-Verlag: Springer Nature, 1996.

[34] Y. Zhan and S. S. Sapatnekar, "High-efficiency green function-based thermal simulation algorithms," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 9, pp. 1661–1675, 2007.

[35] B. Vermeersch, S. Mishra, M. Brunion, O. Zografos, M. Lofrano, H. Oprins, J. Myers, Z. Tokei, and G. Hellings, "Multiscale thermal impact of bspdn: Soc hotspot challenges and partial mitigation," in *2024 IEEE International Electron Devices Meeting (IEDM)*, 2024, pp. 1–4.

[36] M. Naeim, H. Oprins, S. Das, G. Van Der Plas, Y. Dai, P. Chen, C. Kao, D. Biswas, and D. Milojevic, "Thermal analysis of 3d stacking and beol technologies with functional partitioning of many-core risc-v soc," in *2024 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2024, pp. 33–38.

[37] Gmsh. [Online]. Available: https://gmsh.info

[38] I. A. Baratta, J. P. Dean, J. S. Dokken, M. Habera, J. S. Hale, C. N. Richardson, M. E. Rognes, M. W. Scroggs, N. Sime, and G. N. Wells, "Dolfinx: The next generation fenics problem solving environment," Dec. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.10447666

[39] Paraview. [Online]. Available: https://www.paraview.org

[40] B. Zhang, W. Xing, X. Zhao, and Y. Sun, "T-fusion: Thermal modeling of 3d ics with multi-fidelity fusion," in *Proceedings of the 30th Asia and South Pacific Design Automation (ASP-DAC)*, 2025, pp. 1406–1412.

[41] R. Courant, K. Friedrichs, and H. Lewy, "On the partial difference equations of mathematical physics," *IBM Journal of Research and Development*, vol. 11, no. 2, pp. 215–234, 1967.