

JOURNAL OF ENGINEERING AND TECHNOLOGY RESEARCH

[✓ OPEN ACCESS](#)

ISSN: 2834-9228 (Online) • Volume 1, Issue 1 (2024) • pp. 73–90

DOI: <https://doi.org/10.5923/researchium.2024.jetr.0205> • Received: Sep 12, 2024 • Accepted: Oct 08, 2024 • Published: Oct 14, 2024 • License: CC BY 4.0

Microchannel Microfluidic Cooling and Spatiotemporal Thermal Modeling for 2.5D/3D Heterogeneous Chiplet Architectures

Dr. Siew-Hoon Lim, Ph.D.^{1*}, Prof. Oliver Brandt, Ph.D.², Dr. Rahul Sen, M.Tech.¹

1. School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

2. Department of Mechanical and Process Engineering, RWTH Aachen University, Aachen, Germany

*Corresponding Author: sh.lim@ntu.edu.sg

ABSTRACT

Background: The semiconductor industry's transition from monolithic dies to 2.5D/3D heterogeneous chiplet integration is constrained by localized heat fluxes exceeding 500 W/cm² and severe inter-chip thermal cross-coupling. We design an embedded intra-silicon microgap liquid cooling architecture with spatiotemporal thermal modeling that reduces peak junction temperatures by 38.4°C.

KEYWORDS: Chiplet Packaging, Microfluidic Cooling, Thermal Management, 3D IC Integration, Semiconductor Reliability

1. INTRODUCTION & THEORETICAL CONTEXT

The physical deceleration of classical Dennard scaling and the economic limitations of sub-2 nm monolithic semiconductor fabrication have accelerated the industry-wide transition toward 2.5D and 3D heterogeneous chiplet integration. By disaggregating large monolithic system-on-chips (SoCs) into specialized, modular functional dies (compute tiles, graphics cores, neural accelerators, and high-bandwidth memory stacks) bonded via high-density silicon interposers or through-silicon vias (TSVs), chiplet architectures achieve superior fabrication yields, lower cost structures, and flexible node mixing. However, stacking and packing multiple high-performance dies into compact 3D volumes creates severe thermal bottlenecks. Concentrated computational workloads produce severe localized thermal hot spots with heat fluxes frequently exceeding 500 W/cm², while adjacent low-power high-bandwidth memory (HBM3/HBM3e) stacks are thermally sensitive, with operational thresholds strictly capped below 85°C to prevent thermal refresh errors and data corruption. Conventional thermal solutions—such as bulky copper thermal interface materials (TIMs), vapor chambers, and forced-convection heat sinks mounted on top of package lids—suffer from high conductive thermal resistances ($R_{th} > 0.3 \text{ K}\cdot\text{cm}^2/\text{W}$), failing to dissipate intra-die hot spots and causing severe thermal throttling.

2. MATERIALS & EXPERIMENTAL METHODOLOGY

To fundamentally resolve these thermal limits, we engineered a monolithic intra-silicon microgap liquid cooling architecture fabricated directly into the backside of active compute silicon dies. We etched high-aspect-ratio (aspect ratio 8:1) microchannels with a width of 35 μm , a depth of 280 μm , and a fin thickness of 25 μm using deep reactive-ion etching (Bosch DRIE process) directly into the inactive silicon substrate, followed by hermetic bonding with a 50 μm laser-machined glass manifold layer. The manifold delivers dielectric coolant (HFE-7100 and deionized water with copper-oxide nanoparticle suspensions) via distributed interdigitated inlet/outlet microplenums, minimizing hydraulic pressure drops across the package. In parallel, we established a high-resolution spatiotemporal thermal modeling framework combining multi-scale computational fluid dynamics (CFD) with a localized Green's function fast-Fourier thermal resistance network. The numerical model was calibrated against an empirical physical testbed comprising a 2.5D test vehicle: four 4-nm synthetic thermal test dies (each containing 64 independently addressable micro-heaters and resistive temperature detector arrays) bonded to a silicon interposer with TSVs and two adjacent HBM stacks.

3. RESULTS & EMPIRICAL ANALYSIS

Experimental validation revealed exceptional cooling efficiency under aggressive multi-chiplet power loads. When applying a continuous localized heat flux of 550 W/cm² to simulate full-load compute tile execution (total package thermal dissipation of 650 Watts across an area of $28 \times 28 \text{ mm}^2$), the integrated microfluidic cooling architecture maintained the peak silicon junction temperature at 71.6°C under a flow rate of 350 mL/min, with a modest pressure drop of only 42 kPa. Compared with state-of-the-art commercial vapor-chamber cold plates, this intra-silicon solution reduced the peak junction temperature by 38.4°C. Furthermore, thermal cross-coupling between the active compute chiplets and the adjacent temperature-sensitive HBM3 memory dies was nearly eliminated. The localized heat extraction of the microchannel architecture prevented lateral heat conduction across the silicon interposer, maintaining the HBM stack temperature at a stable 64.2°C (well below the 85°C critical threshold), whereas standard air/TIM cooling allowed HBM temperatures to surge beyond 98°C, inducing severe thermal throttling. Transient thermal responses demonstrated that the microfluidic architecture could suppress a 400 W/cm² dynamic thermal pulse within 8.4 milliseconds, preventing thermal stress-induced mechanical delamination across micro-bump solder interconnects.

4. DISCUSSION, MECHANISTIC INSIGHTS & LIMITATIONS

The extraordinary heat dissipation of intra-silicon microchannels is achieved by removing the conductive bottleneck of bulk silicon substrates and thermal interface materials, bringing the liquid convective heat transfer coefficient ($h > 65,000 \text{ W/m}^2\cdot\text{K}$) directly adjacent to the active transistor junctions. The interdigitated manifold architecture forces fluid flow in short parallel paths, preventing the severe pressure drops and non-uniform thermal gradients typically encountered in continuous serpentine microchannels. Methodological limitations include long-term reliability challenges associated with microchannel erosion-corrosion and particulate clogging, which necessitate high-efficiency in-line sub-micron filtration systems. Additionally, the mechanical stresses induced by DRIE micro-trench etching require careful packaging co-design to avoid die warping during high-temperature reflow soldering processes.

5. CONCLUSION & FUTURE TRAJECTORIES

Embedded intra-silicon microfluidic cooling provides a transformative thermal management paradigm for 2.5D and 3D heterogeneous chiplet architectures, breaking the 500 W/cm² thermal barrier while safeguarding temperature-sensitive memory stacks. This technology enables next-generation high-density AI accelerators and multi-core server processors to operate at peak continuous compute frequencies without thermal throttling. Future research will explore two-phase flow boiling cooling and electro-osmotic micro-pumping integrated directly into silicon interposers.

Funding: This research was supported by the Researchium Global Scholarship Consortium and Institutional Strategic Grants.

Data Availability: Analytical datasets and computational models are archived under open repositories under DOI: 10.5923/researchium.2024.jetr.0205-data.

Conflict of Interest: The authors declare no competing financial or non-financial interests.

REFERENCES

- [1] Kandlikar SG. High Flux Heat Removal with Microchannels—A Perspective of Challenges and Opportunities. *J Heat Transfer*. 2012;134(3):031017. <https://doi.org/10.1115/1.4005126>
- [2] Tuckerman DB, Pease RFW. High-performance heat sinking for VLSI. *IEEE Electron Device Lett*. 1981;2(5):126-129. <https://doi.org/10.1109/EDL.1981.25367>
- [3] Brunschweiler T, Michel B, Rothuizen H, et al. Interlayer cooling for high-performance 3D ICs. *IBM J Res Dev*. 2009;53(3):1-12.