

From Atom to System: Multi-scale Thermal, Mechanical, and Reliability Modeling of Advanced Packages for AI

Xi-Wei Lin
Synopsys Inc.
Sunnyvale, CA, U.S.
xiwei@synopsys.com

Ti-Yung Lan
Synopsys Taiwan
Hsinchu, Taiwan
tiyung@synopsys.com

Kelly Morgan
Synopsys Inc.
Sunnyvale, CA, U.S.
kamorgan@synopsys.com

Xiaopeng Xu
Synopsys Inc.
Sunnyvale, CA, U.S.
xxu@synopsys.com

Soren Smidstrup
Synopsys Denmark
Copenhagen, Denmark
soren@synopsys.com

Julian Schneider
Synopsys Denmark
Copenhagen, Denmark
julians@synopsys.com

Abstract— Thermal and mechanical effects are major concerns in advanced package assembly, operation, and reliability. This paper reviews and analyzes three aspects of 3DIC packages using multi-scale, multi-physics modeling methods: a) thermal profiles of HBM/GPU within a representative process and design space; b) thermal mismatch induced stress, warpage, and bonding risks during assembly; c) reliability concerns due to interposer stress, crack propagation, and solder fatigue. Atomic scale simulations are also used to extract material properties for chip or package level modeling.

Keywords— 3DIC, advanced packaging, thermal modeling, mechanical stress, molecular dynamics, reliability, crack propagation, solder fatigue, hybrid bonding, Cu grain structure

I. INTRODUCTION

Advanced packaging is at the foundation of the AI revolution. It brings together logic and memory in an interconnected system that enables the performance, bandwidth, and power efficiency necessary to process big data, train frontier models, and deploy inference for pervasive AI applications.[1-2] It's one of the most complex systems ever invented, involving 2.5D/3D heterogeneous integration of IC components from logic chiplets to HBM, interfaces, substrate, voltage regulator, as well as silicon photonics (Fig. 1). The primary challenges in making and operating such a system are thermal and mechanical effects that arise from packaging processes and system operations, affecting a) manufacturing and materials yield and cost, b) temperature and stress profile, and c) reliability.

Modeling and simulation are essential to tackle those challenges. It requires a multi-scale, multi-physics framework to solve problems ranging from materials design at atomic level to heat generation and thermal mismatch induced stress at chip and package level, impacting device performance, mechanical stress, and structural integrity.

II. THERMAL ANALYSIS

A. Power Distribution, Heat Transfer, and Temperature

Thermal properties of a package depend directly on its power distribution, which varies dynamically with the workload type, such as compute-bound vs. memory/IO-bound. Table I shows a typical power breakdown among components in an AI GPU, running at maximum total design power of ~1000W. Roughly half of the power budget is devoted to memory hierarchy, interfaces, and communication fabrics, while the remaining half to compute and SRAM.

While power generates heat, materials and structures determine heat transfer and temperature profile for a given set of boundary conditions. Figure 2 shows an example of a backside power delivery network (BSPDN) chip with uniform power density (54 W/cm²) in the middle and a heat sink on top. The peak temperature (T_{max}) decreases with increasing backside material thickness and thermal conductivity (κ), and increases slightly before decreasing with increasing Si carrier thickness. Notice that the change in T_{max} is quite small

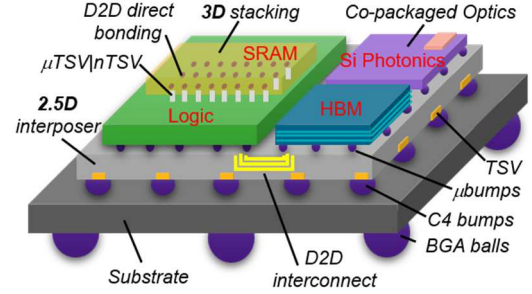


Fig. 1. An advanced package with HBM, Logic, and Si Photonics on 2.5D interposer and substrate, interconnected by TSV, mbumps, and C4 bumps, as well as SRAM on logic by hybrid bonding.

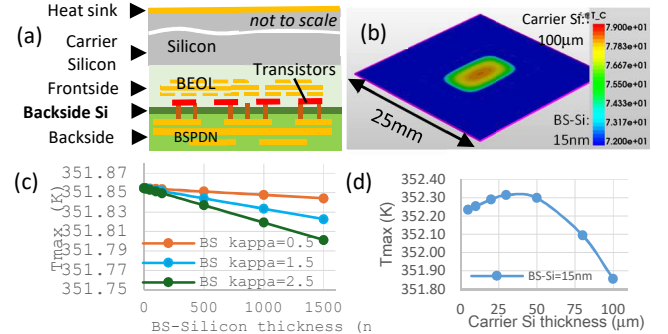


Fig. 2. Effect of Si thickness on thermal profile of a BSPDN chip: a) layer stack; b) temperature map; c) T_{max} vs. backside (BS) thickness at $\kappa = 0.5, 1.5, 2.5$ W/cm²; d) T_{max} vs. carrier Si thickness.

($10^{-2} - 10^{-1}$ K) within the range of thickness change, because Si is a good thermal conductor and not a bottleneck in the thermal “plumbing” system in this case.

TABLE I. POWER BREAKDOWN IN A GPU

Component	Share of Power	Functions
Logic	40-50%	Tensor cores, ALUs, L1/L2 caches
Memory	20-25%	HBM stacks
Interfaces/PHYs	10-15%	Physical layers, Logic to HBM buses
I/O & Interconnect	15-20%	NVLink, PCIe, controllers, etc.

B. Multi-scale Thermal Analysis

In a 3DIC package, heat is generated by transistors at nanometer scale, transferred through local materials and structures, and then spreads across chiplets and interfaces to the whole package at centimeter scale. The smear-material methodology, also known as homogenization, is widely used for multi-scale thermal and mechanical simulations (Fig. 3). It uses smear material to represent local features and maps smear properties from local features to global regions. It drastically reduces mesh size, speeds up runtime, and enables large scale analysis while preserving local properties.

This method is applied to a GPU/HBM accelerator (Fig. 4), where the full package structure is built by process emulation [3] and tiles A through E, corresponding to local patterns from C4 bumps to interposer, controller bumps, channel, and I/Os, are extracted and re-assembled in smear representation by Sentaurus Interconnect [4]. This allows for

process and design space exploration, as illustrated in three examples: a) The maximum junction temperature ($T_{j,max}$) at bottom DRAM die increases with increasing DRAM stack height, because of the trapping of vertical heat flux (Fig. 5); b) $T_{j,max}$ decreases with decreasing μ bump pitch, because denser μ bump array improves thermal conduction (Fig. 6); c) $T_{j,max}$ decreases by $\sim 9\%$ using hybrid bonding, due to better die-to-die thermal contacts than μ bumps (Fig. 7).

Figure 8 shows temperature distribution and thermal coupling between GPU and HBM, which increases with increasing HBM power. This highlights the importance of improving heat dissipation and the need to analyze the package as a whole for thermal optimization.

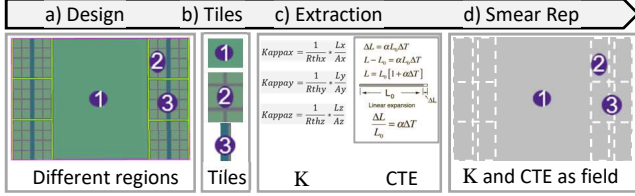


Fig. 3. Smear methodology for local to global thermal simulations. Decompose (a) a design into (b) tiles, each representing a unique local pattern; (c) extract thermal conductivity (K) and CTE; (d) reassemble tile properties on a homogeneous structure for global simulations.

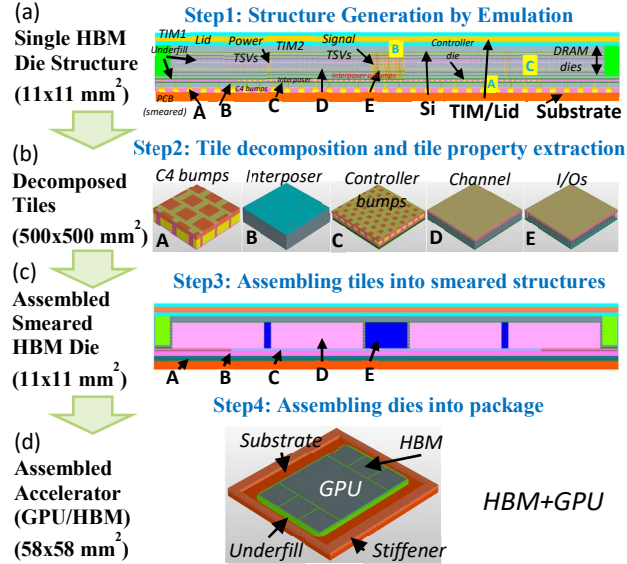


Fig. 4. Local-Global analysis workflow (Steps 1-4) for a GPU/HBM package. (a) Single HBM die with local patterns; (b) tile structures; (c) smeared HBM die; (d) assembled GPU/HBM package.

III. MECHANICAL SIMULATIONS

A. Warpage

Excessive warpage is a particular concern in advanced packaging, as in-plane deformation may lead to overlay errors, while out-of-plane deformation can compromise solder joint co-planarity, induce localized stress concentrations, and lead to downstream assembly defects or reliability failures. Mismatches in coefficient of thermal expansion (CTE) between the silicon die, substrate, mold compound, and other package materials cause differential expansion and contraction upon thermal loading, resulting in bending or twisting that can vary significantly with temperature.

Similar to thermal analysis, homogenization of chip and package substrate is commonly used for warpage simulations. Process simulation for steps such as solder reflow can provide insight into co-planarity and related risks, which may

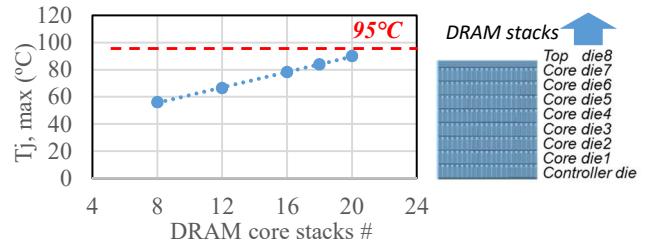


Fig. 5. $T_{j,max}$ vs. number of DRAM stacks in a HBM. The bottom DRAM die 1 has max junction temperature ($T_{j,max}$) and it increases with number of DRAM stacks.

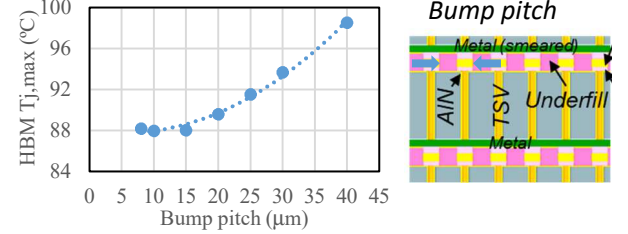


Fig. 6. $T_{j,max}$ vs. μ bump pitch. A tighter pitch helps heat conduction path and lowers $T_{j,max}$.

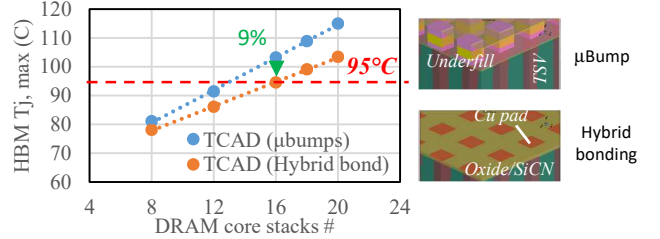


Fig. 7. Hybrid bonding leads to 9-10% temperature improvement, as compared to μ bump.

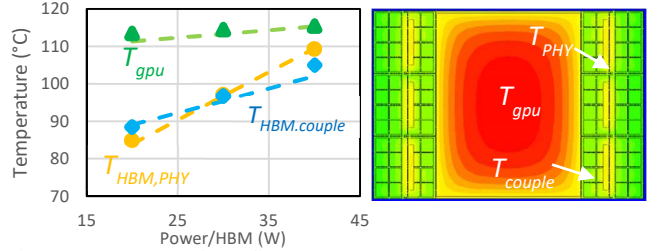


Fig. 8. GPU, PHY, and coupling Temperature vs. power per HBM.

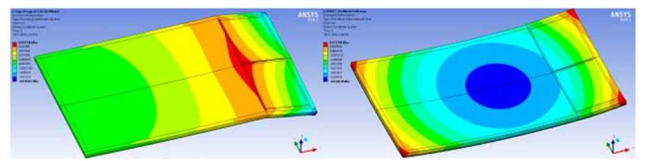


Fig. 9: Simulation of a live-bug view of warpage at reflow temperatures for 2.5D (left) vs. equivalent S-SWIFT (right) modules with 2 HBM devices. Note: sharp hinge and frowning-type warpage of 2.5D module.[5]

contribute to solder bridging or other assembly effects. Figure 9 provides an example of warpage prediction across a reflow temperature profile.[5]

B. Hybrid Bonding

Hybrid bonding (HB), the direct Cu-Cu and dielectric-dielectric bonding used to stack dies, is becoming central to AI chip packaging,[6] as it provides ultrahigh interconnect density, lower RC, and superior thermal conductivity. However, without a solder standoff to absorb CTE mismatch through compliant deformation, HB interfaces are far less forgiving of thermal-mechanical stress. Residual stress from the bonding process, combined with CTE mismatch between Cu, dielectric, and silicon layers, transmits almost directly

across the bond interface. This makes bond-line voids, interfacial delamination, and Cu pillar misalignment or diffusion-driven failure the dominant reliability concerns.

Various steps in the hybrid bonding process can be simulated to assess the joint integrity. One such example is found in [7], where the Smoothed Particle Galerkin (SPG) method from Ansys LS-DYNA is used to assess Cu-Cu direct bonding strength under tensile loading, accounting for the influence of Chemical Mechanical Polishing (CMP)-induced dishing and residual voids or defects at diffusion-bonded interfaces (Fig. 10).

A holistic modeling framework is proposed by [8], showing analysis of other aspects of HB, such as a calibrated feature scale CMP model for pattern dependent dishing and *ab initio* atomic level simulations of cohesive interaction at dielectric bonding surfaces.

C. Multi-grain Copper Reliability in Hybrid Bonding

Cu-Cu bonding is also affected by Cu grain structures. Not only in-grain mechanical properties but also cross-grain bonding depends on grain orientation. It is important to consider the grain structure dependence in void formation and interface morphology,[9] as well as reliability modeling. Figure 11a displays grain orientations and crack opening along the interface between two Cu pads, while Figure 11b shows the stress contours inside Cu pads. A phase field method is used for simulations.[4] It is evident that the crack tip stress distributions are affected by Cu microstructure.

IV. RELIABILITY MODELING

A. HBM and Interposer Integration

Due to CTE mismatches, stress hot spots may build up and cause reliability concerns in a 2.5D package.[10] Here we evaluated three interposer materials for their mechanical compatibility with HBM stack to evaluate the risk of interfacial delamination. Figure 12 compares their strain energy densities at an HBM/Interposer interface corner. With the Si-like CTE, but ~3X stronger modulus, the SiC interposer exhibits superior mechanical compatibility. In contrast, the polymer interposer is the worst in strain energy density, due to relatively large (>3X) difference in CTE.

B. Crack Formation and Propagation

Fracture is also a critical concern in AI chip and package design, as silicon and other brittle materials are highly sensitive to flaws and stress singularities, which can drive crack formation and eventual failure under thermal loading. Mismatches in CTE between dissimilar materials generate stress at interfaces and geometric discontinuities (e.g., die corners, interconnect layers), creating sites for crack initiation.

Rather than relying only on simple stress-based failure criteria, fracture mechanics approaches provide a more rigorous basis for predicting failure. This includes fracture mechanics parameters (e.g., Stress-Intensity Factor, J-Integral, Energy Release Rate) used to characterize the driving force for crack growth, as well as simulation techniques for modeling delamination and arbitrary crack propagation. Delamination studies often use Cohesive Zone Models (CZMs), which characterize damage initiation and evolution at pre-defined interfaces. Figure 13 provides an example of fracture in back-end-of-line (BEOL) structures.[11]

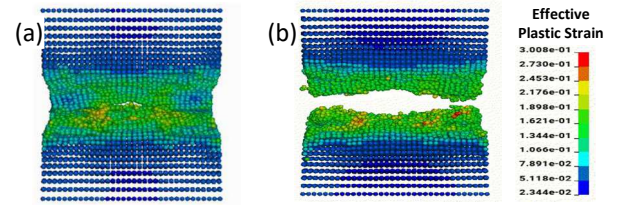


Fig. 10: Dynamic failure process of bonded copper pads at (a) $t=0.0013s$ and (b) $t=0.002s$ under tension loading. Material separation is initiated at the weakest bonding position associated with CMP dishing vacancy.[7]

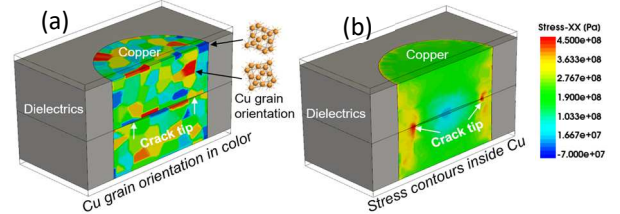


Fig. 11. (a) Cu grains and interfacial crack opening in Cu pads. (b) Stress contours inside Cu.

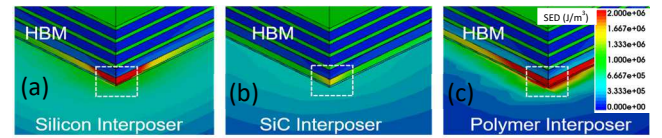


Fig. 12. Strain energy density (SED) contours for (a) Silicon, (b) SiC, and (c) polymer interposers, with max value at 9.93×10^5 , 4.77×10^5 , and 21.91×10^5 J/m³, respectively.

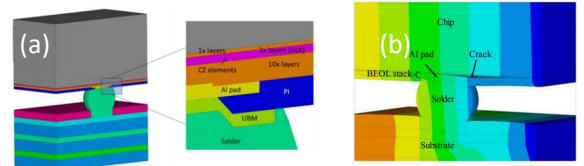


Fig. 13. (a) Schematic of 3D local model. (b) Displacement contours with failed CZ elements.[11]

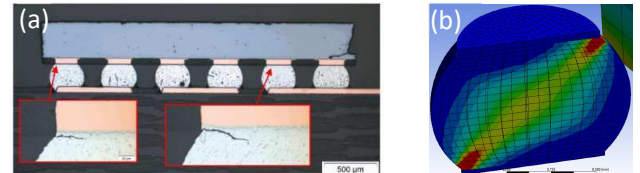


Fig. 14. (a) Cross-sections performed after cyclic fatigue test before total solder ball fracture: crack initiations are visible on each solder ball.[12] (b) Simulated nonlinear plastic work in a critical solder joint subjected to thermal cycling.

C. Solder Fatigue

Solder joint failure from thermal-mechanical fatigue is a dominant wear-out mechanism in electronics, including AI hardware. As temperatures change, packages and printed circuit boards expand and contract unevenly due to CTE mismatch, subjecting solder joints to repeated shear loading. Accumulated damage drives crack initiation and eventual failure (Fig. 14a).[12] Reliability can be improved through implementation of enhanced thermal management strategies, selection of lower-CTE materials and suitable solder alloys and optimized joint geometry/standoff height.

Historically, physical testing has been used extensively to assess such failures. However, a variety of simulation methods (both analytical and finite-element-based) have been introduced to determine reliability/lifetime predictions. In some cases, detailed finite-element models (even local- or sub-models) are used to account for detailed solder, trace, via,

and other geometries and more accurately account for material behaviors, such as creep and plasticity. Figure 14b shows how simulation can be used to perform such assessments.

V. MATERIALS ENGINEERING

A. Cu Grain Bonding Strength Simulation

Atomic level simulations can be used to extract Cu grain boundary (GB) strength and screen robust interface structures relevant to HB reliability. Grain boundaries are constructed using the QuantumATK Grain Boundary Builder and relaxed with the MACE potential.[13] The work-of-separation, defined as the reversible work required to separate an interface, is obtained from constrained NVT molecular dynamics (MD) simulations with a Nosé–Hoover thermostat.[14] It is calculated from the energy difference between the separated and joined configurations, normalized by the interfacial area. It's a measure of GB strength and provides a useful metric for comparing different GB orientations. The results (Fig. 15) exhibit substantial variation, implying Cu microstructure may affect bonding properties.

B. Thermal Conductivity of Thin Films and Interfaces

MD simulations [14] can also be used to provide detailed insight into the atomic-scale mechanisms of thermal transport in thin films and interfaces. Here, thermal transport across an amorphous-SiO₂/crystalline-AIN/amorphous-SiO₂ structure is investigated. The simulations employ a machine-learned moment tensor potential (MTP), trained to density-functional-theory (DFT) data, enabling efficient and accurate modeling.[15] The out-of-plane thermal conductivity and interface thermal conductance are extracted from the resulting temperature profiles using Fourier's law. Figure 16 shows that the AIN thermal conductivity increases with film thickness, due to decreased phonon boundary scattering in thicker films, and reaches ~80 W/mK at 50nm, significantly below the bulk value of ~300 W/mK at room temperature [16]. In contrast, the interface (i.e., boundary) thermal conductance is largely independent of film thickness. It depends, however, on the atomic termination of the interface, with the N-terminated interface exhibiting a higher conductance (~1.4 GW/m²K) than the Al-terminated interface (~0.9 GW/m²K), reflecting differences in interfacial bonding. These results highlight the importance of both size effects and interface chemistry in determining thermal transport across AIN/SiO₂ structures.

VI. SUMMARY

The temperature profile of a package depends on power distribution, geometry, material properties, and boundary conditions. Homogenization is effective to model global thermal and mechanical properties while preserving the influence of local attributes. Thermal profiles of HBM/GPU packages are analyzed for various process and design conditions. Thermally induced stresses are key drivers of warpage during assembly and reliability concerns in HBM/interposer and hybrid bonding. Prior work on reliability due to crack propagation and solder fatigue is reviewed to highlight important reliability mechanisms. Atomic scale simulations are useful for providing physical insights and extracting material properties for package scale modeling.

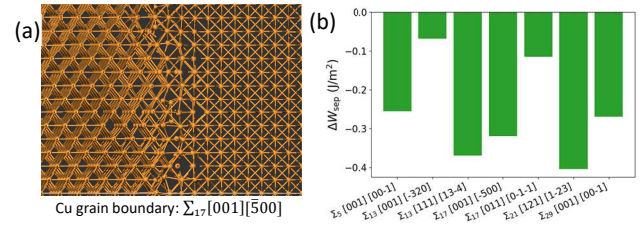


Fig. 15: (a) atomic view of a Cu grain boundary. (b) Cu grain bonding strength (deviation in work of separation relative to the bulk value) varies depending on grain orientation.

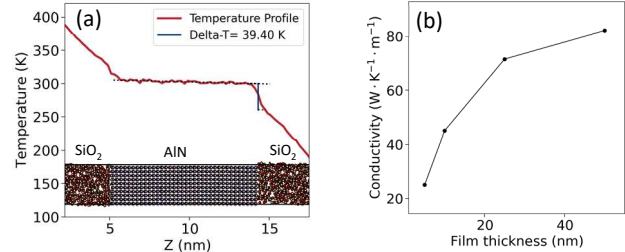


Fig. 16: (a) Local temperature profile in a SiO₂/10nm-AIN/SiO₂ interface model from MD simulations showing temperature gradient and drop at the interface. (b) Thermal conductivity as a function of AIN film thickness.

REFERENCES

- [1] L.C. Lu, "Advancing Package and System Integration for Next Generation AI", IEEE/JSAP Symp. VLSI Tech., June 2026.
- [2] X.W. Lin *et al.*, "Heterogeneous Integration Enabled by the State-of-the-Art 3DIC and CMOS Technologies: Design, Cost, and Modeling", IEEE International Electron Devices Meeting (IEDM), Dec. 2021.
- [3] Sentaurus Process Explorer User Guide X-2026.03, Synopsys.
- [4] Sentaurus Interconnect User Guide X-2026.03, Synopsys.
- [5] G. J. Scott *et al.*, "Heterogeneous Integration Using Organic Interposer Technology", IEEE Electronic Components and Technology Conference (ECTC), June 2020.
- [6] J. H. Lau, "State of the Art and Outlooks of Cu–Cu Hybrid Bonding", IEEE Trans. On Components, Packaging and Manufacturing Technology (CPMT), Dec. 2025.
- [7] B. Ren *et al.*, "Integrated Process-Mechanical Stress Analysis of 2.5D/3D ICs with Two Types of Interconnections in Advanced Packaging", IEEE Inter. Integr. Rel. Workshop (IIRW), Oct. 2023.
- [8] S. Sitaraman *et al.*, "A Holistic Development Framework for Hybrid Bonding", IEEE Electronic Comp. Tech. Conf. (ECTC), May 2022.
- [9] J.-U. Lee *et al.*, "Crystal plasticity-based modeling and experimental validation of the influence of microstructures and grain boundary junction types on the Cu–Cu bonding interface," in Proc. IEEE Electronic Comp. Tech. Conf. (ECTC), May 2025.
- [10] Y. Hwang *et al.*, "High Bandwidth Memory 4 (HBM4) Package Development Challenges", IEEE/JSAP Symp. VLSI Tech., June 2026.
- [11] S. Raghavan, "Experimental and Theoretical Study of On-chip Back-end-of-line (BEOL) Stack Fracture During Flip-chip Reflow Assembly," Ph.D. dissertation, Georgia Inst. of Technol., 2014.
- [12] S. Zanella, "Creep Fatigue Interaction in Solder Joint Alloys of Electronic Packages", Ph.D. dissertation, École Polytechnique, Université Paris-Saclay, 2018.
- [13] I. Batatia *et al.*, "A Foundation Model for Atomistic Materials Chemistry", J. Chem. Phys. Nov. 2025.
- [14] J. Schneider *et al.*, "ATK-ForceField: a New Generation Molecular Dynamics Software Package", Model. Simul. Mater. Sci. Eng. (2017).
- [15] S. K. Achar *et al.*, "Using Machine Learning Potentials to Explore Interdiffusion at Metal–Chalcogenide Interfaces." ACS Appl. Mater. Interfaces, Dec. 2022.
- [16] Z. Cheng *et al.*, "Experimental Observation of High Intrinsic Thermal Conductivity of AIN", Phys. Rev. Materials 4, April 2020.