Towards Cost-Effective Thermal Management Method for Processing in 3D Memory

---

A Dissertation Presented to the Graduate Faculty

of School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Jun-Han Han

August 2022

Abstract

As Moore's law states, the number of transistors in an integrated circuit doubles every two years. The increase in transistor density and operating frequency produces heat problems in semiconductor devices, limiting further advancements. To resolve thermal issues, the performance criteria in modern computing have transitioned from higher frequency and denser transistors to more advanced integration processes, algorithms, and architecture. However, due to data-intensive computing, the performance has been degraded from the bandwidth limitation between processor and memory. Processing-in-memory and 3D chiplet are preeminent solutions for the degraded performance due to the "memory wall." Processing-in-memory prevents performance degradation by migrating a portion of the computation load to the memory, while 3D chiplet integration does this by shortening data movement and increasing bandwidth. The high bandwidth memory is a 3D integrated circuit with memory dies stacked on top of a logic die. The 3D-stacked memory is the core technology in processing-in-memory architecture and is a region where thermal issues usually occur in the 3D chiplets. The computation in the 3D-stacked memory induces a higher power density in the 3D stacks. The higher power density increases the volumetric heat flux, while the heat sink removes the heat only from the top surface. This volumetric thermal hazard also occurs in other 3D chiplets in the same manner. To manage thermal issues in 3D semiconductor devices, a thermal management solution with a cost-efficient volumetric cooling system is essential.

In this dissertation, we propose a new thermal management method with microfluidic cooling to enable processing in-3D-memory. This dissertation presents the thermal management solution in two aspects: device design and thermal simulation. In the device research aspect, we develop a cost-effective microchamber cooling structure for 3D-stacked memory devices. While the conventional microchannel method has a high fabrication cost due to the additional fluidic structures, the proposed method reduces the fabrication cost by exploiting the gap between the upper and lower stacks and the 3D printing package. We demonstrate the validity of the microfluidic chamber cooling method with an experimental setup. The prototype shows the feasibility of the proposed method. In the second part of this study, we simulate the thermal behavior of the 3D chiplet with the microfluidic cooling system. First, we model the thermal characteristics of the 3D stacked memory from a 4-layer stack to a high-rise stack. For further

investigation, we use HotSpot for fast and accurate thermal simulation. We run thermal simulations using HotSpot in different cooling configurations for a processing-in-3D-memory system, where memory layers sit on top of a logic layer. The HotSpot simulation results are validated with a Multi-Physics simulation, COMSOL. When a detailed fluidic behavior is observed, the cooling capacity of the microfluidic cooling system is degraded by the thermal boundary. We minimized the impact of the thermal boundary layer by creating a more detailed thermal model of the fluidic chamber with micropillars inserted.

We investigate the thermal characteristics of modern chiplet systems. The temperature of the 3D stacked high-performance processors has significantly increased from that of the 2.5D integration. We lower the temperature below the operating temperature limit by implementing novel thermal management methods such as alternating flow directions and applying multi-layer cooling. We also investigate a heterogeneous processor-memory chiplet system in 2.5D and 3D integration. While the heatsink cooling system cannot maintain the temperature of the 3D chiplet below the operating temperature limit, the microfluidic cooling system reduces the temperature with a cost-affordable hybrid cooling method. We conclude that our proposed thermal management method provides a cost-effective and 3D chiplet-compatible solution that guides chip designers and architects in understanding the thermal behavior of modern computing devices.

Acknowledgments

I want to first thank my advisor, Professor Mircea Stan, for being an amazing mentor who not only treated me as an independent researcher but also guided me with great support. I would also like to thank the advisory committee, Professors Kevin Skadron, Nathan Swami, Xinfei Guo, Kyusang Lee, and Sarah Sun, for the support and guidance in writing my dissertation. I would also like to acknowledge my past fellow researchers in our projects, Trey(Robert) and Karina, as well as all HPLP members, especially Sergiu, Tommy, Vaibhav, Rahul, Elisa, Ceylan, Sakib, Yunfei, Patricia, Pai, and Mandi, who have supported me to overcome difficulties in research and life projects.

I would like to thank my parents and friends, as well as former colleagues in ETRI back in South Korea, who had encouraged and supported me to come to the United States for this opportunity. I would like to thank my furry cat daughter, Hanu, for being a tranquilizer. I want to thank my newborn child, Teo, for being my motivation to come this far and strive for new opportunities. Lastly, I want to thank my wife, Wendy, for being my best friend and the love of my life who has always been by my side.

List of Figures

List of Table

Contents

1. Introduction

1.1 Motivation

The semiconductor industry has been developing at the pace of Moore's Law for the past decades. The number of transistors in an integrated circuit (IC) doubles about every two years with the continuation of transistor scaling. The advancement in semiconductor technology also led to an increase in operating frequency, performance, and power consumption of integrated circuits. The trend toward "performance at any cost," resulting in an increase in transistor density and operating frequency, came to an end due to the inherent thermal limits with increasing power consumption [1]. The restriction of maximum power density and frequency imposed semiconductor industries on developing superseded methods such as parallel computing, advanced integration, algorithm, and architecture.

In the modern computing era, the "memory wall" between the processor and memory is the primary performance bottleneck for data-intensive workloads. The limited bandwidth between the processor chip and memory chip causes performance deterioration. There are two principal solutions among the emerging technologies, one in architecture and the other in the integration process. One solution is processing-in-memory (PIM) which solves this problem by migrating a portion of the computation load to the memory. The other solution is 3D chiplet integration which prevents deterioration by shortening data movement and increasing bandwidth.

Processing-in-memory and near-data-processing manage restricted power and performance by migrating a portion of the computation load to the memory. For conventional architecture, the processing unit is far from the data storage, causing 62.7% of the total system energy to be spent on the data movement [2]. Migration of the processing unit to near the memory contributes to energy saving due to the shortened data path and reduced data movement. In addition, the hierarchical and parallel structure of dynamic random-access memory (DRAM) enables parallel data processing of PIM architecture. 3D memory, such as high bandwidth memory(HBM) [3], provides a suitable technology for memory-centric computing by stacking heterogeneous layers together. On the other hand, the processing in the 3D memory induces a higher power density in the stack. Additional processing features in the memory system utilize more power than the typical memory read and write operation. The higher power density causes

the PIM architecture to encounter the principal thermal issue of the semiconductor, limitation of power, and performance.

Chiplet technology achieves better performance metrics with reduced power consumption by increasing the data bandwidth. Chiplet integration has multiple dies in a single package integrated on a silicon interposer providing higher bandwidth between the chips and a board. An example of such an interconnect is the embedded multi-die interconnect bridge (EMIB) from Intel to enable die-to-die integration on the package [4]. While chiplet-based systems offer more significant advantages over the monolithic chip approach, it is mainly limited to 2D integration. 3D integration is limited to low-power units such as 3D HBM. The 2.5D chiplet locates 3D-stacked devices and other dies side by side on the interposer. Consequently, the 2.5D chiplet system has a limited bandwidth compared to the 3D integration[5]. Thus, from the interconnection aspect, 3D IC is even more advantageous in performance and power-saving because it significantly reduces the interconnection distance compared to the interposer.

For the 3D integration, multiple dies are stacked as a single chiplet utilizing through-silicon-vias (TSVs). An actual 3D IC chiplet system enables an advanced package to stack memory on a processor or processor on a processor. Several recent 3D integration demonstrations have been explored to allow for opportunities in high-performance computing [6], image sensors [7], or gas sensing [8]. These are also known as monolithic 3D integration. However, one of the most critical factors that obstruct the 3D IC development is the thermal issue resulting from the increased power density. Power density is already an issue for high-frequency 2D monolithic chips in advanced technology nodes[5]. The temperature of 3D IC rises over operating temperature because of the dimension mismatch between heating and cooling. While the 3D IC generates volumetric heat with increased power dissipation, the top surface dimensions limit the conventional cooling capacity. This thermal problem will get progressively worse as more computing-intensive units (such as AI computing units or GPUs) stack on top of each other or a CPU[9]–[12]. Moreover, thermal issues have become more critical in modern interconnection technologies, such as front-end-of-the Line (FEOL), Back-End-of the-Line (BEOL), and Through-Silicon-Vias (TSV), as they increase the thermal resistance in the chiplet and accordingly worsen the thermal issues [13].

Engineers and researchers in academia and industries have been actively seeking cooling solutions for the thermal challenges[1], [10]–[13]. Recent studies have investigated microfluidic cooling as one of the most promising thermal management solutions for future 3D integrated systems [14]–[17]. But there is still no practical solution that can fully resolve the thermal issues in 3D integrated circuits. A cost-efficient thermal management method is required to resolve the issues. In addition, the importance of identifying and developing a detailed understanding of the capabilities and limitations of critical thermal technologies in thermal aspects arose. Chip designers or architects would like to find the thermal characteristics of the chiplet design in the early stage of development. By exploring the design space, chip designers can make early decisions and shorten the development cycle. This is critical, especially for 3D IC, as thermal management is a crucial factor for stacking strategies. Therefore, a fast and accurate simulation framework that integrates the most advanced microfluidic cooling modeling is necessary.

1.2 Conventional thermal management methods

1.2.1 Heatsink

A heatsink is a heat exchanger that transfers thermal energy from a high-temperature chip package to a low-temperature ambient medium. The fin structure enlarges its surface area to maximize the heat transfer to the ambient fluid, such as air. A fan can be attached on top of the heatsink to enhance the airflow so that the heat transfers. Depending on the presence or absence of a fan, it can be a passive or active heatsink. We chose four different heatsinks to evaluate the expected cooling capacity: one passive heatsink and three active heatsinks with different cooling capacities and costs [18].

1.2.2 Throttling

Previous studies investigated memory throttling control using temperature information from a thermal sensor on a Dual Inline Memory Module (DIMM) [3]-[6]. The temperature can be lowered by throttling with reduced memory traffic on the memory bus [21]. Because the throttling method manages the memory performance depending on temperature information, it cannot consider temperature variations over the system. As a result, it leads to excessive

throttling of the memory system and underutilization of the system budget. Liu et al. proposed a control technique for minimizing the temperature variations over the DIMM [20]. This method manages data traffic with expected temperature change based on a superficial thermal model of the DIMM. Meanwhile, Lee et al. investigated the memory cell's thermal behavior with extensive consideration of the device and circuit behavior [22]. From this approach, they optimized operating time with latency control. Despite the efforts of previous researchers, the throttling method has a critical limitation: the deterioration in the memory performance.

1.2.3 Liquid cooling for 2D and 2.5D chiplet

Microfluidics is the technology of manipulating fluids in channels with dimensions from tens to hundreds of micrometers. The heat generated from dies can be removed by the coolant flow on the top surface, the flow through the inside channels, or the flow through the cavity between the dies. The first approach is cooling the chip by jet impingement onto the top surface of the die [15], [23]. This method can achieve highly efficient cooling by eliminating the thermal interface material and applying a high flow rate onto the die. However, this structure can be applied only to the top surface, which is disadvantageous to the 3D-ICs similar to the conventional heat sink methods. Another microfluidic solution is called the intra-chip cooling method, where the flow removes the heat through the channels that are integrated on the backside of the dies. The embedded microchannels are advantageous because of their water-tight structure. However, in the case of 3D-stacked ICs, additional vertical fluidic interconnections such as fluidic via and pipe are required to enable vertical coolant movement. The additional fluid structures increase fabrication steps, costs, and chip dimensions. In addition to the difficulties in the fabrication process, the microchannel cooling method conflicts with TSVs by occupying the die area [24], [25]. The electrical interconnection between the dies, such as TSVs, micro-bumps, or Cu-pillars, occupies most of the chip dimensions in high-bandwidth devices. Therefore, the microchannel cooling method has an inherent limitation to be applied for multi-layer 3D-ICs. Lastly, the inter-chip cooling method is another candidate for the cooling solution of 3D-IC. The coolant removes the heat by the flow between the dies. By exploiting the inherent gap between the dies, this approach does not require additional fluidic structures and

can save fabrication costs. It also has multi-layer scalability, which is suitable for multilayered 3D stacked IC cooling systems.

1.3 Heterogeneous Integration Roadmap

We proposed the thermal study in 2018 to provide a low-cost thermal management solution for processing-in-memory devices aligning with the *Thermal and Power Management* project in the *Center for Research on Intelligent Storage and Processing-in-memory (CRISP)* funded by Semiconductor Research Corporation (SRC)'s *Joint University Microelectronics Program* (JUMP). Our study referenced the request for a thermal solution from the 2015 *International Technology Roadmap for Semiconductors (ITRS)* by *Semiconductor Industry Association* (SIA)[1]. The ITRS roadmap predicts an increased power density of over 100 W/cm^2, while the hotspots have four times larger power densities in the local area.

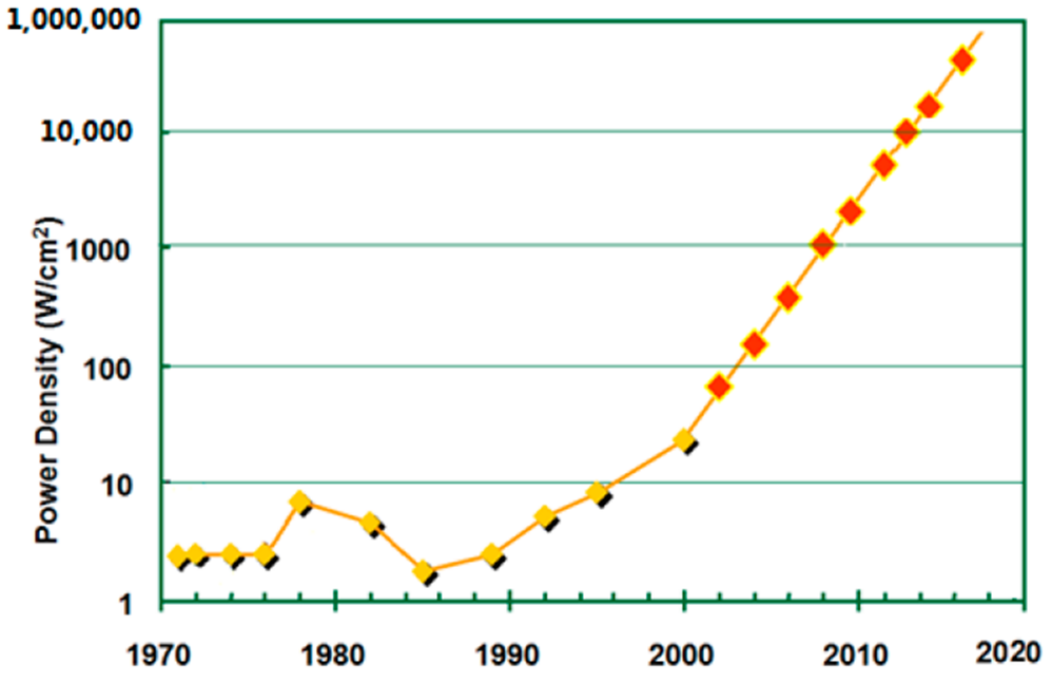


Figure 1. Power density trend for a single package [1]

Concurrently with our study, the IEEE Electronics Packaging Society launched the publication of the Heterogeneous Integration Roadmap (HIR) international roadmap in 2019.

HIR is an organization of task working groups continuing the discontinued ITRS roadmap in 2015[13]. The HIR roadmap described the current challenging thermal situation and demand for novel thermal management methods. The requirements for the thermal management methods 2D and 3D ICs are to manage thermal issues for the average power density of $2W/mm^2$ and to manage two-to-four times higher hotspot power density with power-cost-architectural considerations. For the thermal community, it is also essential to investigate thermal behavior to understand the capabilities and limitations of the promising thermal management methods [13]. Accordingly, developing a thermal management method with high peak-power cooling capacity, low cooling power, low-cost fabrication, low area overhead in silicon design, and providing thermal behavior is required.

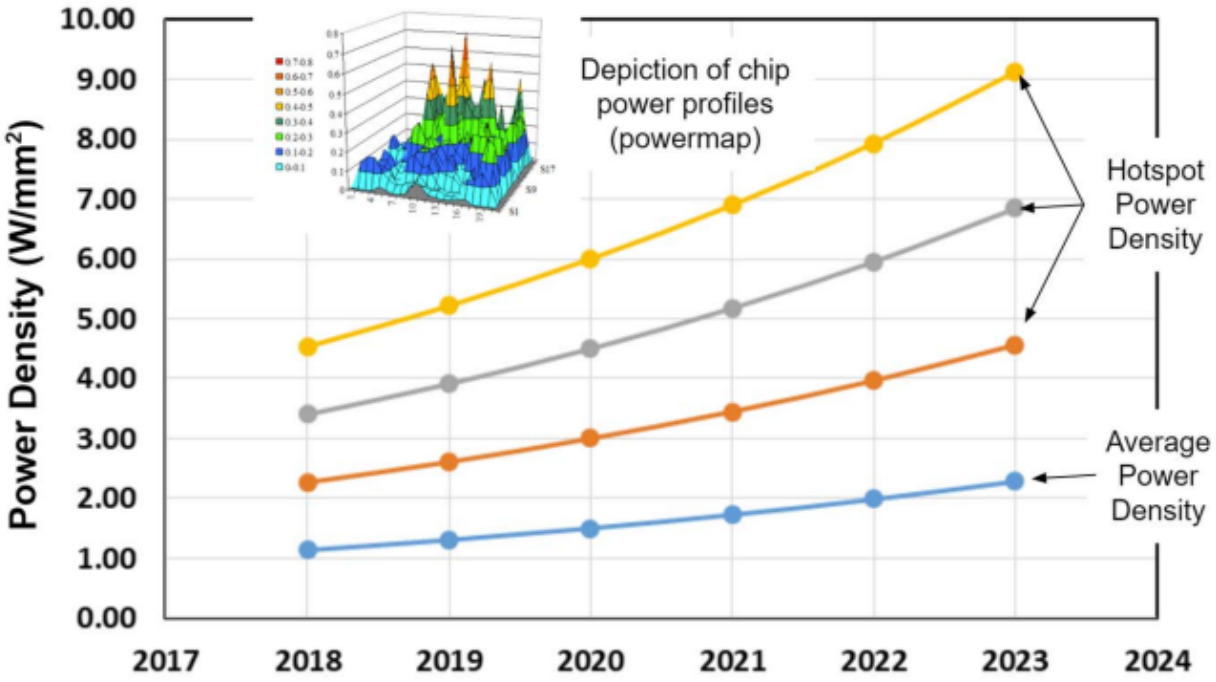


Figure 2. Nominal Thermal Demand Envelopes covering average and hotspot peak power density for both 2D and 3D Architectures [2]

1.4 Thesis Contributions

This thesis identifies the thermal issues that the modern chiplet will need to address due to increasing device and power densities and limited cooling capacity, along with demands for cost-effective thermal management. The current thesis aims to develop low-cost thermal management solutions focused on hotspots investigating thermal behavior in device and design aspects. The main contributions of this thesis are summarized as follows:

1. Proposing a cost-effective microchamber cooling method exploiting an existing die cavity and 3D printing package. The conventional heatsink has several drawbacks in the application of modern chiplets. The cooling capacity is limited due to the limited cooling surface and modern interconnection layers such as FEOL and BEOL. In the case of 3D stacked IC, the heat dissipation transfers from the bottom to the top layer generating a thermal gradient. This vertical thermal distribution causes thermal issues in heterogeneous integration. In the case of processing-in-memory devices, the processor and memory have different operational temperature limitations, and the high temperature of the processor affects the operation of the memory. Microchannel cooling is a promising solution offering an embedded cooling system removing the heat directly from the silicon layer. However, the additional microfluidic structure significantly increases the fabrication cost and chip dimension. We propose a microfluidic cooling system using the gap between the layer as a fluidic channel and a 3D printing package as a body of inlet and outlet. By eliminating the additional cooling structure, the proposed method prevents the increase in fabrication cost and chip dimension. In addition to the cost-effectiveness, the open floor plan of the fluidic chamber is advantageous in cooling behavior. The interconnection between layers such as microbumps or micropillars expands the area of the heat exchange surface.
2. Validating the proposed microchamber cooling method through an experiment. We designed and fabricated an experimental test setup to validate the proposed microfluidic cooling system. The main question is, “does the microfluidic flow through the microchamber cooling structure with a 3D printing package can cool the chip?”. To demonstrate the idea, we fabricated a fluid chamber with a thermal test chip attached to the PCB substrate. The dielectric coolant has been injected with a pipet and pulled out

with a paper pump. While the thermal resistors generate heat flux, the thermal sensing diodes sense the temperature in real-time. Our results show that the proposed method is a promising alternative to conventional microchannel methods with advantages in low-cost fabrication. In addition, we demonstrated low cooling power consumption through the thermal circuit modeling and flow rate calculation. Our cooling system is affordable with a low-cost hydraulic pump with a low power consumption of under ten watts.

3. Implementing thermal modeling. Thermal modeling is an essential tool for understanding the thermal behavior of chiplets. Through the thermal model, we can calculate the temperature of the chiplet for given power consumption. We demonstrated the thermal model of the microchamber thermal test system and processing-in-memory device. We calculated the temperature in the logic and memory layer of the processing-in-memory system with different heatsinks. We also calculated the hydraulic pump's required flow rate and pressure for microfluidic cooling for processing-in-memory devices. We also investigated the thermal behavior of high-rise 3D-stacked IC.
4. Investigating the thermal behavior with HotSpot 7.0. We recently released a new version of the HotSpot simulator with the microfluidic cooling feature. The HotSpot is a thermal circuit model-based, fast and accurate thermal simulation tool for chip designers. We can simulate the temperature of the design in the pre-RTL stage with inputs of the floor plan and power traces. We model and simulate the thermal characteristics of the microchamber embedded in processing-in-3D-memory devices. Next, we compared the cooling performance of the microchannel and microchamber methods. Additionally, we compared the simulation results from Hotspot with results from the multiphysics simulator COMSOL. We found differences in temperature between different simulators and identified an effect from the thermal boundary layer. The thermal boundary layer has a severe impact on microchamber cooling capacity due to the abstract thermal model. In a natural microchamber cooling system, the interconnection structures such as microbumps or micropillars mitigate the thermal boundary layer effects. Lastly, the simulation results of HotSpot have been validated with COMSOL simulation.
5. Providing thermal characteristics of modern chiplet and processing-in-memory systems. The thermal modeling and simulation of modern processors are investigated.

Microfluidic cooling is advantageous to the heatsink, especially when the high-performance processors are stacked face-to-face. The thermal behavior of processing-in-memory architectures is investigated and focused on the transition from 2.5D integration to 3D integration. The 2.5D chiplet has a processor and memories integrated side by side on the silicon interposer layer, and the 3D chiplet has memories stacked on top of the processor layer. Microfluidic cooling is superior in heat spread and alleviates the hotspots with higher power densities. Although we can increase the cooling performance by increasing the flow rate of coolant in the embedded channel or chamber, increasing the flow rate is accompanied by an increase in cooling cost and mechanical stress. Our proposed thermal management method is suitable for high-power 3D chiplets that are not relying on a higher flow rate. Firstly, we can add multiple microfluidic cooling layers between the dies. Because the microchamber cooling exploits the existing die cavities and 3D-printing package, we can integrate multiple cooling layers without additional hydraulic fabrication on silicon. Secondly, we can apply the novel microfluidic features such as the control direction of flows. By alternating the flow direction, the lateral thermal gradient problem can be solved. Lastly, for the current cutting-edge technologies such as processing-in-memory devices, the hybrid cooling method of heatsink on the top and microfluidic at the bottom shows superior thermal management performance.

1.5 Thesis Organization

The remaining chapters of this dissertation are organized as follows:

Chapter 2 presents the proposed thermal management method using a microfluidic chamber. We describe the limitations of conventional cooling methods of the heatsink, throttling, and microchannel. The Microchamber cooling structure exploits the existing die cavity as a fluidic chamber, hence achieving advantages in low-cost fabrication and fluid flow control.

Chapter 3 presents the experimental validation of the proposed microchamber cooling method. Details of experimental setup, device and system designs, and test results are covered in this chapter. Fluidic circuit modeling and design results are described. We show the flow rate range of given pressure with an affordable hydraulic pump with a reservoir.

Chapter 4 presents thermal circuit modeling of the modern chiplet cases. The properties of the chiplets are modeled as thermal elements and circuits to evaluate the thermal behavior. We investigate the thermal behavior of processing-in-3D-memory and high-rise 3D devices by using the thermal circuit modeling method.

Chapter 5 presents the thermal study result with HotSpot 7.0. We release the new version of HotSpot, including the microfluidic cooling feature and use in this chapter. Details of microfluidic behavior in the microchannel and microchamber layer are covered in this chapter. Simulation results of HotSpot 7.0 are validated with a multiphysics simulator.

Chapter 6 presents thermal simulation results of modern chiplets. We evaluate the operating temperature of high-performance 3D IC and processing-in-memory systems with conventional cooling methods and microfluidic cooling. Our results suppose that our proposed thermal management method provides a cost-effective solution for the thermal issues in modern chiplets.

Chapter 7 concludes the thesis. We summarize the major contributions and also discuss the future directions in this chapter.

2. Thermal management with Microfluidics

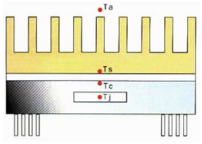
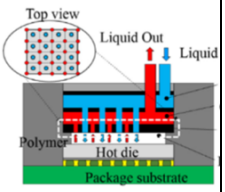
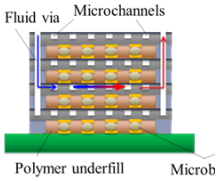
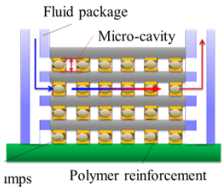
2.1 Microchannel cooling for 3D-IC

2.1.1 Conventional microfluidic cooling methods

Microfluidics is the technology of manipulating fluids in channels with dimensions from tens to hundreds of micrometers. Microfluidic cooling is a cooling method for semiconductors to remove heat directly from the silicon dies of the chip using microfluidics. The heat generated from the dies can be removed by coolant flow on the top surface or through the inside of the 3D stack. For the volumetric flow to cool down the volumetric heat from the stack, embedded cooling is desirable. Previous studies have defined embedded microfluidic cooling as intrachip and interchip cooling methods, in which the coolant flows inside channels or flows through the cavity between the dies, respectively [26]. The microfluidic cooling solutions are described below depending on the location where the coolant flows.

The first approach describes a cooling method that removes the heat from the chip by jet impingement onto the top surface of the die [15], [23]. This method can achieve highly efficient cooling by eliminating the thermal interface material and applying a high flow rate onto the die, which becomes a compatible cooling technique for high-power devices. In addition, the vertical inlets and outlets can eliminate undesired heat transmission due to lateral coolant flow. This structure, however, can be applied only to the top surface and is not applicable to 3D volumetric cooling.

Table 1. Comparison of different microfluidic cooling solutions for 3D-ICs

Cooling type	Heat sink	Direct Jet [23]	Microchannel	Microchamber
				
Structure	Assembly level	Package level	Die level	Package level
Cost	↓~↑	↓	↑↑	↓
Dimension	2D	2D	3D	3D
Hotspot cooling	X	O	X	O
Convection Thermal Resistance (°C/W)	Passive 4.0 High-end 0.2	< 0.1	N.A.	N.A.
Maximum Power		1000 W/cm^2	200 W/cm^2	200 W/cm^2
Fluid flow rate	N.A.	8.3 e^{-6} m^3/s		1.2 e^{-6} m^3/s
Pump Power	N.A.	1.3 W		-
# of Heat Exchanger (A cpu, B GPU, C Memory)	A+B+C	>1	>1	>1
Main-hurdle	3D-IC	3D-IC	Fabrication cost	Dielectric Coolant

The second microfluidic cooling solution is the intrachip cooling method. Heat is removed by the coolant flow through the microsize channels, which are integrated on the backside of the dies. The integrated microfluidic channels can offer stable and effective liquid cooling inside of the die. This method is extensively studied and verified in various prototypes [27], [28], and its simulator development has been completed [29]. However, in the case of 3D-stacked ICs, additional vertical fluidic interconnections between the dies are required to

enable vertical coolant movement [30]. This complex fabrication process becomes increasingly challenging as the number of stacks increases. In addition to the difficulties in the fabrication process, the microchannel cooling method conflicts with TSV by occupying the die area. The electrical interconnection between the dies, such as TSVs, micro-bumps, or Cu-pillars, occupies most of the chip dimensions in high-bandwidth devices. Therefore, the microchannel cooling method is inherently limited for multilayer 3D-IC with a large amount of inter-layer electrical connections.

Lastly, interchip cooling is another technique for cooling 3D-IC. In contrast to the matured intrachip cooling studies, interchip cooling has only been conceptually investigated [31]–[33]. Despite considerable success in microchannel cooling, a cost-effective and multilayer-compatible technology is required. Previously, we proposed a microchamber cooling method with a 3D printing package. We experimentally verified the feasibility of this method. By exploiting the inherent gap between the dies, this approach does not require additional fluidic structures and can save fabrication costs. In addition, this method is suitable for multilayered 3D stacked IC cooling because of its vertical scalability.

The aforementioned microfluidic solutions have advantages and disadvantages depending on their flow locations, as described in table 2. The jet impingement has the highest flow rate as well as a relatively low risk of physical stress. In addition, cooling can happen in local areas preventing the heat from spreading. Since the cooling feature is only on the top surface, the fabrication is low cost. However, this solution has a major disadvantage: it is not applicable to the 3D-IC. On the other hand, the embedded cooling methods, such as intrachip and interchip, are more compatible with 3D-ICs.

2.1.2 Disadvantages of Microchannel cooling method

The microchannel cooling method, also known as the intrachip structure, is considered the most promising microfluidic cooling solution [24]–[30], [34]–[38]. This method is advantageous for cooling high-power IC due to its direct cooling feature. However, it has inherent structural disadvantages in 3D-IC cooling applications.

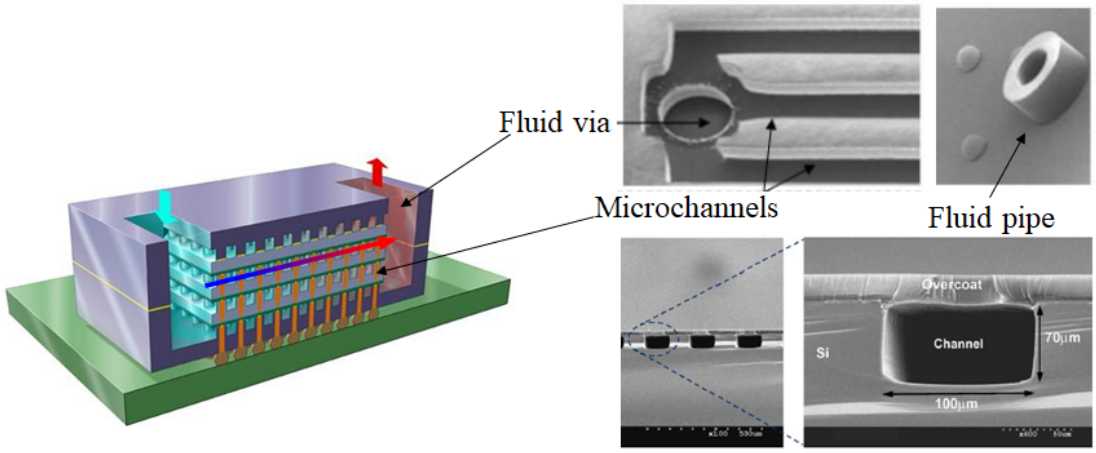


Figure 3. (a) Schematic diagram of 3D-IC with Fluid via and microchannels [6] and (b) images of integrated microchannels, fluidic via, and fluid pipe [11].

One of the disadvantages is the cost and complexity of the fabrication of the microchannels. The microchannel can be fabricated with a trench process on the backside of the wafer and sealed with a capping layer. In addition, fluid vias should be trenched to construct inlets and outlets to circulate the flow. Two additional patterning processes and trench processes are sources of increasing cost. Attaching a capping layer is another process obstacle in achieving low-cost requirements.

Another disadvantage is the vertical flow discontinuity. While microchannels can be configured in the lateral flow, additional structures are required to build vertical flow between the layers. For example, fluidic pipes can be constructed between the layers to connect the fluid. During the reflow process of micro-bumps, the distance between the layers is determined [24], [34]. Therefore, to prevent an electrical interconnection issue, the fabrication process of the fluidic pipes cannot affect the reflow process of the micro bumps. Additionally, since the fluidic pipes are several hundred micrometers in size, the die area needs to be expanded for the fluidic pipes to fit in. This additional structure affects not only the fabrication complexity and cost but also fluid circulation. The flow rates differ layer by layer due to the fluidic resistance difference.

The most significant disadvantage of the microchannel cooling method is the conflict between microchannels and existing interconnection structures. In the lateral floorplan view, the microchannels collide with through-silicon via (TSV) in the limited chip area. While the TSVs

have occupied the backside of a silicon wafer, there is no available space for microchannels to be trenched [39]. In addition to the lateral conflict, the microchannel method also conflicts in the vertical aspect. Because the TSV interconnects layers vertically, TSV needs to penetrate the entire layer. Hence, the capping layer should be attached before the formation of TSVs. The trenching process requires penetration through heterogeneous layers of a silicon wafer, an adhesion layer, and the capping layer. Additionally, the capping layer must contain metal pads for contact and needs to match the thermal expansion coefficient with the silicon wafer during the interconnection process. The conflict between the lateral and vertical structure impedes the realization of microchannels in 3D-IC applications.

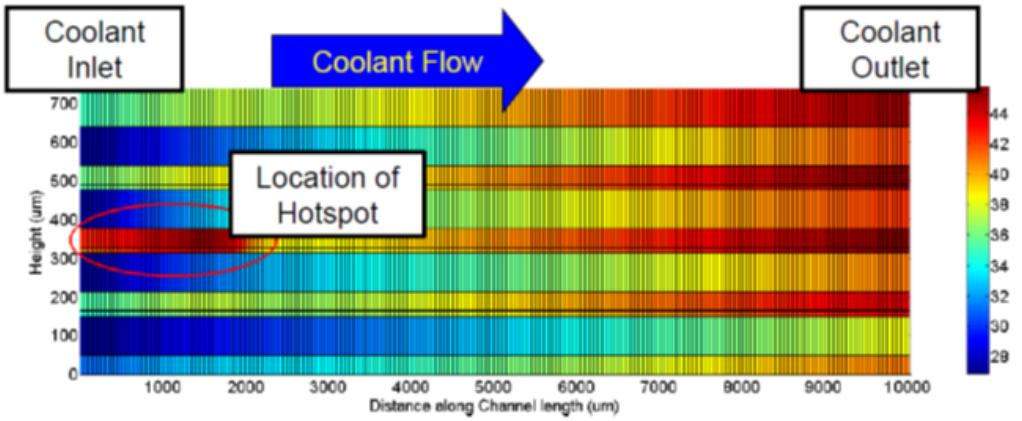


Figure 4. Heat Spreads due to fixed flow direction from 3D-ICE simulation [15].

In addition to the aforementioned structural disadvantages, the microchannel method has cooling disadvantages due to the fixed channel feature. The direction of flow in the microchannels is fixed due to the inlet and outlet being settled. This fixed direction flow generates a thermal gradient along with the flow direction. While the fresh coolant injected through the inlet has a low temperature, the temperature of the coolant gradually increases, exchanging heat with the target chip. Hence, there is a gradual temperature increase from inlet to outlet. Previous researchers tried to resolve this problem by non-uniform microchannel design[40], 4-port design[41], or arbitrary design [42] methods to suppress the temperature gradient. However, the hotspots in the modern chiplets are arbitrary and modified by the computation tasks. In a worst-case scenario, hotspots can be located near the inlet and outlet. The

coolant removes the heat from the hotspot near the inlet and diffuses the heat along with the fluid direction. When it comes to the hotspot near the outlet, the heated coolant exacerbates the hotspot. This thermal gradient problem is the major drawback of the microfluidic cooling system.

2.2 Low-cost microchamber cooling method

The microchannel cooling method has the advantage of direct cooling from the flow through the isolated microchannels. However, it is unsuitable for 3D-IC applications due to the inherent structural disadvantages. This study proposes a cost-effective microfluidic cooling method for 3D-ICs, also known as the microchamber cooling method. Instead of fabricating the microchannels, fluidic vias, and pipes on wafers, the proposed method utilizes the cavity that already exists between the dies and the 3D printing package.

A device structure of the previous microchannel cooling method is described in Figure 5 (a). The conventional approaches require microchannels along with the silicon dice and trenched thermal vias through the silicon dice for circulating fluid. Instead of fabricating additional structures, we propose a microfluidic cooling method that exploits the gap between the silicon dice. By using the empty space as a microfluidic chamber, the proposed method can achieve coolant flow. The device structure in Figure 5 shows the proposed device structure consisting of 3D stacked ICs and a 3D printing package. The method does not require additional micro-structure on the wafers for circulating fluid. Rather, the 3D-printing apparatus constructs inlet, outlet, and channels for the coolant flow of the stacked 3D-IC structure. This approach can reduce the process complexity and the cost by eliminating the additional structures on the silicon wafers.

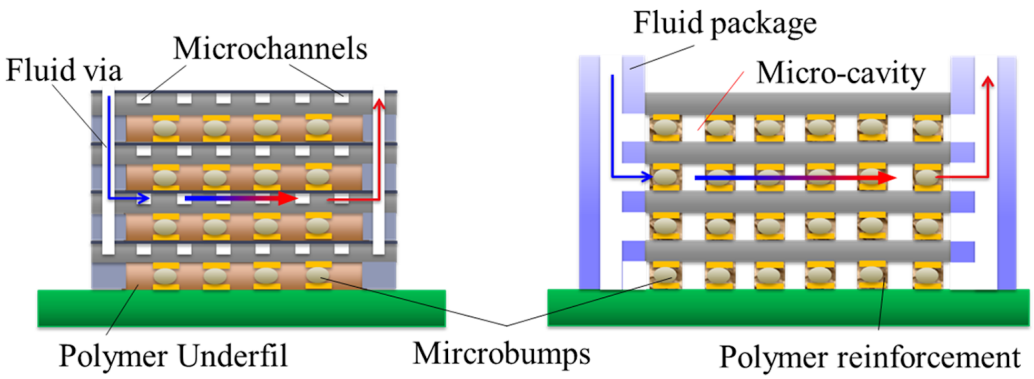


Figure 5. Schematic diagram of the microchannel and microchamber cooling method

In addition to the fabrication and cost advantages, this method has structural advantages in fluid flow control. The microchannel cooling system has a limitation due to the fixed one-directional straight coolant flow[40], [42]. After the microchannel is designed and fabricated on the backside of the silicon, the location and direction of the microchannels are fixed. Therefore, the microchannel cooling system has difficulties in managing arbitrary hotspots. This is especially an issue in some modern technologies, such as Processing-In Memory (PIM) or FPGAs, that can have thermal issues caused by multiple hot spots at changing locations. On the contrary, the microchamber is a multidirectional structure, a vacuous plane with sustaining Cu-pillars or micro bumps. By controlling inlets and outlets on the sides, we can configure optimized cooling features for multiple hot spots at changing locations.

3. Experimental validation of Microchamber cooling with 3D printed package

3.1 Device design

3.1.1 Microchamber cooling Structure

When the silicon dies are stacked up to construct a 3D-IC, the electrical signals between two layers are connected through micro bumps or Cu-pillars. This electrical interconnection leaves a gap between the layers, as shown in figure 6. Coolant flows through this empty chamber and can directly remove the heat from the chip. The microchannel device requires an airtight path, inlet, and outlet to circulate the coolant flow. On the other hand, the microchamber cooling system enables these features by using a 3D printing package. Figure 10 shows a chip surrounded by a 3D printing package. The 3D printing package encapsulates the airtight microchamber while the inlet and outlet are excavated for coolant circulation. The details of the 3D printing package design will be discussed in the following sections.

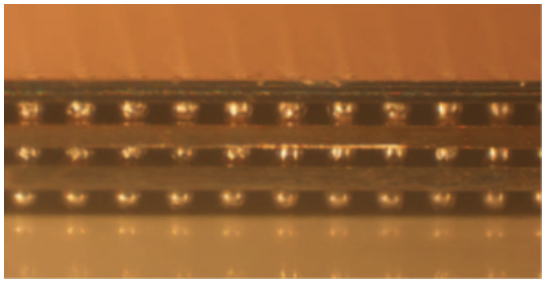


Figure 6. Side view of 3D-stacked IC [18]

3.1.2 Microchamber integration

For the thermal test, we chose an existing commercial chip instead of designing a custom chip. The chip used in this study is TTC-1002 by Thermal Engineering Associates, Inc. and provides control of heat flux generation and temperature measurement. Figure 7 shows the image of the thermal test chip. The thermal test chip has a 2540 μm × 2540 μm dimension: two heat generating resistors and four thermal sensing diodes. The six components allow the independent

heat flux generation and temperature sensing features in the test system. Twenty-four micro bumps of 80 µm in height were attached to the silicon chip before assembling the thermo-fluidic chamber. The test chip is fabricated on top of the PCB substrate using a reflow oven LPKF ProtoFlow. The stacked thermal chips in Figure 8 show the gap between the chips. The height of the chamber is measured using a digital caliper and a profilometer. The average gap between the four samples is 68 µm. Note that due to the reflow process, the height of the micro bumps decreased from 80 µm.

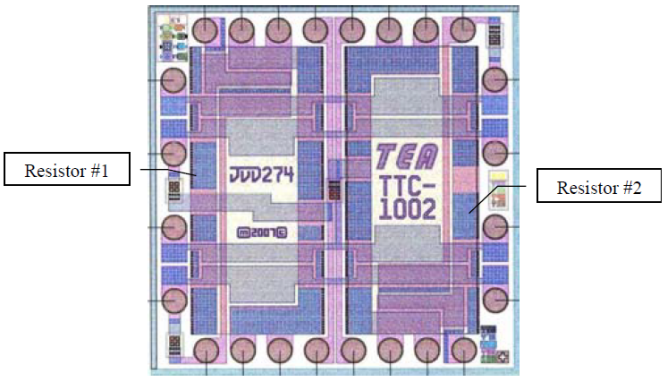


Figure 7. Layout image of thermal test chip [19]

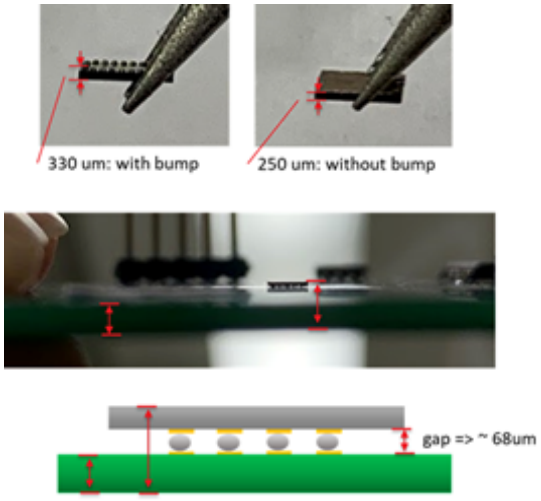


Figure 8. Images of thermal test chips, microfluidic chamber, and schematic diagram of the microfluidic chamber.

3.1.3 3D Printing Package design

For the 3D-printing package, we designed and fabricated three types of packages. The packages consist of the body attached to the substrate and the insert, which covers and seals the top surface of the chip. The inlet and outlet are fabricated into the insert located on top of the channel region. Figure 9 (a) is a rectangular shape insert with a 0.1-inch thickness to seal the gap in the upper direction. This design can encapsulate the gap, but the fabrication of the inlet and outlet hole was out of range for the accuracy of the 3D printer. Figure 9 (b) has two separate thin bodies and the gap between the two forms of a fluidic channel. The encapsulation on the top can be made with a flat glass attachment. However, this structure has a void between the top glass and the chip to secure an assembly margin. Figure 9 (c) is used in the experiment, which is an improved version of the type A package. The insert has a trapezoid shape aligning with the fluidic channel and 0.02-inch thickness. We could reduce the thickness of the insert while maintaining the encapsulation using UV polymer sealing on the top.

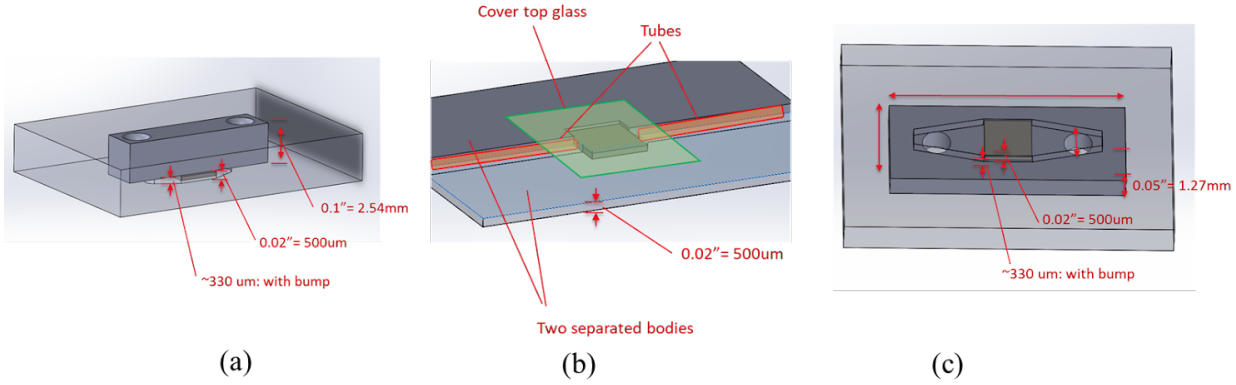


Figure 9. Three types of 3D-printing packages

3.2 Experimental setup

3.2.1 Driving system

To investigate the fluidic-cooling experiment, the thermal test chip was directly mounted on the printed circuit board. The gap between the chip and the board constitutes the thermal chamber in this configuration. A Fused Deposition Modeling (FDM) type 3D printer Stratasys

F170 is used for printing the apparatus. Figure 10 shows the 3D design result using SolidWorks, and Figure 11 shows the image of the actual device. The device has an inlet and an outlet for circulating the coolant. It also embeds a distribution structure for the coolant and acts as a cue for the fluid chamber structure.

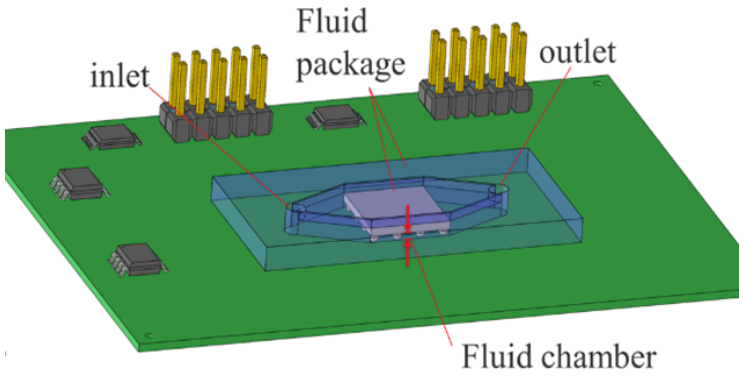


Figure 10. Design of microfluidic chamber experiment system

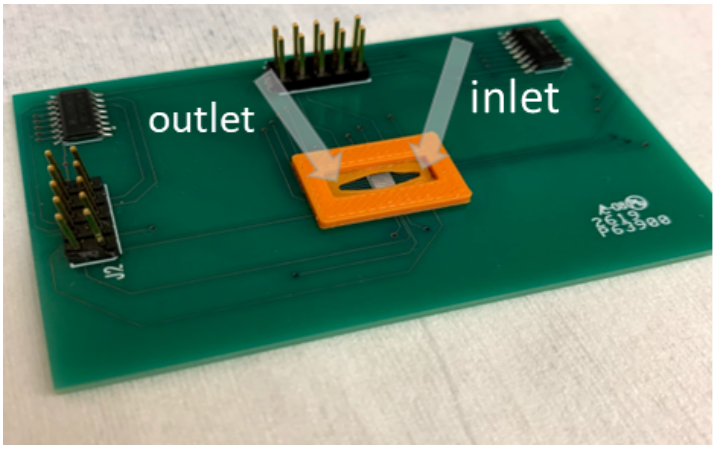


Figure 11. Image of the microfluidic chamber device using 3D printing package.

The temperature of the test chip can be controlled and sensed by the electrical system. Two thermal resistors on the chip generate the heat flux up to 6W each. An external power source is connected to the system to sustain a stable power supply. Texas Instrument AMC 7823 is an analog monitoring and control system with 8 DACs, 8 ADCs, and a constant current source.

Under a constant current driving condition, the forward voltage of sensing diodes varies with the temperature change. The forward voltages of the diodes are read through the 12-bit ADC of TI AMC 7823. The sensing diodes have a nominal forward voltage of 0.71 V under a current feeding of 1mA per diode. TI AMC 7823 is reconfigured to generate 1mA of the precision current source.

Heat flux generation and temperature sensing operations are controlled by a Raspberry Pi 3 Model B system. The TI AMC 7823 and the Raspberry Pi communication are connected through a Serial Peripheral Interface. While the system has 8 ADCs to sense the forward voltage, only one current source exists. Hence, the current feeding is required to be shared with four sensing diodes. The current is shared with time-division multiplexing using the TS3A5017 analog switch. The switches are controlled by the Raspberry Pi with a time margin of 10 ms.

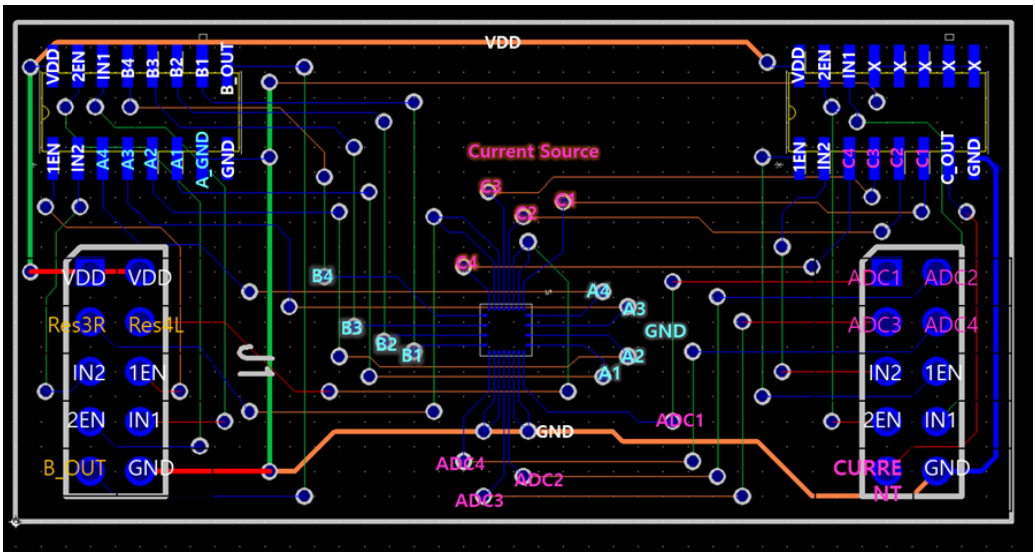


Figure 12. PCB Layout image of the thermal testing board

3.2.2 Cooling and sensing system

A dielectric coolant of EC120 ElectroCool® was used as the coolant for the fluid chamber. The dielectric coolant has a high resistivity of 1x1014 Ω/cm and dielectric constants of 2.1. Dielectric coolant is required to ensure proper electric insulation of the micro bumps that act as pillars between stacks. The dielectric coolant has a kinematic viscosity of 5.02 cSt, which is higher than the 0.658 cSt of deionized water at 40°C.

A constant current of 100 μA drives the thermal sensing diode, and a heat flux of 0.5W is applied to the resistor. The forward voltage of the diodes decreases when the resistor generates heat. The forward voltages of the diodes are sensed using the 12-bit ADC. The change in temperature can be represented by the difference in the forward voltage with a correlation constant K. The equation is written as

$$K = \left| \frac{T_{High} - T_{low}}{V_{low} - V_{High}} \right|$$

The unit of constant K is °C/mV, and the calibrated value in the experimental set-up is 0.2°C/mV.

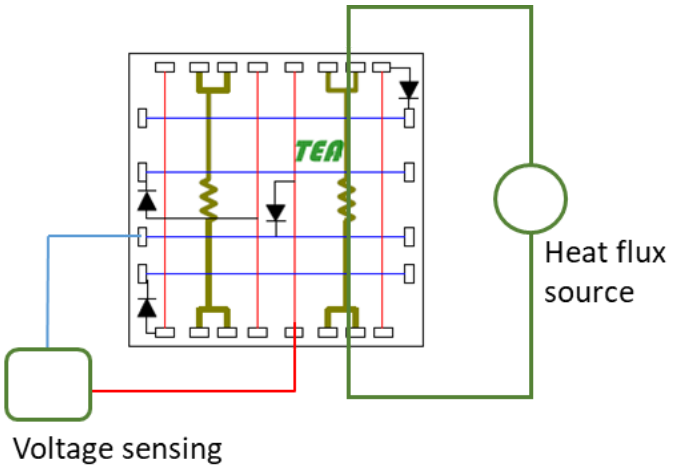


Figure 13. Schematic diagram of driving and sensing system image

3.2.3 Fluidic pump

The microchamber cooling system is composed of a fluidic pump, driving–sensing system, thermal test board, and control system. A Fluidic pump is a source to generate the flow through the microchamber and remove the heat from the chip. A paper pump is commonly used in microfluidic experiments using passive elements and is simply absorbing fluid using a pulp's capillarity. Although the paper pump is valid for preliminary research, it lacks precise control of fluidic flow. Next, microfluidic syringe pumps are advantageous in precise flow rate control. However, the syringe pump has limitations in maximum flow rate and transient cooling because the coolant is flowing from one syringe to the other. Lastly, a pump with a reservoir is

advantageous in the cooling system. The liquid cooler is commercially available in computer systems and has a closed loop system composed of the radiator, cooling fan, reservoir, and pump. The closed loop system provides a continuous flow rate and chilled coolant through the radiator. In addition, the system already has a compact size and appropriate power supply level, which can be applied to consumer electronics. The disadvantages of the system are the precision of the flow rate while the flow rate is controlled by pulse-width modulation (PWM) signal, not by the hydraulic aspect. We need fluidic circuit modeling to calculate the expected flow rate in the cooling system.

3.2.4 Fluidic circuit design

The flow rate and pumping requirement for the cooling system is another design factor to consider in the microfluidic system. Although a higher flow rate could be beneficial for cooling capacity, a higher flow rate can cause mechanical stress on the channel or burden on the pump. Figure 14 shows the designed cooling system and its fluidic circuit model. The flow rate of the fluidic cooling system can be calculated by using the fluidic circuit and Hagen-Poiseuille equations[43]. The microchamber has 68 μm in height and is square in shape with an area of 8 x 8 mm^2. The hydraulic resistance (Pa-s/m^3) of the rectangular-shaped microchamber can be calculated from the equation,

$$R_{hyd\_channel} = \frac{12\eta L}{1\text{-}0.63(h/w)} \frac{1}{h^3 w}$$

where viscosity η is 1.05 × 10^{-3} Pa·s, length L is 8mm, height h is 68 μm, and width w is 8mm. Also, the hydraulic resistance of the inlet-outlet pipes can be calculated from the equation,

$$R_{hyd\_pipe} = \frac{8}{\pi}\eta L \frac{1}{a^4}$$

where viscosity η is 0.00105 Pa·s, length L is 0.01m, and radius a is 0.5 mm. The hydraulic resistance of the system is 6.65 x 10^{10} Pa-s/m^3, where the internal hydraulic resistance of the pump is 2.15 x 10^8 Pa-s/m^3, and the hydraulic resistance of the inlet and outlet is 2.68 x 10^{10} Pa-s/m^3. The required pumping hydraulic pressure for the flow rate is described in figure 15.

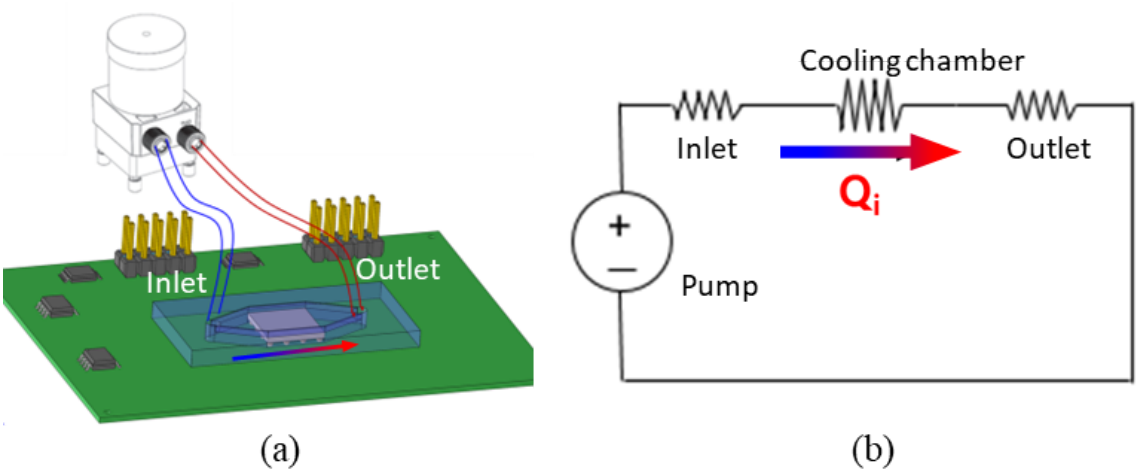


Figure 14. a) 3D-printing design and (b) schematic fluidic circuit diagram of microchamber cooling system

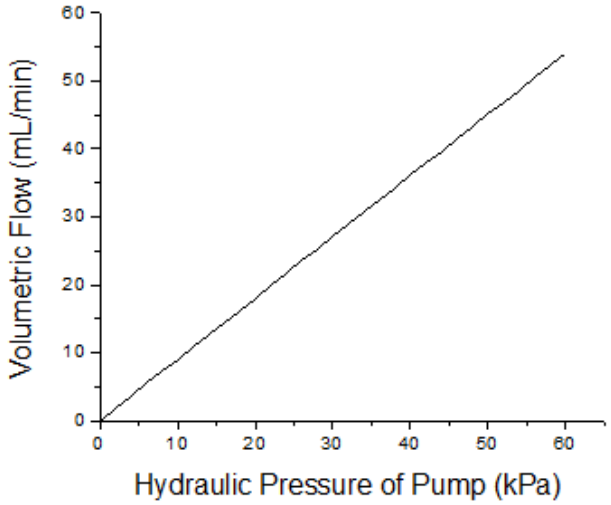


Figure 15. The flow rate through the microchamber with given Pump power.

3.2.5 Prototype

The previous experimental results show the validity of the proposed cooling method. In order to deliver the feasibility of the proposed method, we built a prototype in figure 16 that conceptually shows the thermal management system with microfluidic chamber cooling. The thermal management system controls electrical, thermal, and microfluidic parts simultaneously. The system drives the thermal test chip and senses the thermal information from the test chip.

The thermal test chip generates heat flux while sensing the temperature through the sensing diodes. The thermal sensing data can be used to control the flow rate. We expect the thermal feedback, 3D-printing package, and closed-loop cooling system to be a low-cost thermal management solution for the modern high-performance ICs such as 3D-IC and processing-in-memory devices.

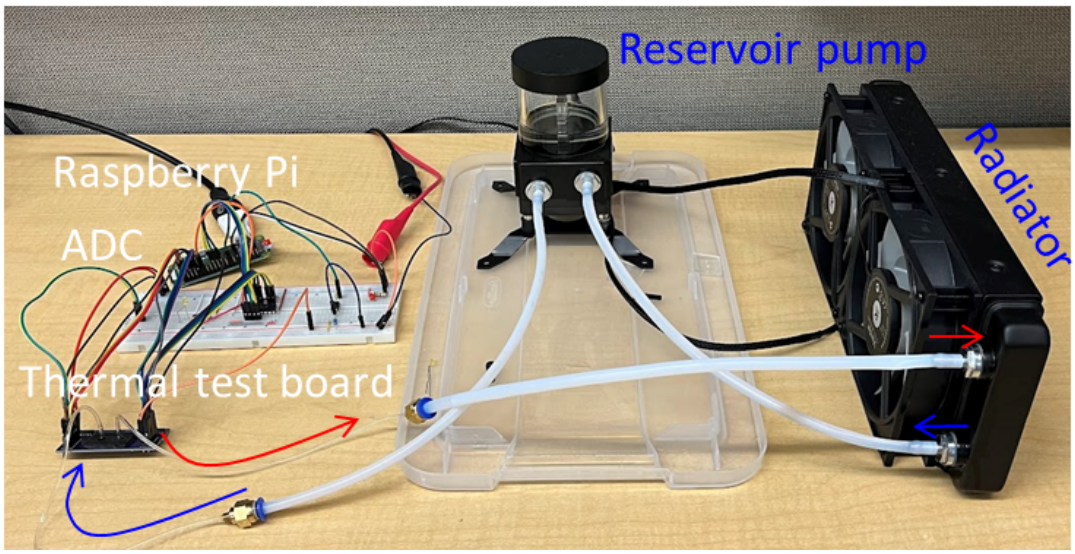


Figure 16. Image of fluidic circulation prototype

3.3 Result and discussion

3.3.1 Microfluidic cooling through the fluidic chamber

Figure 17 shows the converted temperature obtained as the product of the forward voltage and constant K. After 90 seconds of the heat flux generation period, the flux generator is turned off. The black and red curves are the baselines to show the temperature behavior with and without heat flux generation. 10 µl of the dielectric coolant at 10ºC is injected through the fluid chamber to wet the surface and facilitate the fluid flow. A micropipette tip was used as a coolant reservoir at the inlet, and a passive pumping mechanism was placed at the outlet. The passive pumping mechanism we used consisted of a paper pump that acts upon the capillary effect and the porosity of the paper to withdraw coolant from the fluid chamber.

When the chilled coolant flows for 60 sec, the temperature drops 10°C for 3 seconds and is re-heated by the heat flux generator. The blue curve of figure 18 shows the cooling effect of the coolant flow. This change of temperature due to the coolant flow confirms that the pumping effect of the coolant through the microcavity effectively cools down the thermal test chip.

The green curve of figure 18 shows the microfluidic cooling behavior of the fluidic chamber method. During the 90 seconds heat flux generation period, coolant flow is applied at 60 seconds. The paper pump is used at an outlet, and a 1ml pipette tip is attached to the inlet. The 1ml of coolant gradually runs down by gravity until the pressure between inlet and outlet is balanced, and then the paper pump is placed at the outlet with a primer coolant drop to enable paper pumping by capillary effect. During 14 seconds of the cooling period, despite the 0.5W heat flux being generated, the temperature gradually decreased by 14 °C and started to be re-heated by the heat flux generator after the end of the coolant flow. These preliminary results show that our thermo-fluidic chamber cooling method is feasible for a 3D stack of IC.

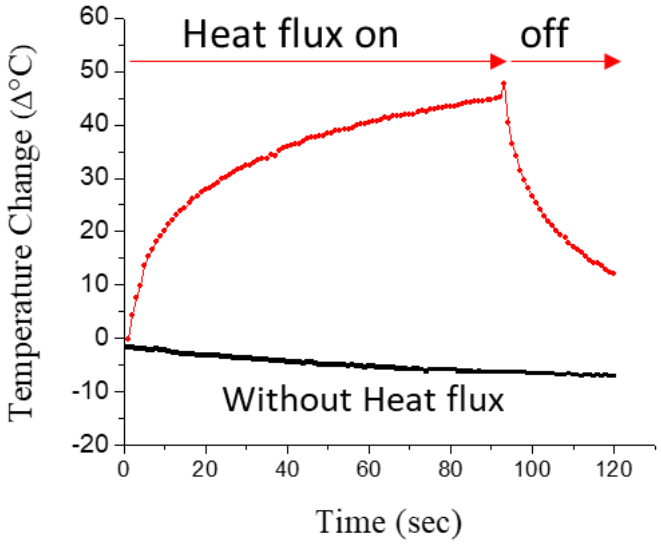


Figure 17. The converted temperature change from the forward voltage of the thermal sensing diode of the thermal test chip baselines without microfluidic cooling

(a)

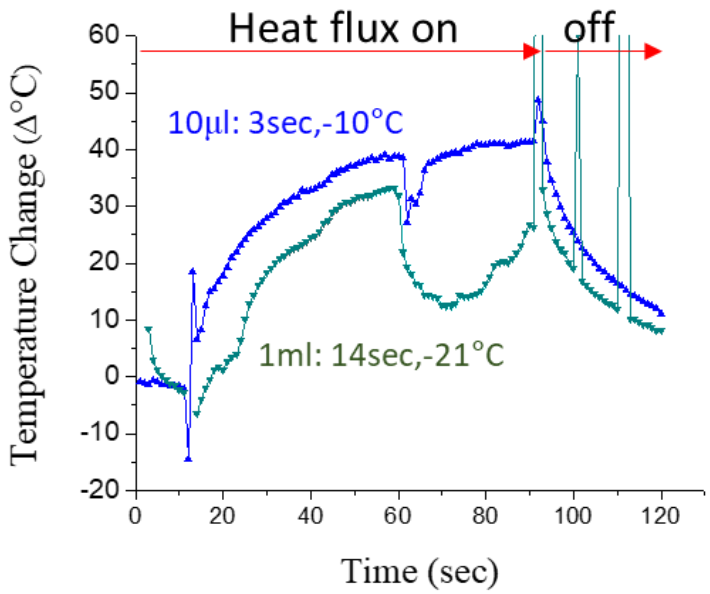


Figure 18. The converted temperature change from the forward voltage of the thermal sensing diode of the thermal test chip with microfluidic cooling.

3.3.2 PDMS package

The aforementioned experiment demonstrates that the microfluidic chamber method with a 3D printing package is a valid alternative to the micro-channel method. This method can achieve controllable coolant flow with a low-cost fabrication process. Structural and material improvements are required to achieve a watertight 3D-printing package for microfluidic cooling of 3D-IC.

One of the main parts of the fluidic system is a watertight package. We investigated a Polydimethylsiloxane(PDMS) package made with SU-8 molding. Figure 19, 20, and 21 show the fabrication steps, designing results, and fabrication results.

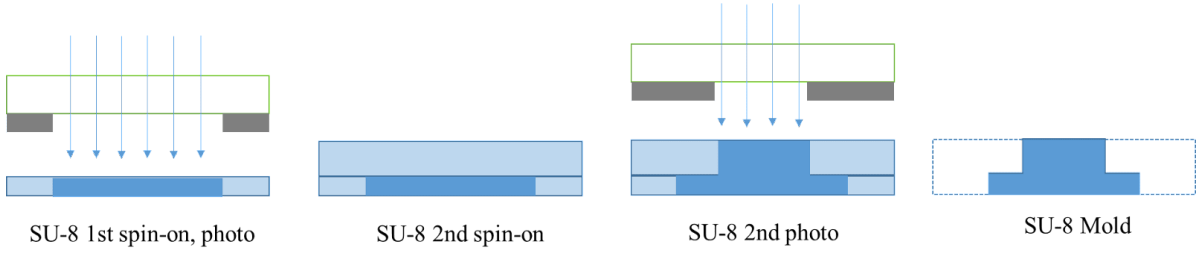


Figure 19. The process step of the SU-8 molding and PDMS package for microfluidic cooling

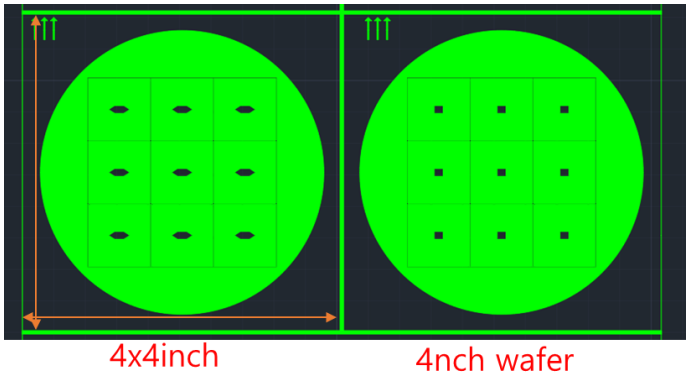


Figure 20. Photomask design of SU-8 molding

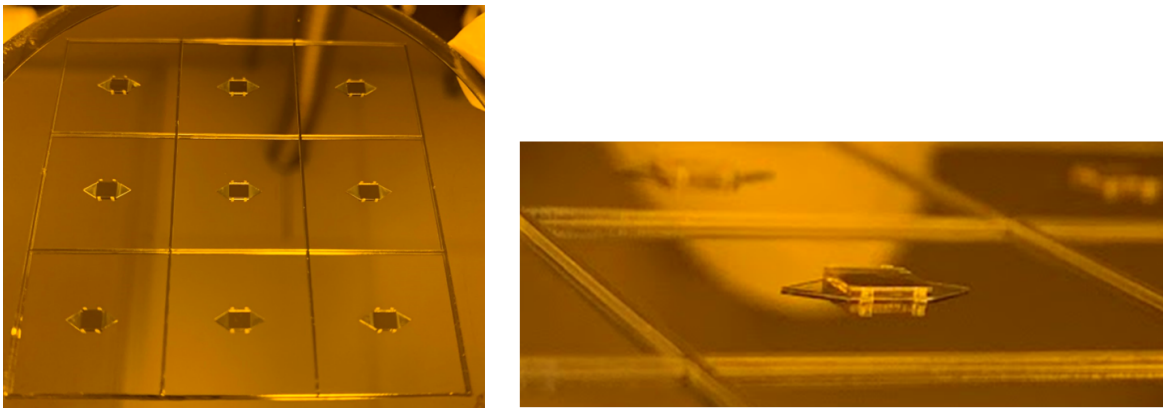


Figure 21. Images of the fabricated SU-8 molding

As a UV curing adhesive, we used Norland Optical Adhesive(NOA) 78. This curing adhesive is widely used in the display industry offering easy handling with high viscosity of 9000 CPS and a tight bond between plastic-to-plastic or glass-to-plastic. The disadvantage of the NOA 78 is yellowing color after curing, but it did not affect the microfluidic experiment. However, the adhesion between PDMS and PCB surface was not enough to sustain the watertight package. The surface flatness and existing via holes under the curing area are not suitable for the attachment of the PDMS package. In addition, controlling issues of thickness and placement of UV polymers may increase the fabrication cost of the proposed method. Because of these fabrication issues, we consider the 3D printing package as a low-cost cooling package candidate. The conventional package has a rough surface which affects microfluidic flow. By

using a specific microfluidic 3D printer, a solid 3D printing package with higher resolution enables a low-cost cooling solution. We suggest the following improvements to ensure the encapsulation and fluidic channel region. First, the 3D printing package depends upon a UV transparent material. The current 3D printing package blocks the UV rays and obstructs the curing process. Second, the surface around the thermal test chip of the PCB requires a flat area without holes. Lastly, the precise UV dispensing and curing process is required. Excessive or misplaced UV adhesive causes failure in encapsulation or encroachment to the channel region. These requirements can be improved with commercially available adhesive dispenser systems and do not significantly increase the fabrication cost of the system.

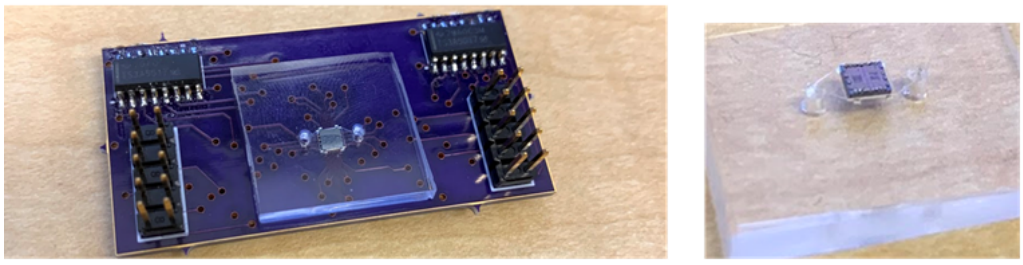
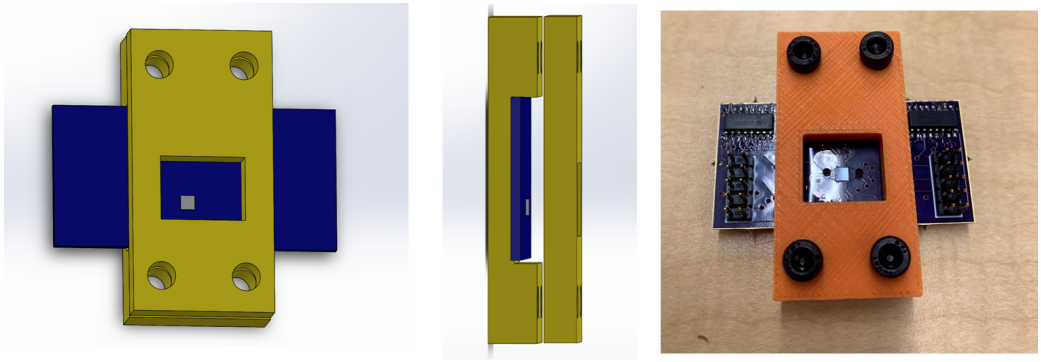


Figure 22. Images of the PDMS package are attached to the thermal testing board.



(a) (b) (c)

Figure 23. (a),(b) Design of PDMS package with a 3D-printing jig and (c) Fabrication result of PDMS package with a 3D-printing jig

3.3.3 Multiplexing flow method

The validity of the low-cost fabrication of the proposed cooling system has been shown in the previous chapters. As well as the cost-effectiveness, the microchamber is advantageous in

the cooling capacity aspect. The conventional microchannel method has the disadvantage of fixed flow and consequent thermal gradient problems. On the contrary, the microchamber has an open floor plan, and it can be applied to the multiplexing flow control method. The multiplexing method controls the on and off of inlets and outlets. Accordingly, the flow directions and locations can be controlled by the thermal management system. In figure 24 (a), the hotspots on the left pane are located, one near the inlet and the other near the outlet. In this case, the coolant gets heated at the first hotspot and spreads the heat along with the streamline. Moreover, at the second hotspot, the heated coolant exacerbates the second hotspot. Figure 24 (b) shows the inlet and outlet multiplexing method, avoiding the thermal gradient by changing the flow streams. This thermal management method is a promising way to solve the arbitrary hotspot problem of modern ICs.

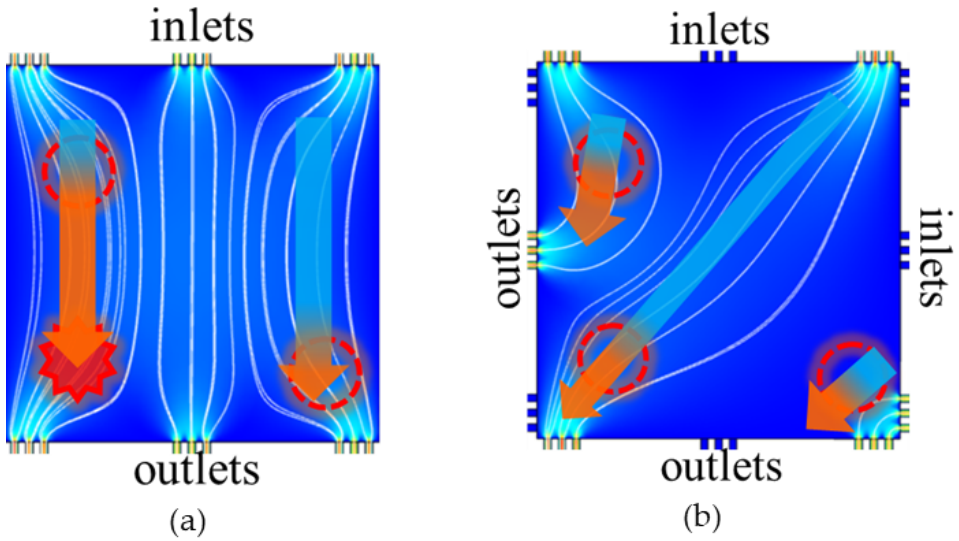


Figure 24. (a) thermal gradient problem with fixed flow direction and (b) multiplexing flow control to manage arbitrary hotspots.

4. Thermal modeling

4.1 Thermal Circuit modeling

4.1.1 Thermal resistance

Thermal modeling is a significant process of thermal management to evaluate a variety of techniques for investigating the thermal issue. The properties in the ICs can be modeled as thermal circuits by using a duality between heat transfer and electrical phenomena [44]. Thermal resistance is defined as the ratio of the temperature difference, dT, to the heat transfer Q.

$$R = \frac{\Delta T}{Q}$$

In addition, the heat flux generated in the ICs can be described as a power source in the circuit, and the heat transfer can be described as a thermal current flow through a thermal resistance. Hence, the temperature difference can be described as voltage. For example, the processor core, cache, and I/O blocks can be modeled as separate heat flux generators according to their power consumption. In the DRAM case, each vault and memory controller unit can be modeled as a heat source. Based on the thermal circuit elements, the 3D-stacked DRAM can be modeled as a circuit. The thermal circuit of the 3D stack of DRAM vaults comprises heat source, conduction resistance, and convection resistance. In this model, the heat generated from the memory stack and heat dissipation to the top surface are ignored. The conduction thermal resistance $R_{Conduction}$ varies with the thermal conductivity of the silicon wafer. The convection thermal resistance $R_{Convection}$ represents the heat transfer from the silicon wafer to the ambient. The following equations express these thermal circuit elements.

$$R_{Conduction} = \frac{t_{Si}}{k_{Si} \cdot Area}$$

$$R_{Convection} = \frac{1}{h_{Air} \cdot Area}$$

4.1.2 Thermal Modeling of thermal test system

The temperature of a system with microfluidics can be calculated with thermal circuit modeling including fluidic circuit modeling in the cooling aspect. In this chapter, we evaluate the

the required volumetric flow at given power consumption to maintain the operating temperature under 85°C. The flow rate in the channel was calculated using the Hagen-Poiseuille equations [43] with channel height of 68µm, length of 8mm, and width of 8mm, as well as inlet and outlet pipes. The power is applied and consumed in the heat flux generator. The heat flux is generated and conducted across the chrome layer with Conduction of 93.7 W/(m·K). The heat is conventional through the coolant liquid boundary layer from the test chip. The thermal modeling results are shown in figure 25. Figure 26 and 27 show the thermal circuit and fluidic circuit calculation result with different coolants. From the thermal circuit calculation result, the EC120 coolant can cool down a power density of 140W/mm^2 consumed in the thermal testchip maintaining the operating temperature under 85°C under the volumetric flow of 1.67 × 10-7 m^3/s with 200 mbar pump system.

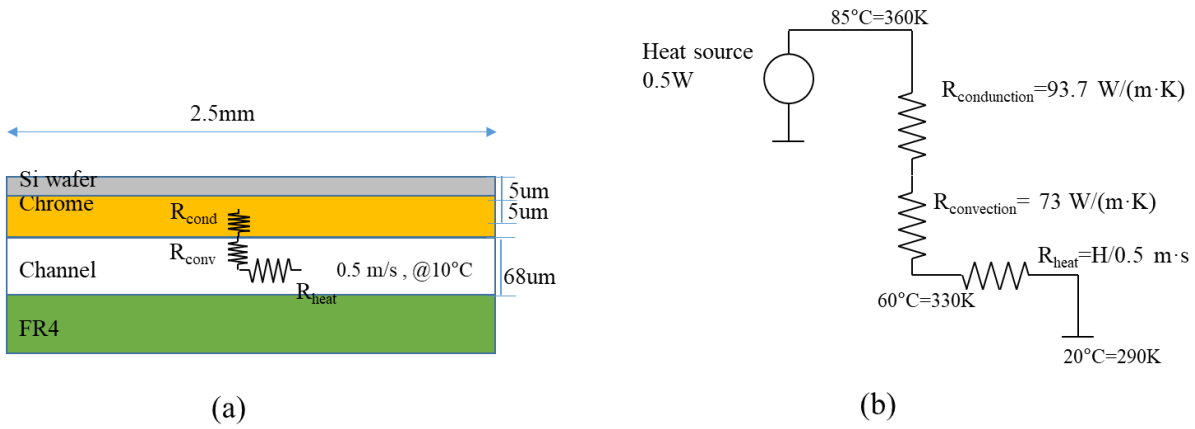
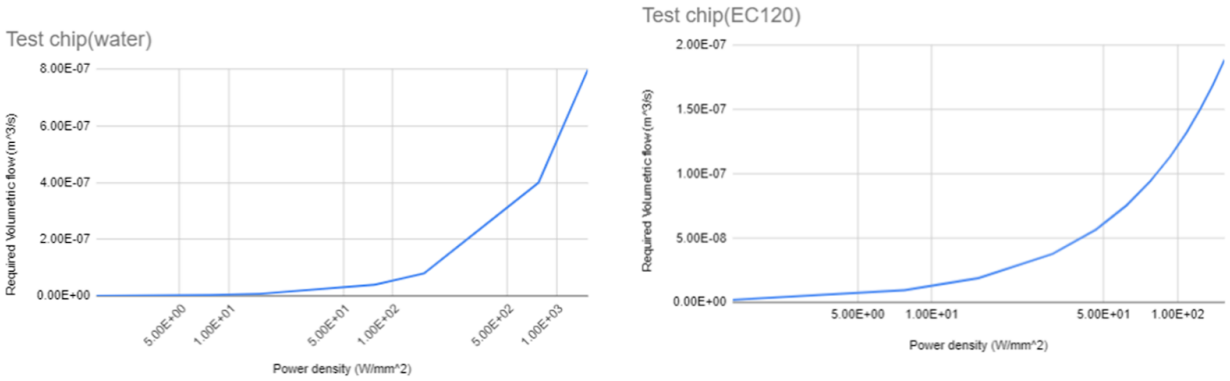


Figure 25. (a) Cross-section diagram and (b) Schematic diagram of a thermal circuit model of the thermal test system



(a) (b)

Figure 26. Required volumetric flow to maintain the operating temperature with increasing power density in the test chip with coolant of (a) water and (b) EC120.

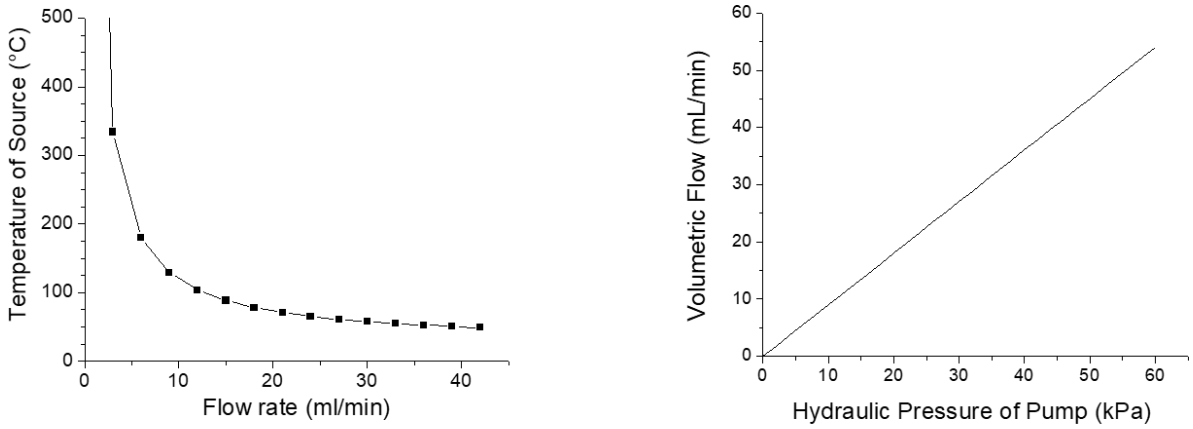


Figure 27. The temperature of the source with given flow rate and the volumetric flow rate for given pressure supply.

4.2 Thermal circuit model of 3D IC

4.2.1 Thermal issue of 3D-IC

A typical example of 3D-IC would be a 3D-stacked memory such as High Bandwidth Memory (HBM), which can achieve higher memory density and shorter data movement by stacking dies [19], [21], [45]. The stacked logic and memory layers are a key basis for integrating in-memory computing architecture. However, one of the critical challenges of integrating in-memory-computing architecture in 3D-stacked memory is the thermal issue.

In-memory computing brings about severe thermal problems in the 3D memory, generating volumetric heat in the stack, and causing increased heat flux. However, the cooling capacity is restricted to the top surface area of 3D-IC. In turn, the lack of cooling capacity limits power consumption in the 3D-IC. The power limitation suppresses the performance of in-memory computing architecture.

4.2.2 Thermal Modeling of 3D-stacked memory

The conventional passive heat sink is attached to the top surface, as shown in figure 28 (a). The logic and memory dies attached to the silicon interposer and a substrate. Figure 28 (b) shows the thermal circuit model of the system. The thermal convection to the bottom direction is negligible, and the heat dissipation in the logic layer is modeled as a heat source.

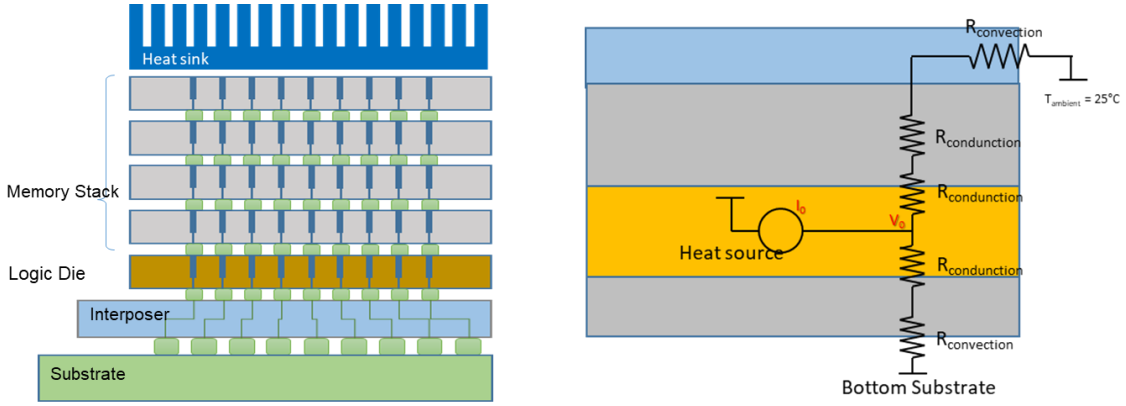


Figure 28. (a) Cross-section diagram and (b) Schematic diagram of a thermal circuit model of the 3D-stacked memory

The thermal circuit-based temperature calculation is a fast and straightforward way to predict the memory's temperature with different power consumption configurations. For example, the logic layer at the bottom can integrate arithmetic functions to increase processing performance, in particular in near-memory computing architecture. The additional integration of the processing function increases power consumption in the logic layer. The calculation result of the 3D memory is described in figure 29. The temperature of the logic layer rises as power

density increases. The increasing trends have different slopes due to the cooling features. The passive heat sink has the least cooling capacity of 4.0 K/W. The maximum available power budget of the memory system with heatsinks is described in table 2. The active heat sinks have more robust cooling capabilities of 2.0, 0.5, and 0.2 K/W for low-end, commodity-server, and high-end-server active coolers [18]. As the cooling capacity of the heatsink increases, the power budget in the memory system increases.

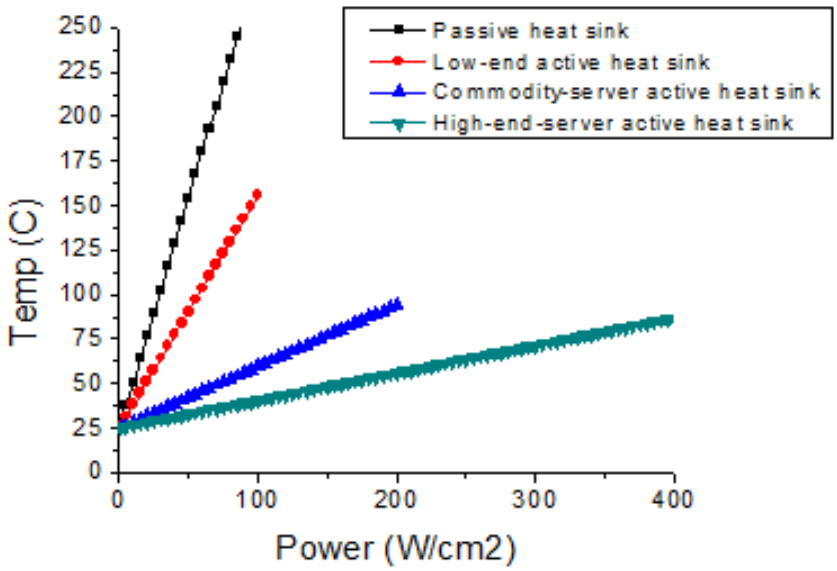


Figure 29. The temperature of the logic die in the 3D-stacked memory

Table 2. Simulation result of 4-layer stack

Cooling Method	Max. power < 85°C	
	Total	Logic layer
Passive heat sink	23.2 W/cm^2	14.9 W/cm^2
Low-end active heat sink	45.9 W/cm^2	29.4 W/cm^2
Commodity-server active heat sink	173 W/cm^2	110.7 W/cm^2
High-end-server active heat sink	388 W/cm^2	248.3 W/cm^2

4.2.3 Thermal Modeling of Processing-in-Memory

Recently, advanced memory technologies such as HBM achieved performance enhancement by stacking memory layers on top of the logic layer. The 3D memory system has a limited power budget because of the inherent thermal issue. While heat is generated in the volume, the cooling capacity is limited to the heatsink attached on the top surface. Due to the heating and cooling imbalance, the temperature increases. The maximum temperature allowed for operating a DRAM system is 85°C. If any hotspots in any layer exceed the limit temperature, it causes thermal failure in the memory system. Consequently, the power budget is determined by the maximum temperature of the memory layers. The total power density of the memory system with four layers of memory on the logic die is limited to 17.3 W/cm^2 and 11.2 W/cm^2 dissipated in the logic layers [18]. This tight power budget of 3D-stacked memory restricts integration of in-memory computing architecture. Recent research shows that in-memory computing architecture requires more power budget in the memory systems [46], [47]. Server-on-chip [46] proposed two memory dies stacked on a logic layer structure, which consumes 16W/cm^2 in the logic layer. Tesseract [47] accomplished a 30X speedup with PIM while the power density in the logic layer is 33.2 W/cm^2. Despite the computing advantages, the increase in power consumption obstructs the integration of Processing-in-3D-memory. Consequently, an extensive understanding of the thermal behavior of the 3D-IC is required. The modeling and simulation of thermal behavior can be investigated from the most superficial level of thermal circuit to the most elaborate level of multiphysics simulation. We propose a simple thermal circuit model for PIM integrated 3D-stacked IC. With a simple model simulation, we can predict the power budget of many layers of 3D-stacked in-memory computing. Moreover, we can offer a power consumption guideline for in-memory computing in either the logic or memory layers.

The processing unit can be integrated into the logic layer or in-memory layers [48]–[50]. In the first case, most of the power of in-memory computing architecture is consumed in the logic layer alone. When we assume that the power consumption is weighted on the logic layer, the thermal circuit can be modeled with one heat source at the logic layer. The near memory computing model is the same configuration as the figure 28. The power dissipation from the logic layer is modeled as a heat source, but from the memory layers are neglected. If the in-memory computing function is distributed over the memory dies, then each memory layer's

power consumption should be considered as well. Figure 30 shows the thermal circuit model of the in-3D-memory computing device. This model offers a glimpse of the power budget design of in-3D-memory computing devices.

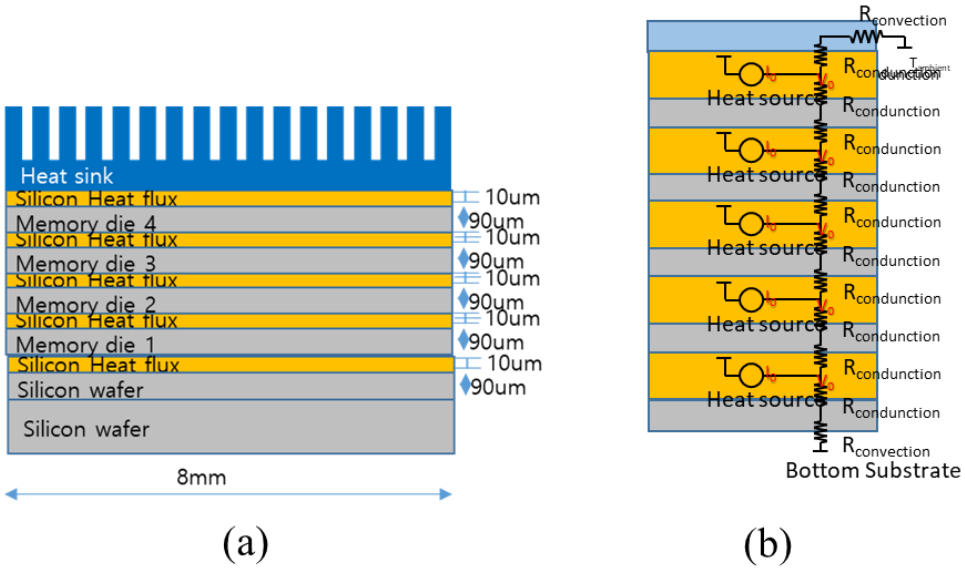


Figure 30. Schematic diagram of (a) thermal circuit model of the 3D-memory with PIM on the memory dies and (b) simplified circuit

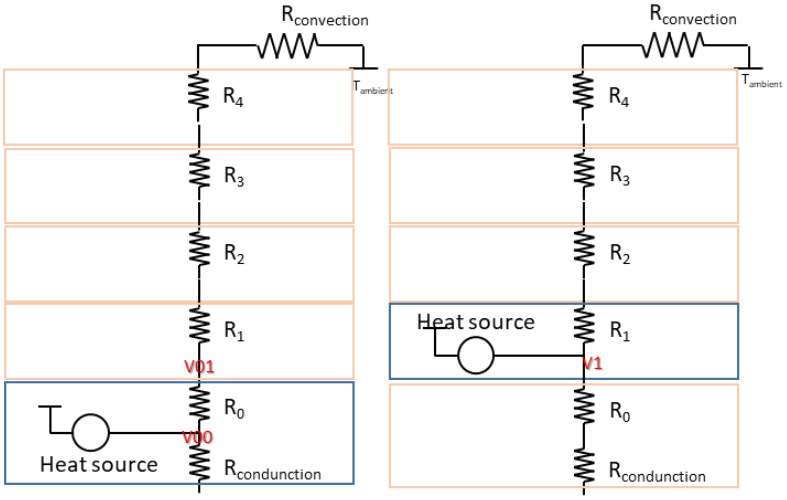


Figure 31. Superposition to calculate the temperature of each nodes

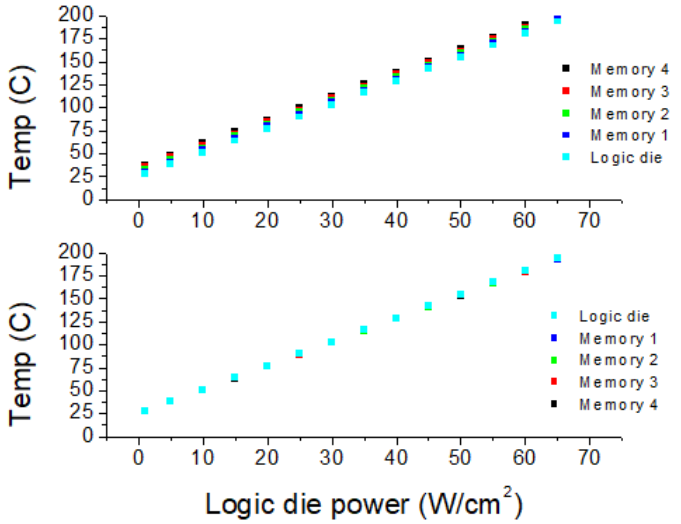


Figure 32. The temperature of the logic and memory layers with power density in each logic layer of 1 W/cm^2 (up) and0.01 W/cm^2 (down)

Figure 31 shows the superposition of heat sources to calculate the temperature of each node. Figure 32 shows the resulting temperature of logic and memory layers with increasing logic die power density. When the power consumption in each memory layer is negligible, the temperature of the logic die and memory dies are identical. However, when the logic layer consumes at least 1 W/cm^2, each model predicts a different logic layer temperature. Figure 33 shows the results from two different simulators. The fourth memory layer on the top is the highest, while the logic layer temperature is the lowest. This temperature variation is due to the superposition circuit calculation method. However, the overall temperature value and trends of the two simulators are identical. The result of the thermal circuit model calculation is validated using the simulation results from 3D-ICE [29].

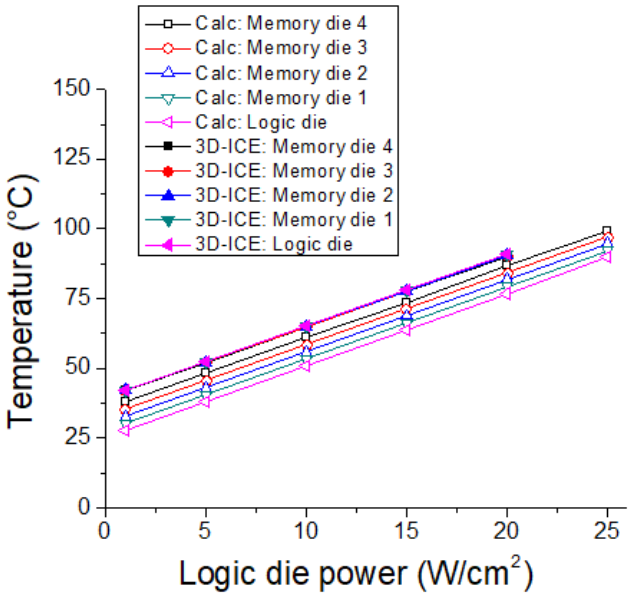


Figure 33. Comparison of simulation result from thermal circuit calculation and the 3D-ICE

4.2.3 Thermal Modeling of high-rise 3D-IC

Based on the thermal modeling investigation, we simulated high-rise 3D-stacked memory for potential power budget design to integrate in-memory computing. The proposed thermal modeling method is easily scalable to the many layers of 3D stacks. We simulated 16, 64, and 128 layers of 3D-stacked memory for the potential power budget design to integrate in-memory computing. A passive heat sink can sustain the operating temperature for a maximum stack of 4 layers. For more than four layers of the stack, an active heat sink is required. Figure 34 shows the abstract thermal model of 3D memory of 16 layer stack. In the case of 16 layers, the low-end active heat sink allows the memory system to operate under 85°C with 2W/cm^2 in each logic layer and 14.4 W/cm^2 total power. However, the low-end active heat sink is only sufficient for conventional memory operation. The simulation results are described in figure 35 and table 3. The low-end active heatsink cannot manage the system for the normal memory operation power budget. More cooling capacity is required to accommodate the high power dissipation of in-memory computing architecture.

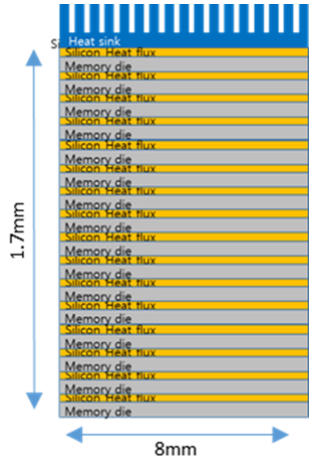


Figure 34. Cross-section diagram of 16-layer 3D-memory with PIM on the memory dies

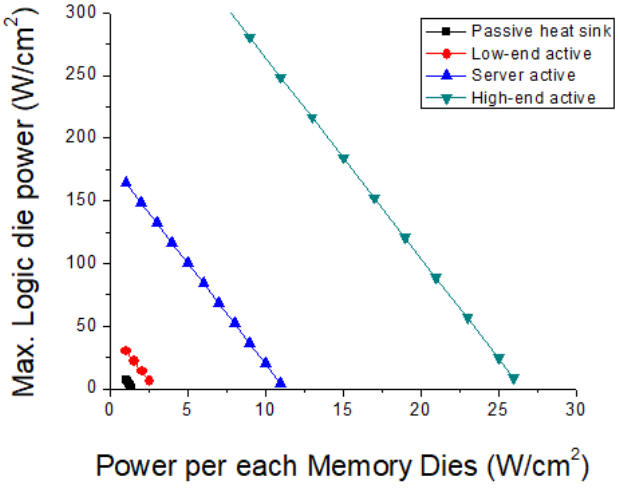


Figure 35. The simulated temperature of the logic layer with different heat sink features.

Table 3. Simulation result of 16-layer stack

Cooling Method	Max. power (< 85°C)	
	Memory layer	Logic layer
Passive heat sink	1 W/cm^2	2 W/cm^2
Low-end active heat sink	2 W/cm^2	14.4 W/cm^2
Commodity-server active heat sink	10 W/cm^2	20.0 W/cm^2
High-end-server active heat sink	25 W/cm^2	24.5 W/cm^2

In high-rise stacks, such as 64 layers or 128 layers, the passive or low-end active heat sink has a low cooling capacity to keep the operating temperature under 85°C. Table 4 and 5 show the maximum power budget of the memory system of 64 layers and 128 layer stacks. In the case of 64 layers, the robust commodity-server active heat sink can maintain the system temperature with power dissipation of 2W/cm^2 and 20 W/cm^2 in the memory and logic layers, respectively. For 128 layers, the high-end-server active heat sink can allow power to dissipate at 1.4 W/cm^2 and 16 W/cm^2 in memory and logic layers, respectively. The high-rise 3D memory requires high-end active heatsink for the normal memory operation. Although the high-rise 3D stacking is advantageous for processing parallelism, the thermal limitation due to the lack of cooling capacity of conventional air-cooling heat sinks restricts in-memory computing integration in the 3D-stacked memory.

The power budget limitation of in-3D-memory computing devices is investigated by using thermal modeling and simulation. The in-memory computing architecture can be modeled as a logic layer concentrated model or a distributed processing load model. Since in-memory computing architecture dissipates more power in the memory layers than the conventional memory operation, the distributed processing model model becomes more appropriate for investigating thermal behaviors. The proposed modeling method is validated through comparison with a simulator 3D-ICE. We find the operating condition with different power consumptions in memory and logic layers, which sustains the operating temperature of the memory system below 85 °C.

Table 4. Simulation result of 64-layer stack

Cooling Method	Max. power (< 85°C)	
	Memory layer	Logic layer
Passive heat sink	Not working	Not working
Low-end active heat sink	Not working	Not working
Commodity-server active heat sink	2.5 W/cm^2	20.0 W/cm^2
High-end-server active heat sink	6.5 W/cm^2	8.5 W/cm^2

Table 5. Simulation result of 128-layer stack

Cooling Method	Max. power (< 85°C)	
	Memory layer	Logic layer
Passive heat sink	Not working	Not working
Low-end active heat sink	Not working	Not working
Commodity-server active heat sink	1 W/cm^2	14.9 W/cm^2
High-end-server active heat sink	1.4 W/cm^2	16.0 W/cm^2

One of the advantages of in-memory computing is the processing parallelism in the memory system. However, due to the limited cooling capacity and power budget, the number of layers should be limited to 16 layers. The limited number of stacks degrades the parallelism in the memory system. The high-rise 3D-stacked memory has a thermal limitation in integrating in-memory computing architecture due to the inherent thermal issues. However, the robust server-level active heat sink cannot accommodate sufficient power for in-memory computing architectures. The potential solution for the thermal limitation in the 3D-stacked memory is the microfluidic cooling method. Contrary to the heatsink that removes the heat only from the top surface, microfluidic cooling can remove the heat directly from each layer. The proposed modeling method can also be used to simulate microfluidic cooling integrated with 3D stacked memory. Thermal issues notwithstanding, in-3D-memory computing architecture is a promising hardware solution for the demand for big data computing. Thermal modeling and an appropriate cooling solution will set forward the integration of in-memory computing architecture in the 3D-stacked memory.

5. Thermal Simulation with HotSpot 7.0

5.1 Microfluidic simulation feature of HotSpot 7.0

5.1.1 Thermal simulation tool

Temperature-aware design is a key design consideration in modern electronics. Without a thermal simulation tool, thermal research was typically done by a post-chip study which takes the worst-case power scenario. The worst-case design arranges a package to handle the severe hotspots, causing an increase in package and cooling costs. On the other hand, insufficient cooling capacity causes chip failure. Thermal simulation tools can help chip designers to predict the resulting temperature by the chip designs. HotSpot is a fast and accurate thermal simulator that uses floorplan and power traces as inputs and generates the temperature of the chip as output based on thermal circuit modeling [51]–[53].

3D-ICE is a thermal simulation including microchannel cooling for 3D-ICs[29], [41]. The limitation of the 3D-ICE is the fixed geometry support in microchannel and microfluidics. The fluidic flow in the simulator is limited to one fixed direction from north to south. This fixed unidirectional flow brings out the thermal gradient problem. Consequently, recent studies [40]–[42] to solve the thermal gradient problem are not applicable to the simulator.

Multiphysics simulation tools such as ANSYS or COMSOL can be used as thermal simulation tools. However, these tools require a separate thermal modeling process. Thermal modeling is a separate time-consuming design process based on the layout of the chip and the power numbers from the simulation or experiment. As well as the modeling process, the simulation process requires a high degree of resources and time for calculation. These tools are suitable for investigating the detailed thermal behavior but not suitable for fast simulation in the pre-RTL stage.

5.1.2 Thermal modeling of HotSpot 7.0

While microfluidic cooling is very promising, it comes with a complex design space. Perhaps the most impactful decision in microfluidic cooling is choosing where exactly to place the channels. Microchannels can be placed in a heat sink or heat spreader, but we believe that for

3D ICs, they will be most effective when placed between layers. It has also been shown that the geometry of the microchannels can have significant effects on both the effectiveness of the cooling and the efficiency of the pumping system [54]. Furthermore, there are closely related cooling techniques to consider, like pin fin cooling and pumping coolant through the micro-gaps present between layers, such as the inter-chip cooling proposed in DARPA's ICECool project [26].

With such a large design space to explore, the research community needs a tool for modeling and simulating microfluidic cooling. We need a tool that is flexible enough to support many different microfluidic cooling techniques, including those that are already discussed and novel techniques yet to be discovered. It should also be employed early in the design process (pre-RTL) to enable researchers to quickly investigate new designs and hone in on the promising ones.

To create a tool that allows for thorough early-stage design space exploration, we start with HotSpot 6.0 [51]–[53], an existing thermal simulator capable of simulating 3D ICs, and extend its thermal model to support microfluidic cooling. HotSpot requires very little information (such as a 3D IC's dimensions and power dissipation values) to perform a simulation, making it suitable for the pre-RTL design space exploration that we want to achieve.

HotSpot 6.0 models a 3D IC by discretizing it into an array of 3D cells, then modeling each cell as one node in a thermal circuit. This circuit can then be analyzed using well-established circuit analysis techniques to find the temperatures at each node, which corresponds to the temperatures within each cell throughout the 3D IC. One of the limitations of HotSpot 6.0, however, is that it can only model heat transfers due to conduction. We extend the thermal model to include the modeling of heat transfers due to convection to provide the microfluidic cooling feature. In the following sections, we discuss how conductive heat transfer is modeled in HotSpot 6.0 and the extensions that we've made to allow modeling convective heat transfer.

5.1.3 Modeling conductive heat transfer

Conductive heat transfer may occur between any two adjacent solid cells or any two adjacent fluid cells in the discretized 3D IC. To model this, we connect each node in the thermal circuit to all adjacent nodes of the same type (solid or fluid) through thermal resistances.

We model heat flow from the center of one cell to the center of the adjacent cell. Since adjacent cells may be different materials, we can think of the thermal resistance connecting the nodes as being two thermal resistances in series: one modeling the conductive heat transfer from the center of the first cell to the edge of the first cell, and one modeling the conductive heat transfer from the edge of the second cell to the center of the second cell. The value of the total thermal resistance is given by the equation, in which t is the distance between the centers of the cells, k_1 is the thermal conductivity of one cell's material, k_2 is the thermal conductivity of the other cell's material, and A is the cross-sectional area of the interface between the cells.

$$R_{th} = \frac{t}{2k_1A+t} + \frac{t}{k_2A}$$

For transient simulations, we also need to take into account each cell's thermal capacitance. We model this by connecting each node to the ground through a capacitance whose value is given by the equation, where C is the cell's volumetric heat capacity, t is the cell's thickness, and A is the cell's cross-sectional area.

$$C_{th} = CtA$$

5.1.4 Modeling of convective heat transfer

We adopt a model very similar to the 4RM-based Compact Transient Thermal Model (CTTM) used in 3D-ICE [29], [41] for convective heat transfer. This thermal model uses the fluid flow rate between every pair of fluid cells. However, finding these fluid flow rates is nontrivial since, in HotSpot, we do not make any assumptions about the microchannel geometry. To find out the flow rates, we model each microfluidic cooling layer using a pressure circuit by taking advantage of the similarities between Ohm's Law and the Hagen-Poiseuille Law [43]. In

the pressure circuit, voltage is analogous to pressure, current flow is analogous to fluid flow, and electrical resistance is analogous to hydraulic resistance. Next, we describe how we use the pressure circuit to find these flow rates and how we then use the flow rates to derive component values in the thermal circuit.

In order to derive the component values for the thermal circuit, we need to know the fluid flow rates between any given pair of fluid cells. To find these, we model each microfluidic cooling layer as a pressure circuit. The pump is modeled as a voltage source, and the hydraulic resistances of the channels are modeled as electrical resistances. The resistance between a pair of cells is given by (3), where h, w, and L are the height, width, and length of the cells, respectively, and η is the dynamic viscosity of the fluid.

$$R_{hyd\_channel} = \frac{12\eta L}{1\text{-}0.63(h/w)} \frac{1}{h^3 w}$$

After calculating the hydraulic resistances between every pair of fluid cells, we perform node-voltage analysis to find the pressure in each fluid cell. Once we have found the pressure and resistance between every pair of fluid cells, we find the flow rates using a simple application of the Hagen-Poiseuille Law. Figure 36 (b) shows a schematic of the pressure circuit for the microchannel given in Figure 36.

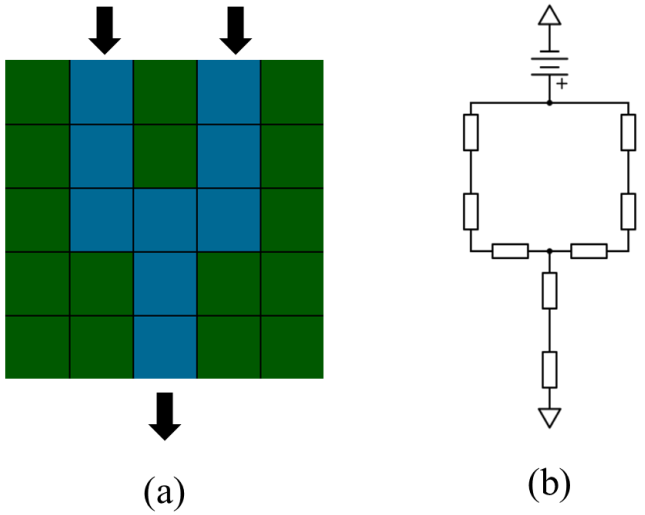


Figure 36. (a) Example Microfluidic Network and (b) Corresponding Pressure Circuit

Convective heat transfer from a solid cell to an adjacent fluid cell is modeled using a thermal resistance between the two cells. We again assume heat flow between the centers of the cells so that the thermal resistance can be thought of as two resistances in series: one representing conductive heat transfer from the center of the solid cell to the edge of the solid cell and one representing convective heat transfer from the edge of the fluid cell to the center of the fluid cell. The total thermal resistance is given by the equation, where t is the distance between the centers of the cells, k is the thermal conductivity of the solid cell, h is the heat transfer coefficient, and A is the area of the interface between the cells.

$$R_{th} = \frac{t}{2k_1A+t} + \frac{t}{k_2A}$$

To model the heat flow between fluid cells due to the movement of the fluid, we connect adjacent fluid cells via current sources. A current source connecting one fluid cell to another fluid cell models heat transfer from the first fluid cell to the second. The values of the current sources are given by the equation, where C is the volumetric heat capacity of the fluid, V is the volumetric flow rate of the fluid (which we found using the pressure circuit), and T_i is the temperature at the interface between the fluid cells.

$$J = CVT_i$$

In practice, we approximate T_i as the average of T_1 and T_2, where T_1 is the temperature of the cell from which the coolant is flowing, and T_2 is the temperature of the cell to which the coolant is flowing: $T_i = (T_1 + T_2)/2$. Figure 37 (b) shows a schematic of the thermal circuit for the microchannel shown in Figure 37 (a). For simplicity, we have omitted the thermal capacitances, vertical thermal resistances to different layers, and the thermal resistances between fluid cells (adjacent fluid cells are connected via both thermal resistances and current sources).

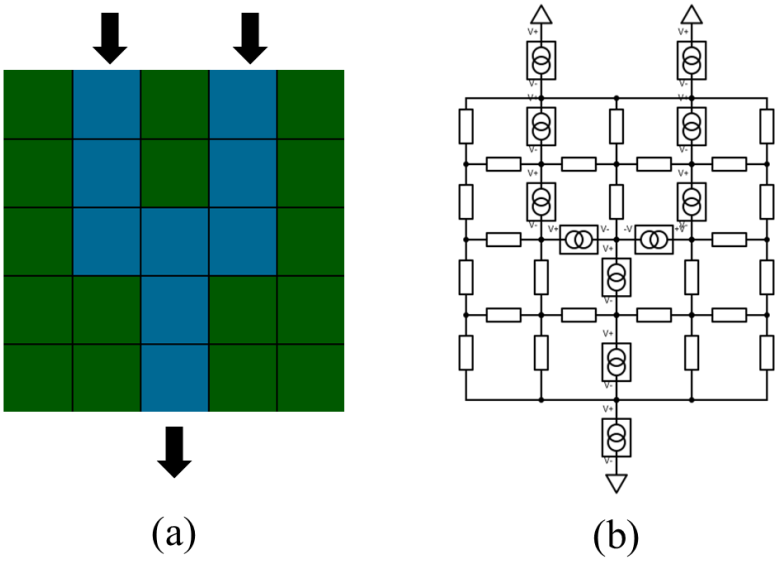


Figure 37. (a) Example Microfluidic Network and (b) Corresponding Thermal Circuit

5.1.5 Implementation

The implementation of the thermal model into the HotSpot simulator is a credential to Robert E. West. He led the development of HotSpot 7.0 and released it to GitHub [55]. HotSpot 6.0 uses a 4th-order Runge Kutta (RK4) algorithm with adaptive step sizing as its transient differential equations solver[51]. However, we've found that our changes to the thermal model have increased the time required for the RK4 solver to converge. In some larger simulations, this convergence time is prohibitive. As a result, we choose to implement a differential equations solver using the backward Euler method. We choose this method because it is the algorithm used in 3D-ICE [29], [41], which we use as a basis for our changes to HotSpot 7.0[56].

The choice of differential equations solver comes with a tradeoff. The RK4 method is more accurate since it has a local truncation error of $O(h^5)$, while the backward Euler method has a local truncation error of $O(h^2)$. However, the backward Euler method is A-stable, while the RK4 method is not. We've found that the RK4 method provides a better balance of accuracy and simulation time for simulations not including microfluidic cooling, while the backward Euler method provides a better balance for simulations that include microfluidic cooling. The rest of this section describes the backward Euler solver and how we've implemented it in HotSpot in

more detail. To use the backward Euler solver, we first represent the thermal circuit using the following equation.

$$GT + C\frac{dT}{dt} = P$$

For a circuit with n nodes, G is an n × n connectivity matrix that describes how all cells are connected via thermal resistances and current sources, C is an n×n diagonal matrix that contains the thermal capacitance of each cell, T is an n × 1 vector of the temperatures of each cell, and P is an n × 1 vector of the power dissipation of each cell. Before the simulation, G, C, and P are known, and we wish to find the temperatures, T.

To solve the thermal circuit with the backward Euler method, we use the equation below, where h is the step size, and f is a function representing the derivative dT /dt.

$$T_{n+1} = T_n + hf(t_{n+1}, T_{n+1})$$

After solving the original equation for dT /dt, inserting the result into the above equation and rearranging it to isolate T_{n+1}, we get below.

$$\left(\frac{1}{h}C + G\right)T_{n+1} = \frac{1}{h}CT_{n+1} + P$$

Finally, we write the equation as the matrix equation $AT_{n+1} = B$ and solve it using the SuperLU matrix solving library [57], [58].

Solving the thermal circuit in steady-state simulations is much simpler. In steady-state simulations, we assume that all thermal capacitances have been fully charged and thus are effectively open circuits. This allows us to zero out the capacitance matrix C and solves the resulting equation with SuperLU.

5.2 Microfluidic behavior in the 3D-IC

5.2.1 Thermal Modeling of processing-in-3D-memory

We used thermal circuit-based simulations such as HotSpot to investigate the temperature of 3D-IC. Although the thermal circuit-based simulations are fast and accurate, a deeper investigation into the microfluidic behavior is required for verification. We used COMSOL Multiphysics to verify the thermal behavior of microfluidic in the thermal chamber. The COMSOL simulation has identical conditions to the thermal circuit configurations. Figure 38 shows the 3D design result of the thermal model of the 3D memory system using SolidWorks. The memory stack is composed of memory dies, logic dies, a fluidic chamber, and a silicon interposer. The memory vault area is 64 mm^2, and the power consumption in the logic die is 64 W and 128W, representing a power density of 100 W/cm^2 and 200 W/cm^2, respectively. Figure 39 shows the COMSOL simulation results of 3D-stacked memory with microfluidic cooling. When the memory stack consumes 64 W of power with a power density of 100 W/cm^2, the memory stack reaches 85.7°C with 1 ml/min of flow rate. With a higher power density of 200W/cm^2, the maximum temperature was 147°C and 67°C with a volumetric flow rate of coolant of 1ml/min and 10 ml/min, respectively.

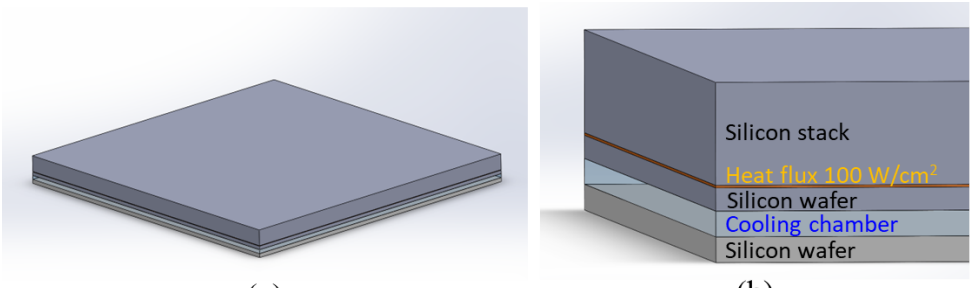


Figure 38. 3D thermal modeling result of the PIM integrated 3D-stacked memory

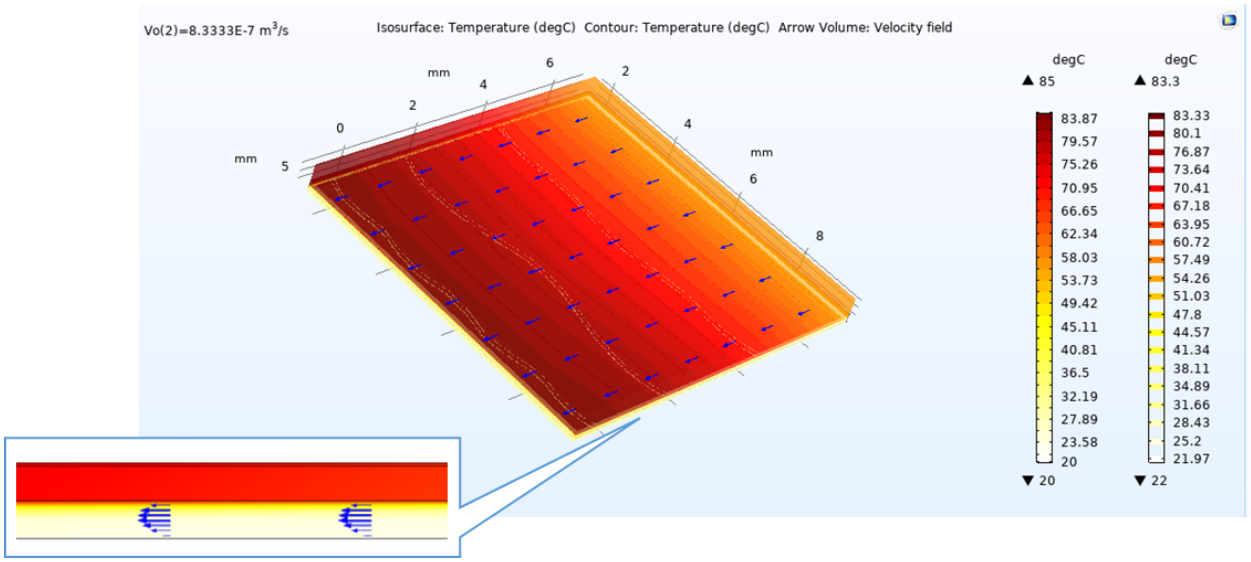


Figure 39. 3D COMSOL simulation result of the PIM integrated 3D-stacked memory

5.2.2 Thermal behavior comparison on Microchannel and Microchamber

The microchamber cooling method has many structural merits over the microchannel method by exploiting the existing cavity. In addition, the double layer of microfluidic cooling shows significantly superior cooling behavior than the conventional methods. Microchamber cooling is advantageous to multilayer cooling because of the simple structure and the vertical scalability. We compared the cooling capacity of the microchamber and microchannel cooling. Figure 40 shows the temperature of the logic layer of 3D memory with different microfluidic cooling structures. The first simulation condition has 100μm of channel and wall widths. The second and third have larger channel widths of 400μm and 900μm. In the last case, there is one chamber with a width of 7900μm. The other simulation variables, such as power dissipation, wall thickness, and flow rates, remain the same. Although the width of channels is different, the thermal behaviors show identical results. Consequently, the proposed microchamber cooling method is more suitable for high-performance 3D-stack IC than the conventional microfluidic cooling methods.

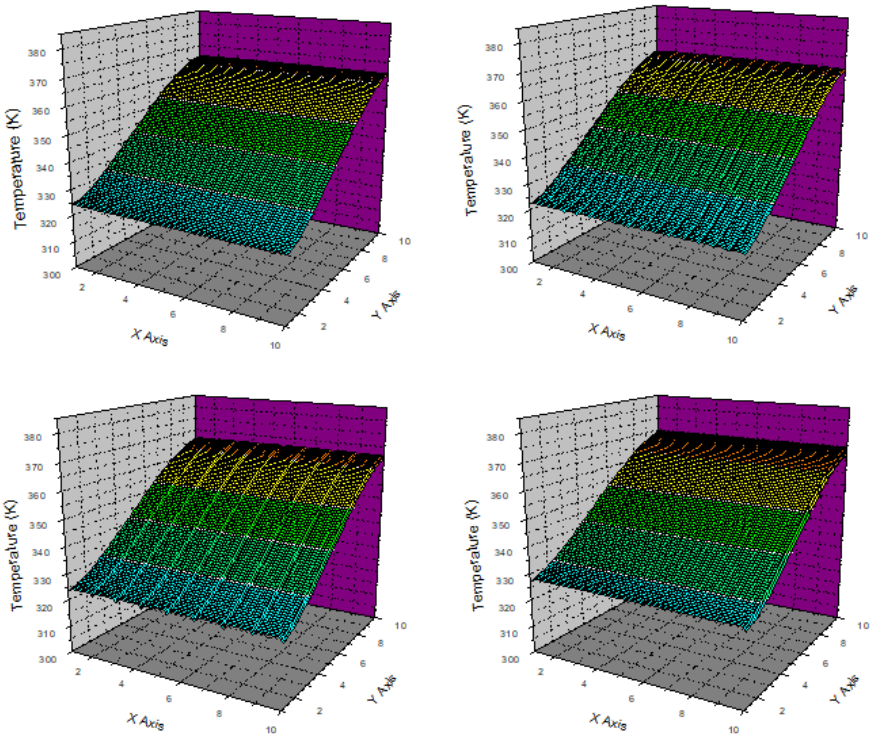


Figure 40. 3D-ICE simulation results of microchannels with widths of (a) 100µm, (b) 400µm, (c) 900µm, and (d) microchamber with a width of 7900µm.

5.2.3 Thermal boundary layer

We investigated the thermal behavior of microfluidics in the channel and chamber by using 2D and 3D COMSOL simulations. Figure 41 shows a cross section view of the COMSOL 3D simulation result. As we see in the close-up view in figure 42, the results show that the temperature of the coolant in the channel has a temperature gradient from the vertical direction of the channel wall. This thermal gradient region is called a thermal boundary layer.

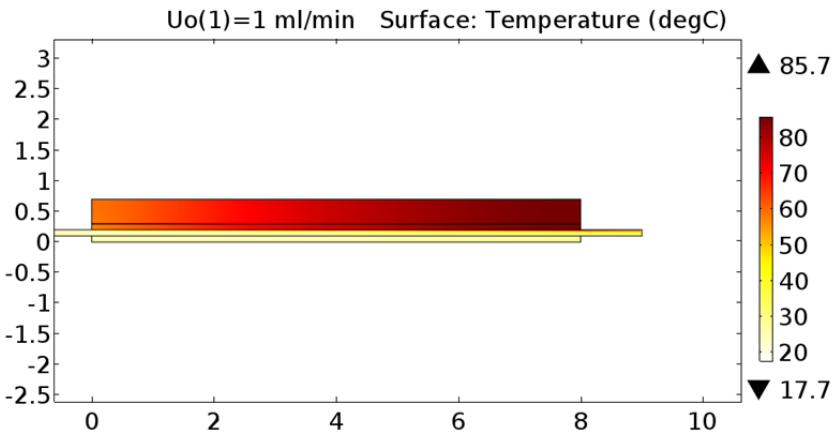


Figure 41. Side view of the COMSOL simulation result of the PIM integrated 3D-stacked memory

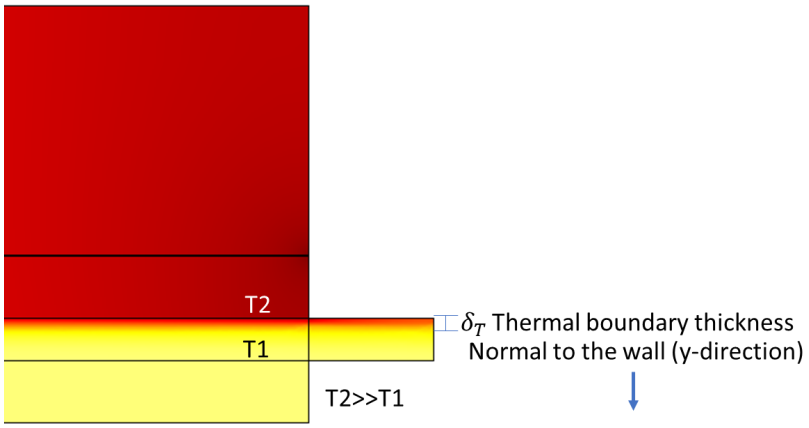


Figure 42. Detailed image of the thermal boundary layer in the side view of the COMSOL simulation result of the PIM integrated 3D-stacked memory

The thermal boundary layer created at the slick heat exchanging surface with laminar flow condition affects the cooling capacity of the microfluidic cooling method [59], [60]. As the coolant flows in the channel, the fluid at the interface between the wall and the fluid satisfies a no-slip boundary condition hence the velocity drops to zero[61]. As the distance from the interface increases, the velocity of coolant increases, and the temperature decreases. Figure 43 shows a schematic image of the thermal boundary layer. The thickness of the thermal boundary

layer can be determined where the temperature of the coolant reaches 99% of the free-stream temperature T0. The thermal boundary layer effect decreases the cooling capacity due to the limited heat exchange between the coolant and the silicon. For ideal turbulent flow, there is no thermal boundary layer, and all portions of the coolant contribute to the cooling behavior. On the other hand, when the thermal boundary layer exists, it means that only the coolant within the boundary layer contributes to thermal exchange. When the flow rate is slow, more portion of the coolant is participating in heat transfer, and the thermal boundary is thick. When the flow rate is fast, the coolant moves too fast before the heat is transferred, and the thermal boundary is thin. With a higher flow rate in the channel, the thickness of the thermal boundary layer decreases while significantly decreasing the heat transfer from the solid wall to the fluid. Moreover, while the 3D-ICE or HotSpot has not considered the thermal boundary layer in the calculation, the thermal boundary layer causes a simulation error between the tools.

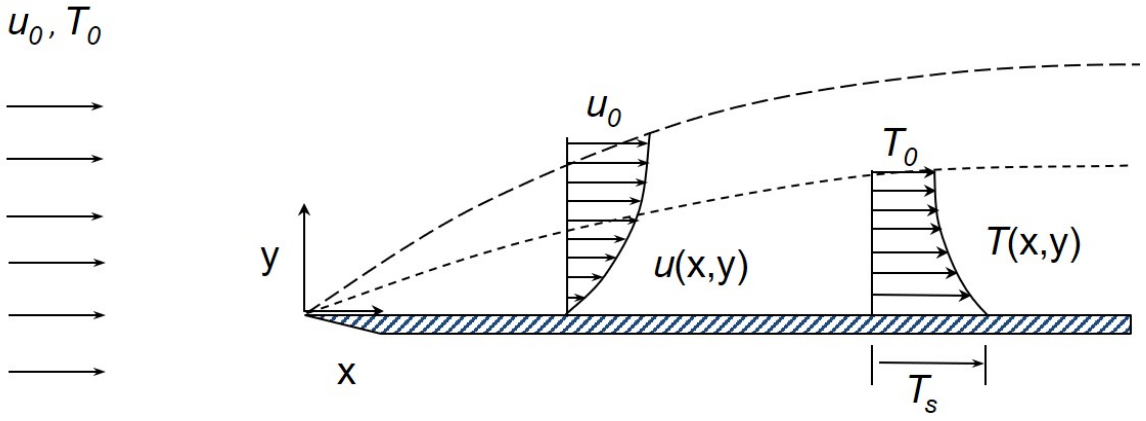


Figure 43. Schematic drawing depicting fluid flow over a heated flat plate [42].

5.2.4 Microchamber with micropillars

The thermal boundary layer is pervasive when the heating object is a flat plate. The previous 3D thermal modeling of the PIM integrated 3D-stacked memory has a simplified structure and flat surfaces. In fact, the microfluidic chamber is filled with microbumps or

micropillars. These electrical interconnections contribute to the turbulent flow in the channel region and mitigate the thermal boundary layer effect. The figure 44 and 45 show the designed fluid chamber area with micropillars. The pillar has a 100 µm diameter and a 100 µm height.

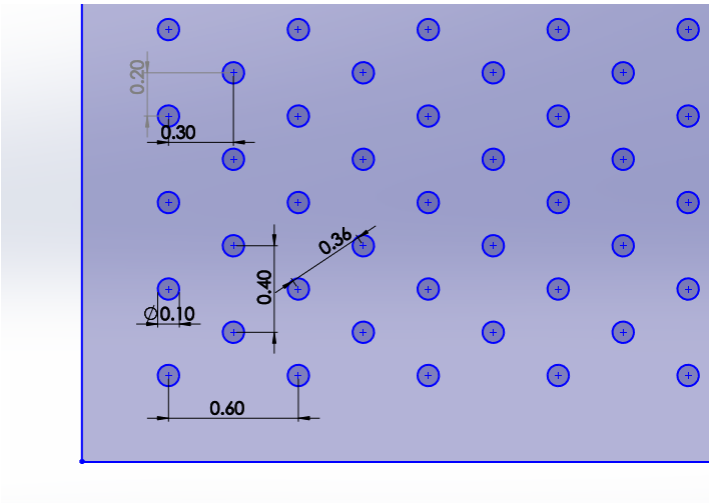
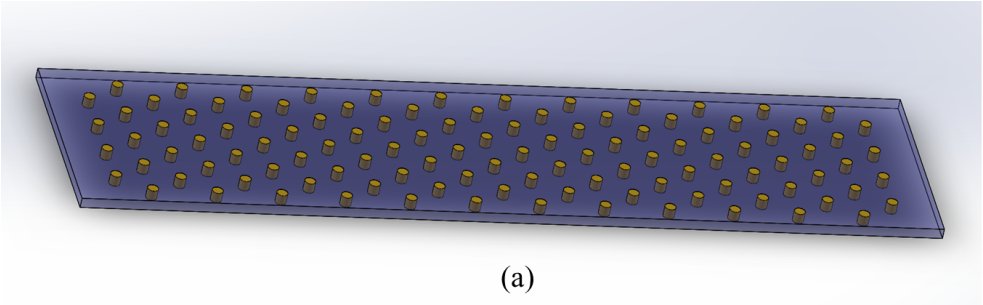
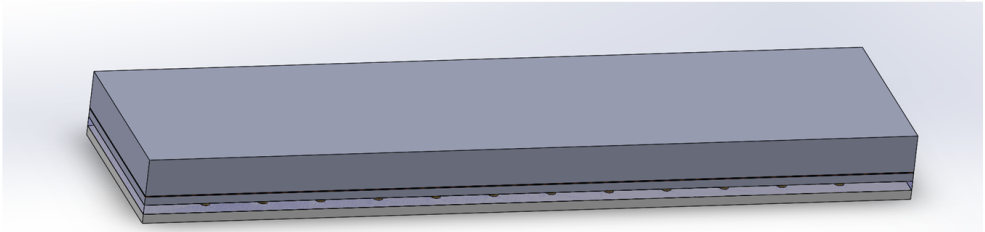


Figure 44. The design and spacing of micropillars in the fluid chamber



(a)



(b)

Figure 45. 3D thermal modeling of (a) fluid chamber with pillars and (b) PIM integrated 3D-stacked memory

We applied the interconnection pillars between the logic layer and the interposer for a more detailed simulation. Figures 46 and 47 show the COMSOL simulation results without and with pillars in the microchamber. The 3D COMSOL simulation result shows that the maximum temperature of the 3D-stacked memory has decreased from 83.87°C to 66.3°C. The temperature difference between the multiphysics simulation and thermal circuit simulation in figure 48 is significantly reduced by adding the pillars. The electrical interconnections between the dies can behave as protruding structures, increasing the cooling area. Cu-pillars between dies can increase the cooling capacity because copper has a higher thermal conductivity of 401 W/(m·K) than the 149 W/(m·K) of silicon. The thermal boundary layer effect is reduced, resulting in the microfluidic cooling through the chamber becoming more effective.

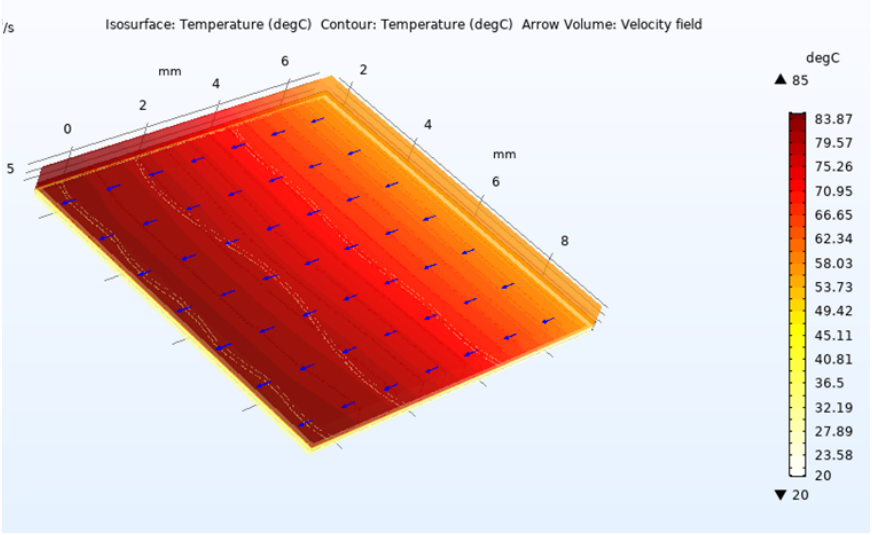


Figure 46. 3D COMSOL simulation result of the PIM integrated 3D-stacked memory

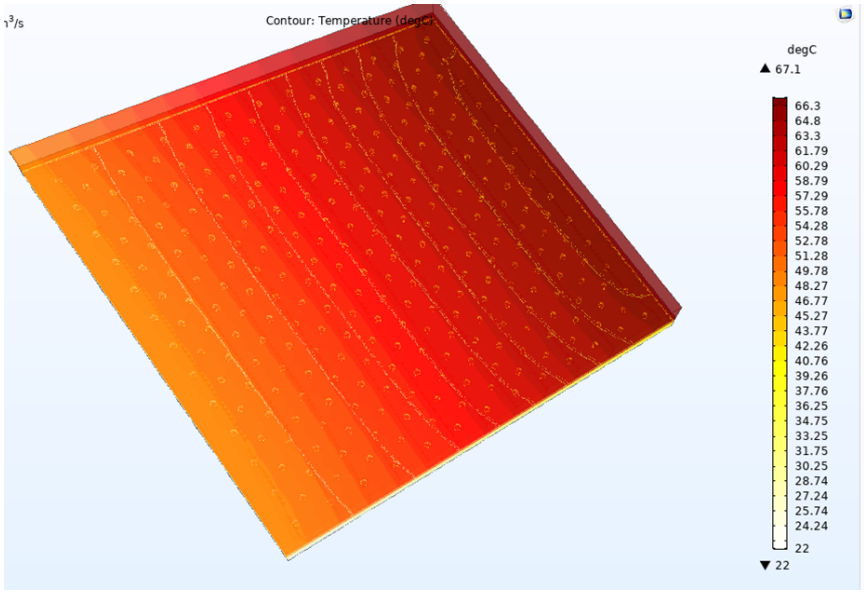


Figure 47. 3D COMSOL simulation result of the PIM integrated 3D-stacked memory with micro pillars in fluid chamber

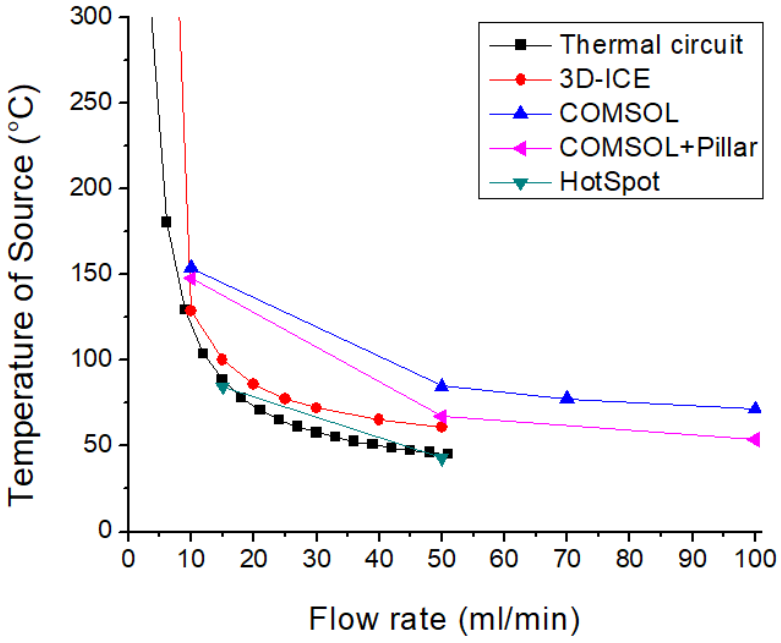


Figure 48. Simulation results comparison.

5.3 Simulation Result Validation of HotSpot

5.3.1 Microchannel thermal modeling

In previous chapters, we investigated the microfluidic cooling behavior of the proposed microchamber structure. The fluidic chamber has a thermal boundary layer which causes the mismatch between the simulation methods. The resulting gap between simulators has been reduced by adding a detailed 3D modeling structure in COMSOL simulation and minimizing the impact from the thermal boundary layer. Granted that the proposed method is advantageous in fabrication and cooling aspects, the novel structure is not suitable for the validation of the new simulation tools. Therefore, we designed a conventional microchannel device to validate the microfluidic cooling feature of HotSpot 7.0. Figures 49 and 50 show the 3D designs of the microchannel cooled processing-in-3D-memory system. The channels and channel walls have the same height of 100 µm and width of 200 µm, except for the edges with 300 µm width. Hence there are 19 microchannels integrated into the backside of the silicon substrate.

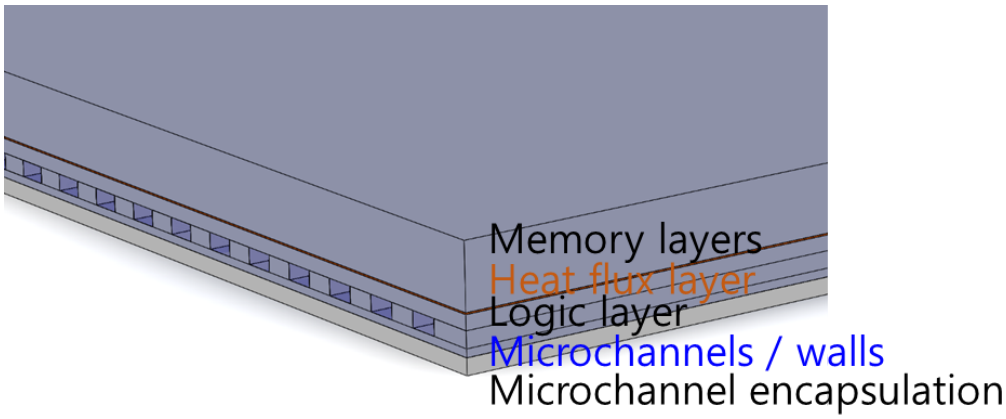


Figure 49. 3D thermal modeling of the PIM integrated 3D-stacked memory with microchannels

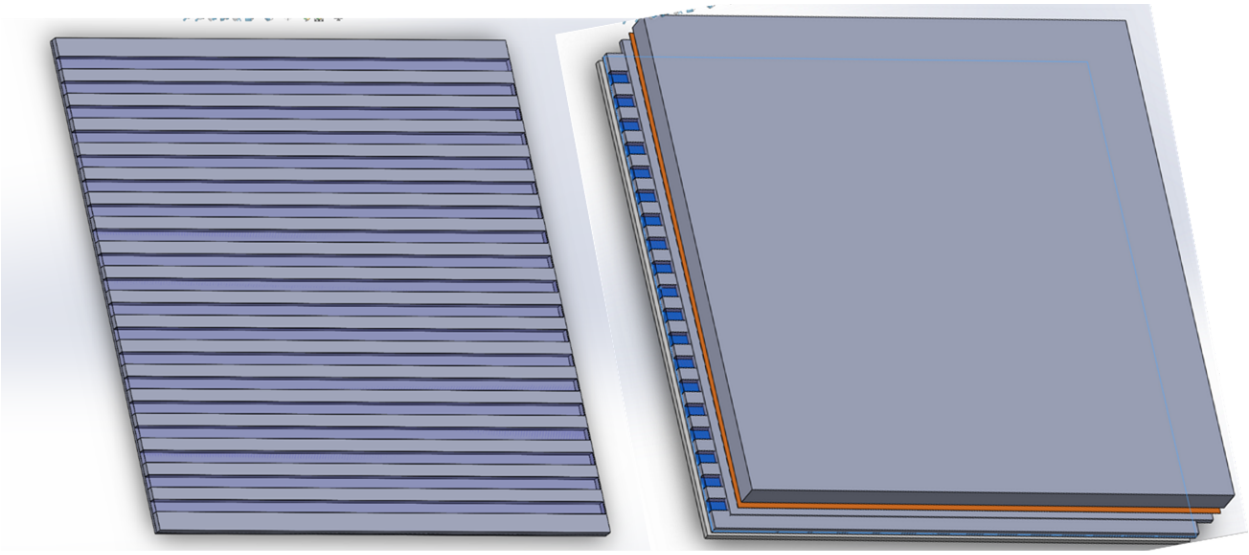


Figure 50. 3D thermal modeling of (a) microchannels and (b) PIM integrated 3D-stacked memory

The power dissipation of the device is 100 W/cm^2 in the logic layer. And we investigated the volumetric flow speed of 10ml/min and 1ml/min. In the case of the volumetric flow of 1ml/min, the total volumetric flow in the cooling systems is 19ml/min, and this requires around power consumption of 4W from the hydraulic pump.

5.3.2 HotSpot simulation result validation

Figures 51 and 52 show the simulated temperature distribution with COMSOL and HotSpot simulators. The thermal modeling, system configuration, material properties, power dissipation, and cooling condition are the same. The height and width of microchannels are 100 µm and 200 µm, respectively. The power dissipation in the logic layer is a power density of 100 W/cm^{2,} and the volumetric flow in the channel is 10ml/min. The maximum temperature logic layer from COMSOL and HotSpot is 40.8°C and 49.3°C, respectively. In the same way, in the memory layer, 40.8°C and 49.3°C, respectively. The temperature distribution shows identical

thermal behaviors, such as more effective cooling in the center rows than the top and bottom rows, thermal gradient from inlet to outlet, and temperature step across the channel regions.

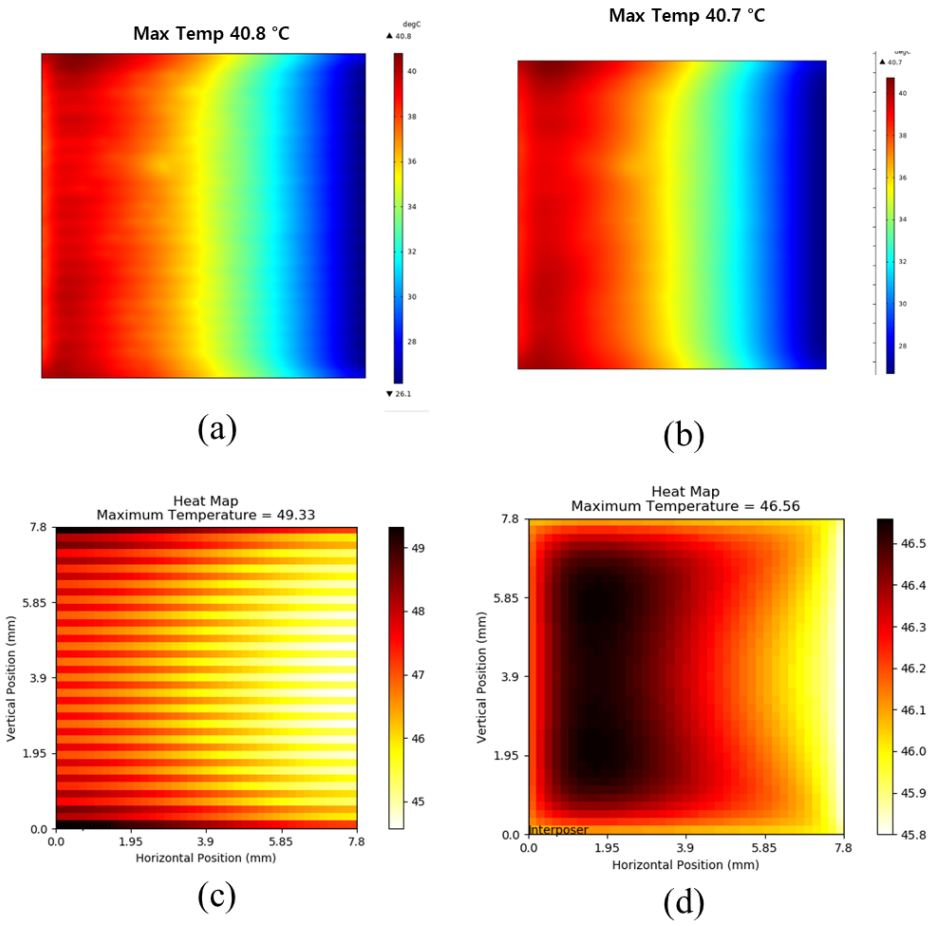


Figure 51. Simulated temperature with COMSOL of (a) logic and (b) memory layer, and with HotSpot of (c) logic and (d) memory layer with the volumetric flow of 10ml/min.

With 1ml/min volumetric flow, the maximum temperature logic layer from COMSOL and HotSpot is 96.6.C and 76.8°C, respectively. With a lower flow rate, the quantitative result shows bigger differences, but is still identical to qualitative thermal behaviors. The color pattern in the center area shows a round shape in both simulation results, while it was more straight in the higher volumetric flow condition.

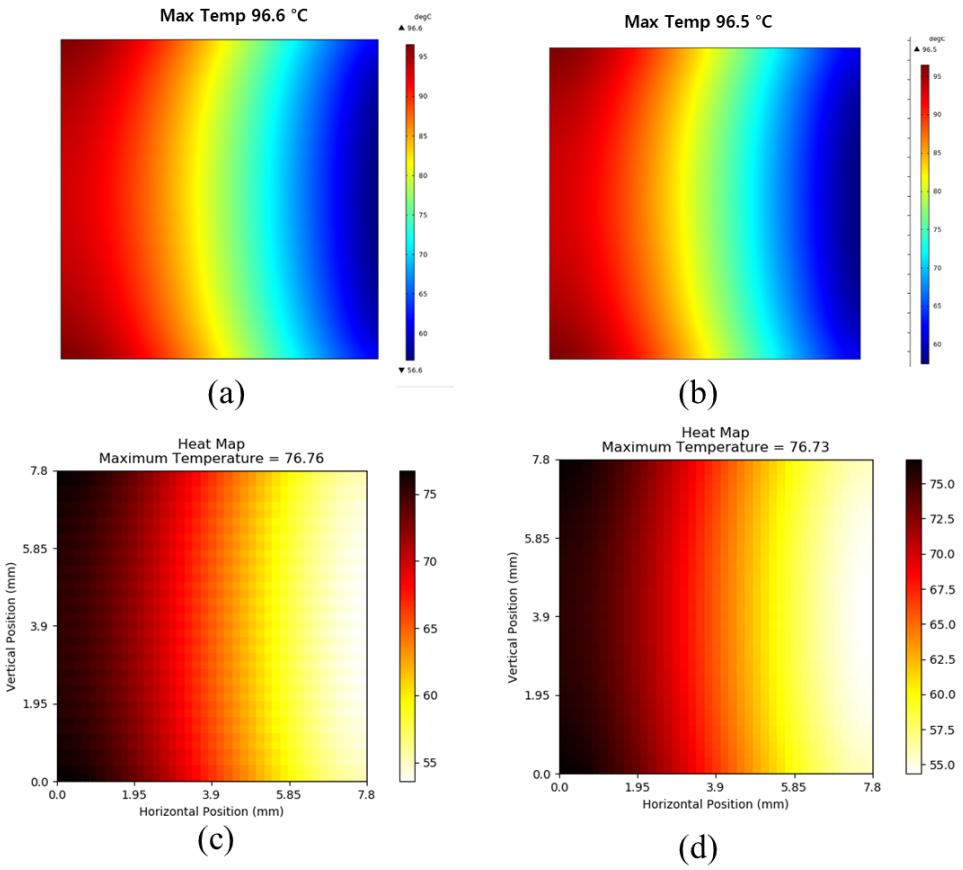


Figure 52. Simulated temperature with COMSOL of (a) logic and (b) memory layer, and with HotSpot of (c) logic and (d) memory layer with the volumetric flow of 1ml/min.

The graphs in figure 53 show the maximum and average temperature of the simulated temperature distribution. We have fewer data points in the COMSOL simulation result because the simulation process takes longer time than HotSpot. For HotSpot simulation, it takes only seconds but takes several hours, in some cases a few days COMSOL to produce simulation results. The temperature trend provides that the HotSpot simulator can be an alternative thermal simulation method to the multiphysics simulator including the 3D-IC and microfluidic cooling features.

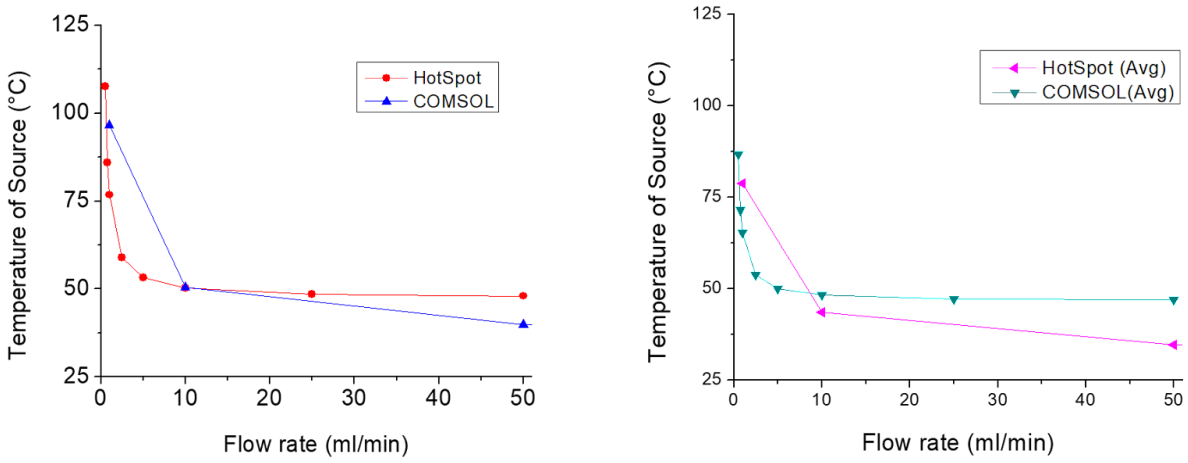


Figure 53. Simulated (a) average and (b) maximum temperature with HotSpot and COMSOL

5.4 Design flexibility of HotSpot

5.4.1 Flexibility of the fluid flow

The fixed unidirectional flow of the microfluidic cooling causes a thermal gradient problem in the chiplets. As shown in figure 54 (a), the temperature on the left side is as low as 55°C, while the temperature on the right side is over 90°C. This thermal gradient is produced when the chilled coolant is injected through the inlets, removing heat from the surface, and the heated coolant sinks out through the outlets. In the worst case scenario, if the hotspots are located on the right side of the floor plan, the microfluidic cooling not only decreases the cooling capacity, but also exacerbates the thermal problem.

The microfluidic feature in HotSpot 7.0 has a design flexibility that allows users to control the fluidic directions in the channels. After placing the inlet and outlets in a fluidic floor plan, the HotSpot calculates the pressure and flow rate in the channels. Figure 54 (b) shows the temperature distribution of microfluidic cooling with the alternating flow direction. The direction of the fluidic flow alternates every even and odd row. The thermal gradient left to right disappeared, and the uniformity of the temperature distribution increased significantly.

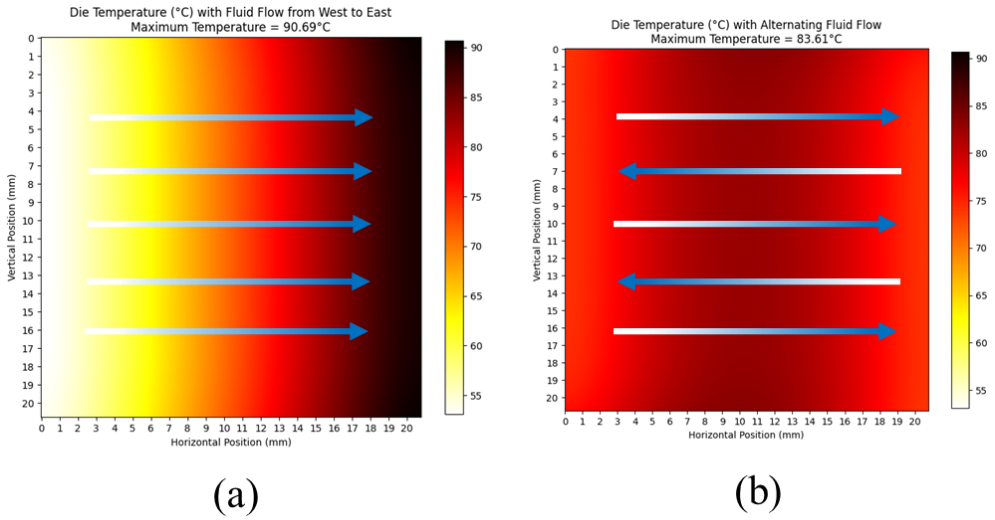


Figure 54. Temperature distribution diagram of the microfluidic cooling system of (a) fixed 1-directional flow and (b) alternate directional flow

5.4.2 Flexibility of the fluid geometry

The HotSpot 7.0 also has design flexibility in microchannel geometry design. While conventional simulation tools have strict fluid design criteria, the HotSpot provides design freedom to the chip and thermal designers. Figure 55 shows two types of design results of microchannel geometry. There is a designed diversity in the aspect of the number and location of inlets and outlets and effective channel width and geometries. HotSpot calculates the pressure and flow rates in the microfluidic channels and calculates the temperature of the chiplet with a microfluidic cooling layer.

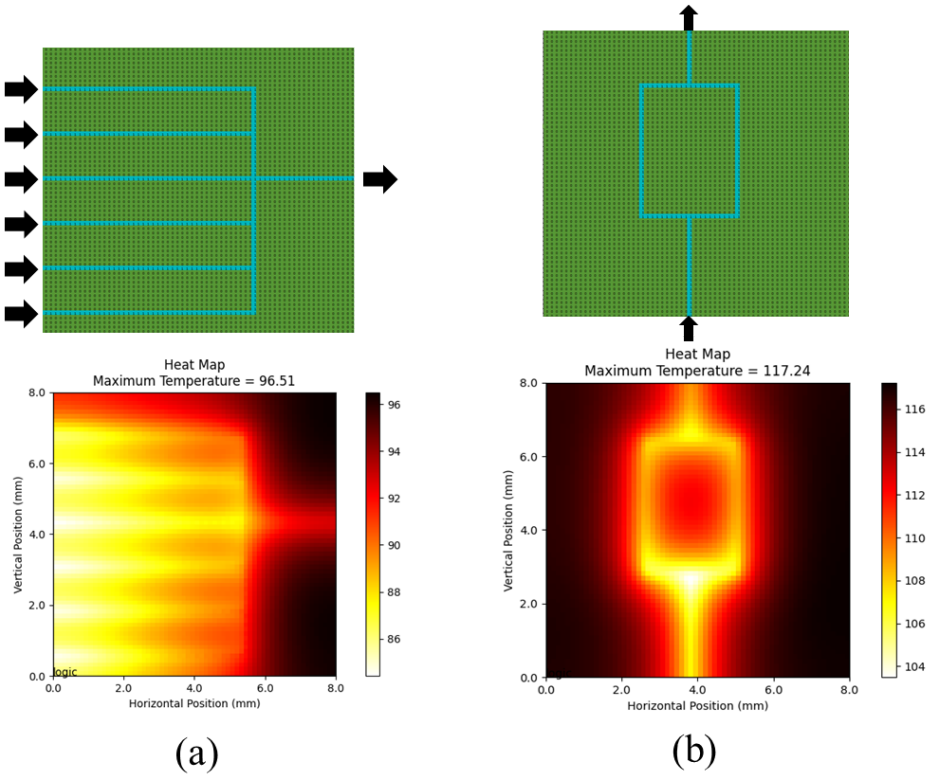


Figure 55. Two types of design results and thermal simulation results.

5.4.3 Layer scalability

HotSpot 7.0 included a layer design flexibility in the layer configuration of chiplets and the cooling layer. The design can place the microfluidic cooling layer on the bottom and top surface of the chiplet for thermal design purposes. In addition, it is possible to add a heatsink or additional microfluidic cooling layers to the system. This feature provides design flexibility in thermal aspects.

6. Thermal investigation of 3D-IC

6.1 Thermal simulation of high-performance multicore processor

6.1.1 Chiplet modeling

We have modeled a 2.5D and 3D integration of a hypothetical high-performance processor, which produces extreme heat. The hypothetical processor architecture is modeled based on the Intel i7-3960X [62], [63]. Figure 56 shows the layout image and the floorplan of a still square with a side length of 20.8 millimeters generated with ArchFP. Each processor has integrated side by side for the 2.5D chiplet integration, and for the 3D chiplet, one processor is stacked directly on the other processor. We also created a hypothetical power map based on the i7-3960X's published TDP of 130 watts. The power-thermal characteristics are investigated under the heat sinks with convection thermal resistance of 4.0, 2.0, 0.5, and 0.2 K/W. And we use the same two configurations for the microfluidic cooling layer with 31 microchannels with a height of 100 um and a width of 325 um.

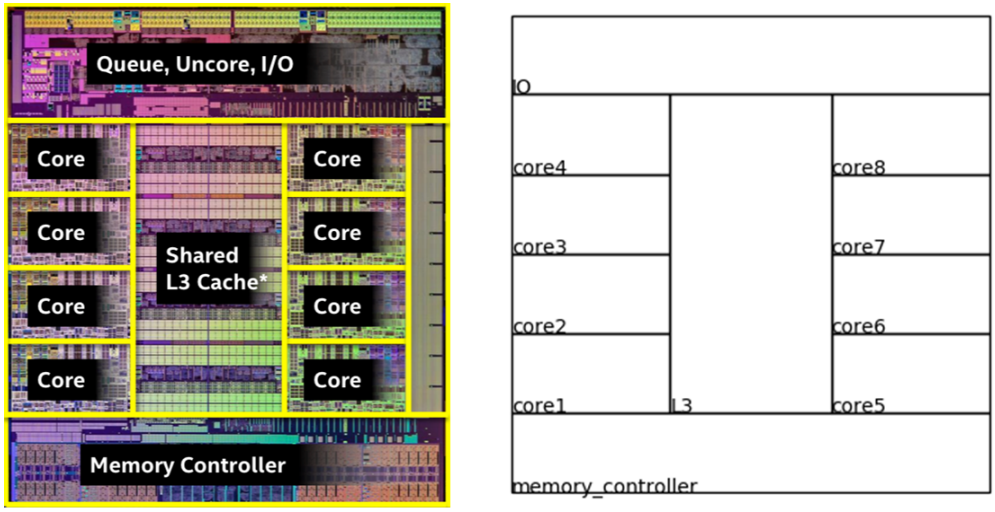


Figure 56. (a) Image[43] and (b) floor plan diagram of i7 processor

6.1.2 Thermal simulation with HotSpot

The thermal characteristic of high-performance processors is investigated using the HotSpot simulator. The hypothetical power map assumed that six of eight cores are active and consuming 20.5W per core and two cores are disabled for thermal management. The other blocks, IO, L3-cache, and memory controller, are consuming 3W, 2W, and 2W, respectively. When the power consumption of the processor is 130W, the maximum temperature of the 2.5D chiplet with passive, high-end, and server heat sinks are 600.54°C, 143.79°C, and 102.34°C, respectively. In addition, we simulated the maximum temperature of the 2.5D chiplet with increasing power consumption. We normalized the power consumption with a maximum TDP of 130 watts. As shown in figure 57, the thermal distribution over the chiplet has the same trends and hotspots with different heatsinks. With increasing power consumption, the heatsink with lower cooling capacity, in other words, with higher convection thermal resistance, shows trends that steeper increasing temperature over the increase of the power consumption.

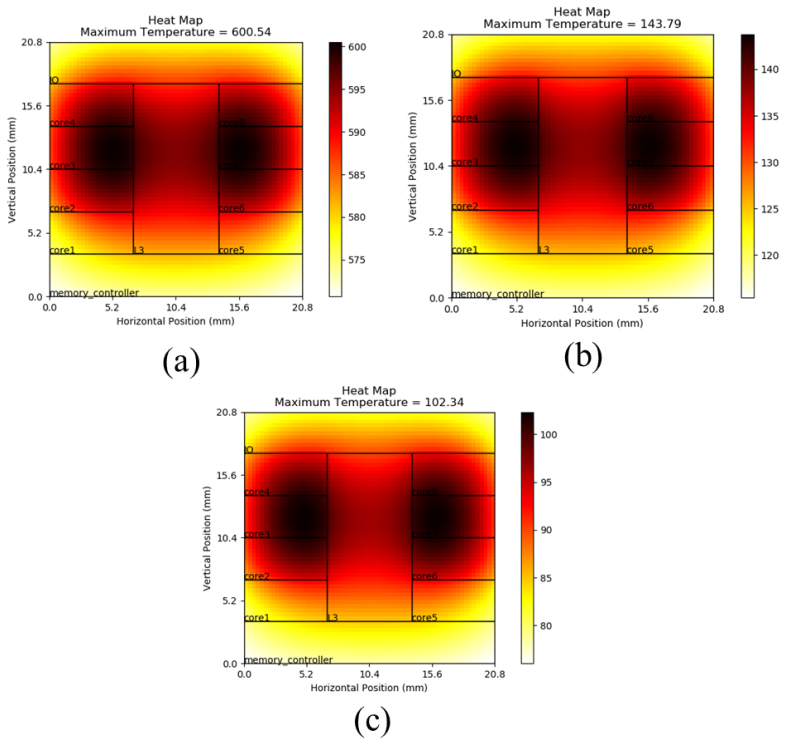


Figure 57. Temperatures throughout a 2.5D high-performance chiplet with (a) passive heat sink, (b) high-end heatsink, and (c) server heatsink

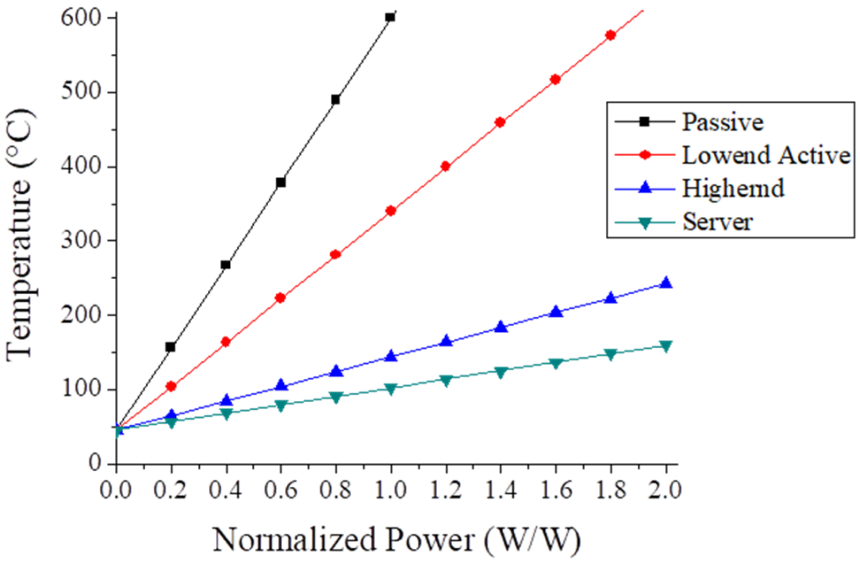


Figure 58. The temperature of 2.5D high-performance chiplet with heat sinks

Figures 59 and 60 show the thermal simulation results of a 2.5D chiplet with microfluidic cooling. The microchannels have a flow rate of 53.5 ml/min under a 200 mbar pressure supplied by a hydraulic pump. The coolant flows from inlets on the left side to the outlets on the right side. Hence, a thermal gradient appears along with flow direction. The hotspot cores on the left plane have a lower temperature of around 70°C, and hotspots on the right plane have a temperature of 106.29°C at maximum. Hence, the maximum temperature of the microfluidic cooling is higher than the server heatsink, while the average temperature decreases from 102.34°C to 97.78°C.

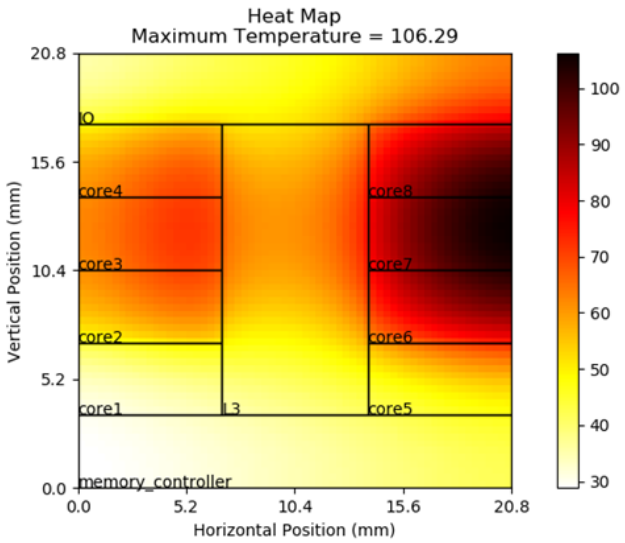


Figure 59. Temperature distribution of high-performance processor with microfluidic cooling

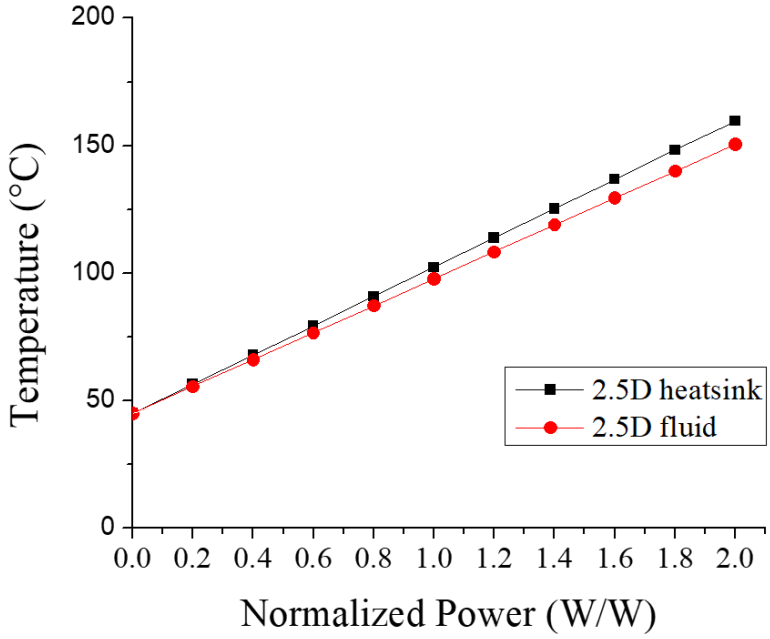


Figure 60. The temperature of 2.5D high-performance chiplet with a server heatsink and microfluidic cooling

6.1.3 3D-stack of high-performance processors

The 3D-stack of high-performance processors has advantages in parallel computing, higher transistor density, and shortened data movement length. However, it also comes with a higher power density and thermal problems. Figure 61 shows the thermal behavior of the 3D chiplet with the server heatsink and microfluidic cooling. The overlapped hotspots exacerbated the inherent thermal problem. While the temperature of the steady state diverged with the heatsink, the microfluidic cooling case has increased the maximum temperature from 106.29°C to 166.57°C. Although the temperature of the hotspots is higher than operation temperatures, the microfluidic cooling system sustained the severe thermal cases with superiorities of heat distribution and direct cooling capacity. The severe hotspot problem due to the thermal gradient of the microfluidic system can be solved with an alternating flow method. The advanced thermal management technique will be described in chapter 6.3.

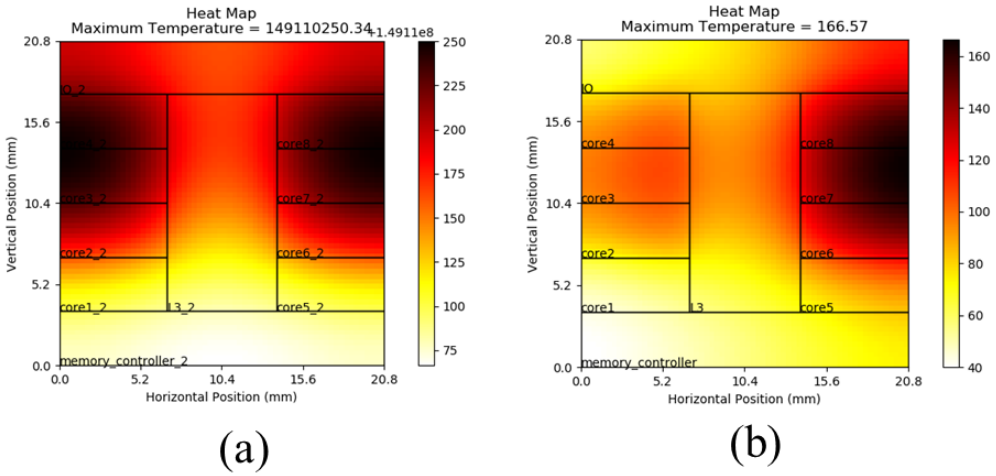


Figure 61. Temperatures throughout a 3D high-performance chiplet with (a) the server heatsink and (b) the microfluidic cooling

6.2 From 2.5D to 3D chiplet integration

6.2.1 Thermal issue of Processing-in-memory stack

Heterogeneous stacking of processor and memory in a single package shortens the data movement and enables the processing to be placed in the main memory system. Without integrating the PIM, 3D stacked memory has a tight thermal budget issue due to the inherent thermal issue. In the case of four memory layers and a logic layer stack with 64 mm^2 areas the total power density can be 17.3 W/cm^2 and 11.2 W/cm^2 dissipated in the logic layer alone [7]. In the thermal analysis, the maximum available power density in the logic die is 13.3 W/cm^2 with passive heat sink cooling that allows the operating temperature to remain under 85°C [9].

The processor-memory stack causes an increase in power consumption, exacerbating this power budget issue in the 3D memory. Recent studies indicate that the PIM integrated 3D-stacked memory suffers from the power budget issue. Milojevic et al. reported that the logic die consumes 16W/cm^2, stacked with two memory dies. The temperature of the logic and memory dies with a passive heatsink increased up to 200°C and 175°C, respectively [14].

Additionally, Ahn et al. accomplished a 30X speedup with PIM while the power consumption increased by 40% compared to a conventional system [15]. Despite these performance improvements, the power density of the logic layer is 33.2 W/cm^2, which violates the power dissipation limit of DRAM. In brief, Processing-in-memory is an efficient computation method for big-data processing that allows for performance scalability with the size of the data. However, the realization of this concept in a 3D-stacked memory is extremely challenging without a thermal management solution.

6.2.1 Chiplet modeling

We model the processor-memory chiplet system with two cases. The first case is a 2.5D chiplet integration, and it has one processor chip in the center and six high bandwidth memory (HBM) on the sides. The 3D integration counterpart is configurable by stacking the memories on the processor. The size of the processor and memory chips are 552mm^2 and 92mm^2, respectively. The floorplans shown in figures 62 and 63 of the chiplets are generated using ArchFP. We also hypothesized the power consumption as 100W for the processor and 18.4 W for the HBM. It is simulated with microfluidic cooling as well as conventional heatsinks. We assumed that the microchannel had integrated into the interposer layer. The 2.5D chiplet has 32 microchannels of a height of 100 um and a width of 346 μm. And the 3D chiplet has 62 microchannels of 100 um height and a width of 191 μm. The microchannel cooling features of 2.5D and 3D chiplet have a flow rate of 35.5 ml/min and 36.3 ml/min under the 200 mbar pump supply pressure, respectively. 200 mbar is affordable fluidic pressure to the 9V reservoir pumps of commercialized liquid cooling systems.

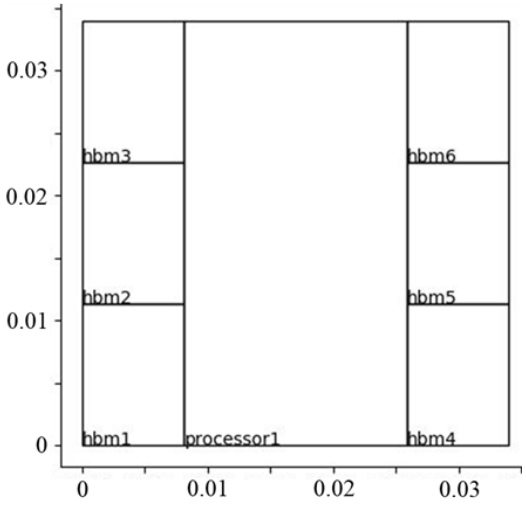


Figure 62. Floor Plan diagrams of 2.5D chiplet integrating processor and memories

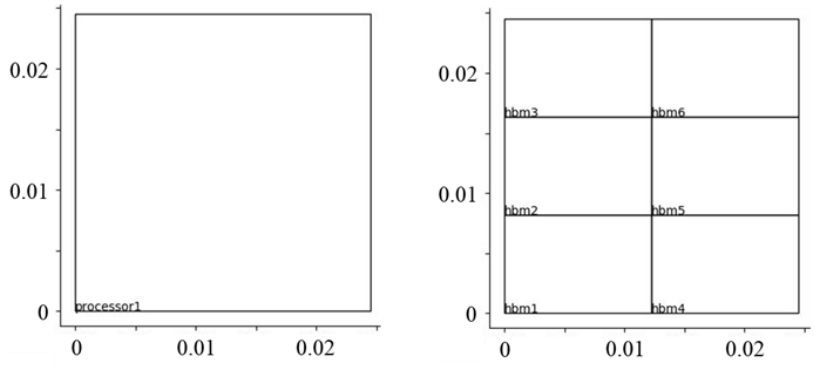


Figure 63. Floor Plan diagrams of (a) processor and (b) memories of 3D chiplet integration.

6.2.2 Thermal simulations chiplet with heatsinks

To investigate the thermal implications of the transition from 2.5D to 3D chiplet, we perform thermal simulations using HotSpot 7.0. By exploring different cooling features, we evaluate allowed maximum power densities and operating temperatures in 2.5D and 3D chiplet. In our first example, we model a chiplet composed of a processor and 3D memories. Figure 64 and 65 show the maximum temperature of the 2.5D and 3D chiplet with increasing power consumption in the processor. The power consumption of the memories remains at a constant 18.4 W.

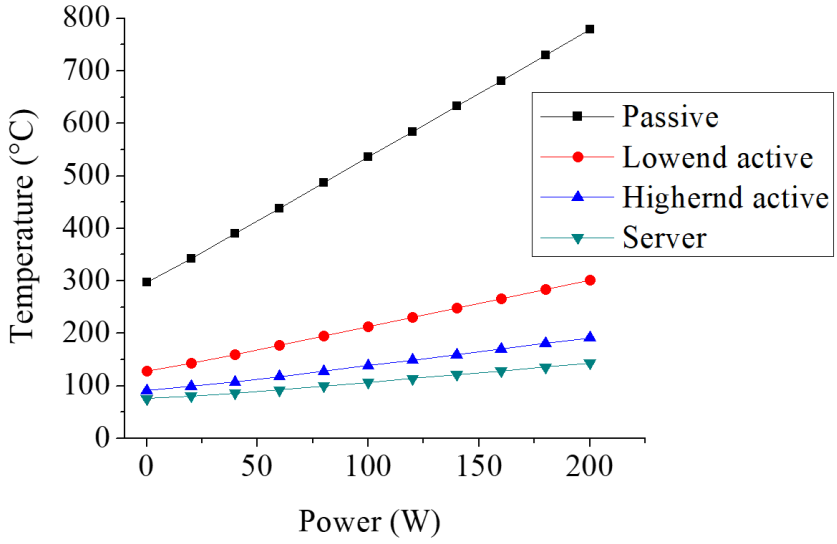


Figure 64. The temperature of 2.5D chiplet with heatsink cooling

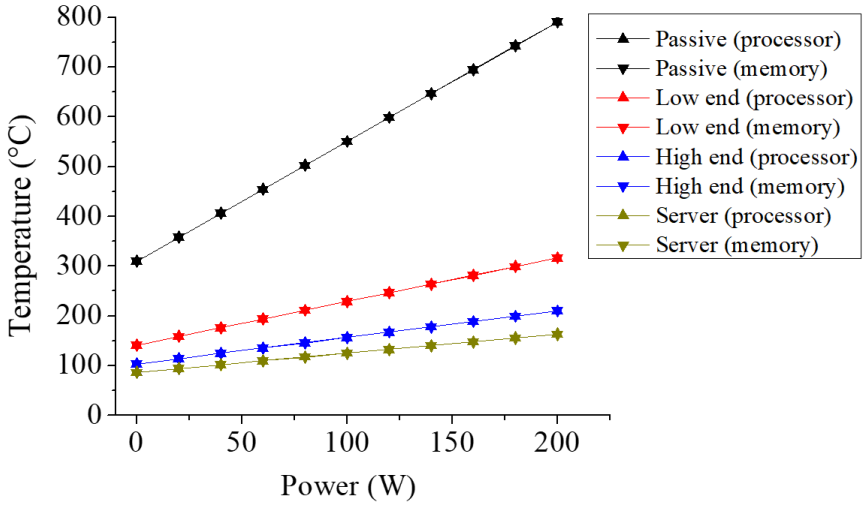


Figure 65. The temperature of 3D chiplet with heatsink cooling

The passive and low-end active heat sinks cannot remove the heat flux generated from the 3D stacked memory alone. When the power consumption of the processor is 100W, the maximum temperature of the 2.5D chiplet with high-end and server heat sinks are 138.66°C and 106.78°C, respectively. The maximum temperatures of 3D chiplet at the same condition are

157.01°C and 125.09°C. The temperature of the 3D chiplet is higher than 2.5D due to the higher power density, but in either case of the chiplet, the cooling capacity by the air-cooling heat sinks has reached the limit. Therefore, novel thermal management methods such as microfluidic cooling are required to maintain the operating temperature under the limit.

6.2.3 Thermal simulations chiplet with microfluidic cooling

Figures 66 and 67 show the temperature comparison of a 2.5D chiplet with a heatsink of convection thermal resistance is 0.2 K/W and microfluidic cooling. The maximum temperature of chiplet with microfluidic cooling is approximately 9.5°C lower than the heat sink. Microfluidic cooling has superior heat spreading characteristics as well as higher heat removal capacity. Although microfluidic is advantageous to heat sinks in cooling behavior, it still lacks the cooling capacity to maintain the operating temperature under the limit. Increasing pump pressure could be a solution, but adding another cooling layer is more effective. When the microfluidic cooling is integrated at the interposer with the top surface heat sink, the operating temperature decreases to 47.2°C.

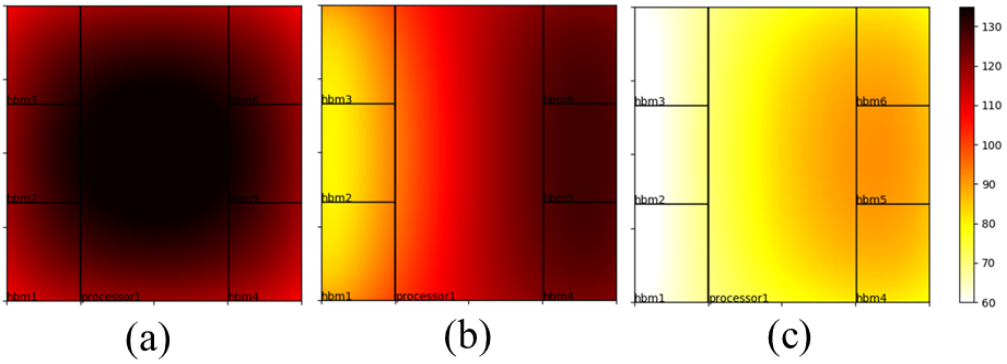


Figure 66. Temperatures throughout a 2.5D chiplet with (a) server heat sink, (b) microfluidic cooling, and (c) hybrid of heat sink and microfluidic cooling at power dissipation of 100 watts.

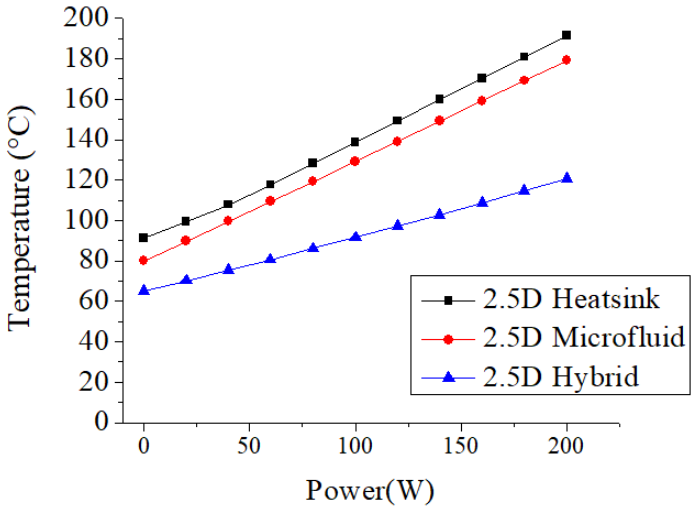


Figure 67. Temperature comparison of 2.5D chiplet with server heat sink, microfluidic cooling, and hybrid of heat sink and microfluidic cooling.

Next, the 3D chiplet simulation results are shown in figures 68 and 69. The 3D integration has a critical thermal issue in the memory layer. As the heat generated from the processor is diffused through the layer to the heatsink layer on the top, the vertical thermal gradient produces an increase in the temperature from the memory layer exceeding the operating temperature limits.

Microfluidic cooling reduces the maximum temperature of the chiplet by 29.17°C with a flow rate of 36.3 ml/min under the 200 mbar supply. While the microfluidic cooling is located in the interconnecting interposer layer, the vertical thermal gradient problem has also been alleviated. The figure 68 (c) and (f) shows the hybrid cooling system with the server heatsink on the top surface. The hybrid cooling system decreases the maximum temperature of the chiplet by 63.83°C lower than the heatsink system. For the 3D chiplet system, the advantages of microfluidic cooling can have a larger impact when compared to 2.5D integration.

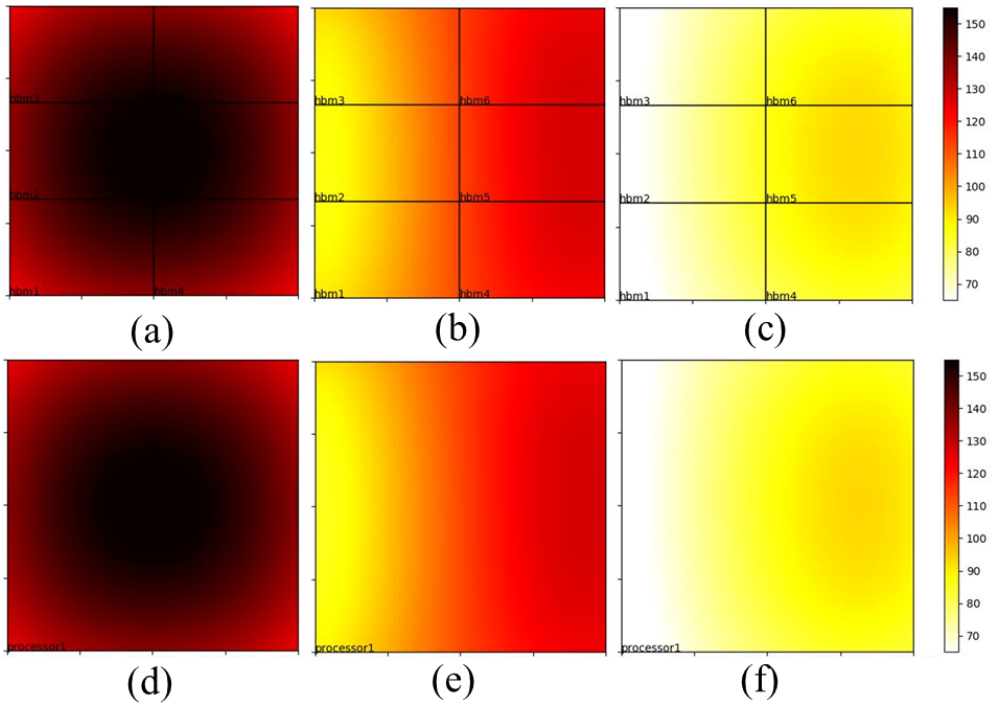


Figure 68. Temperatures throughout a 3D chiplet with (a) and (d) server heat sink, (b) and (e) microfluidic cooling, and (c) and (f) hybrid of the heat sink and microfluidic cooling at a power dissipation of 100 watts.

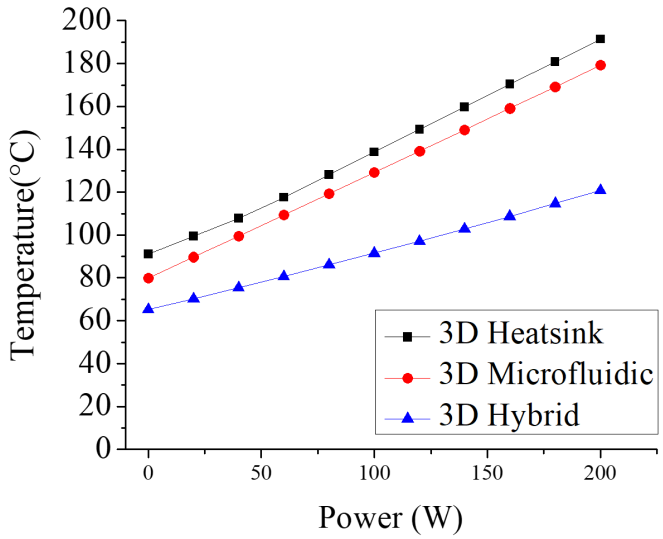


Figure 69. Temperature comparison of 3D chiplet with server heat sink, microfluidic cooling, and hybrid of the heatsink and microfluidic cooling.

6.3 Thermal management of 3D chiplet

6.3.1 Multilayer cooling

The thermal behavior of 3D stacks of high-performance processors is investigated in chapter 6.1. Although microfluidic cooling is superior to the conventional heatsink methods, there is still a lack of cooling capacity to handle the high-power density. One possible solution is increasing the flow rate with a higher pressure supply. With 400 mbar pressure of hydraulic pump supply, the maximum temperature of the device decreases from 166.57°C to 123.1°C. However, increasing the flow rate is not a suitable solution for the consumer electronics application because of limited resources such as cooling power budget, size of the pump, and device robustness. Multi-layer cooling is a promising solution to remove heat from the overlapped high-temperature hotspots in 3D chiplets. By adding another microfluidic layer between the processors, the maximum temperature drops to 85.41°C and 90.13°C under the same flow rate condition. Although multi-layer cooling is superior in the cooling aspect, this method may increase fabrication costs in the microchannel cooling method. The microchannel method requires additional fluidic structures, such as vertical pipes and fluidic vias, for vertical coolant flow between the layers. On the contrary, the microchamber method can configure multi-layer cooling in a cost-effective way by exploiting the existing gap between the layers and the 3D-printing package.

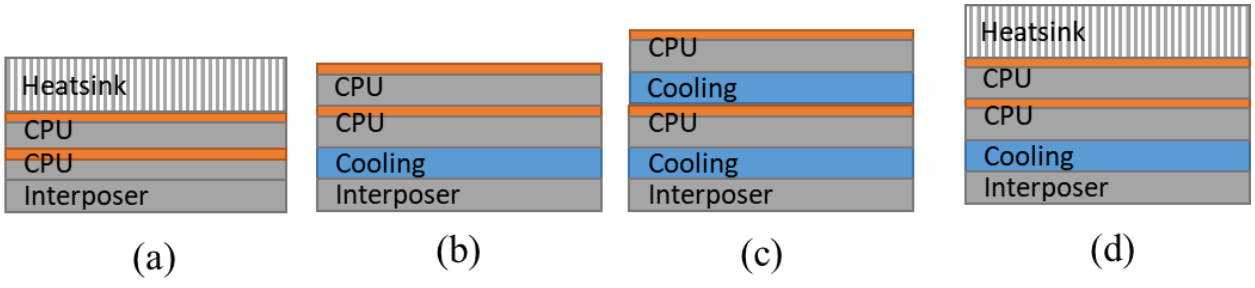
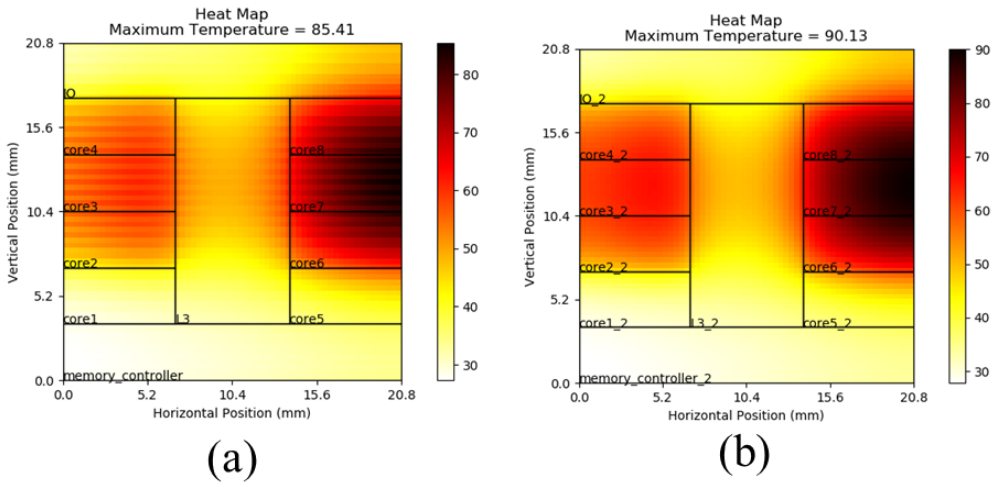


Figure 70. Cross section diagram of (a) heatsink cooling (b) microfluidic cooling, (c) multilayer cooling, and (d) hybrid cooling



(a) (b)

Figure 71. Thermal distribution of (a) lower and (b) upper processors of 3D-stacked high-performance processors with multi-layer cooling

6.3.2 Alternating flow direction

The thermal gradient is an inherent problem in microfluidic cooling methods. When the coolant flows through the microchannel or microchamber, the thermal energy is emitted to the coolant from the silicon surface. Then the heated coolant delivers the thermal energy along with the flow direction. In the worst case, this thermal gradient can be a source of a thermal failure of the chiplet. We proposed a microchamber system for the possibility of flow control in the chamber with the inlet-outlet multiplexing method. The multiplexing method is beneficial in real-time fluid control and hotspot management. For the thermal gradient problem in a microchannel system, the alternating flow direction method is an alternative to flow control. As shown in figure 72, The alternating flow direction method can significantly reduce the thermal gradient problem and peak temperature in hotspots in microchannel cooling.

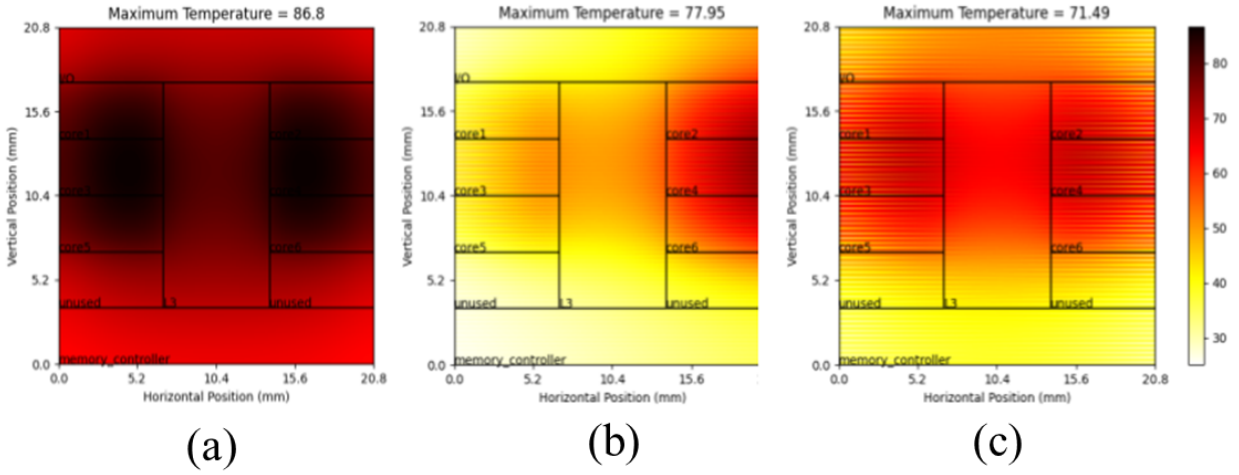


Figure 72. Temperature distribution results with (a) heatsink, (b) microchannel cooling, and (c) microchannel cooling with the alternating flow direction.

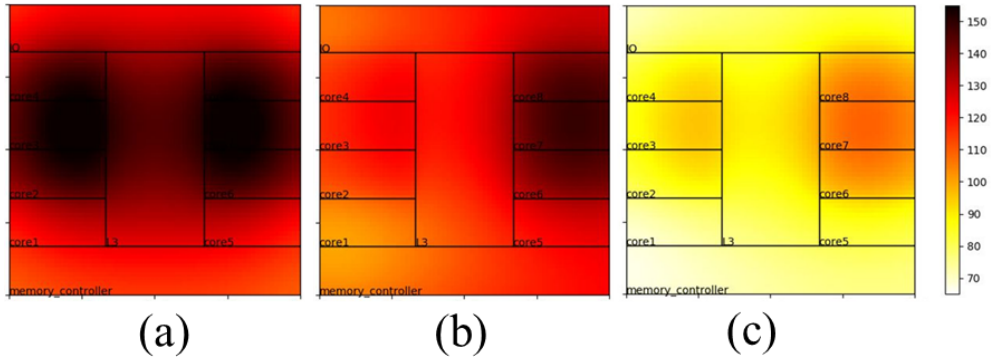


Figure 73. Temperature distribution results with (a) heatsink, (b) microchannel cooling, and (c) hybrid of the heatsink and microchannel cooling

6.3.3 Hybrid cooling

The aforementioned thermal management techniques are cooling-efficient and cost-effective methods for high-performance 3D chiplets with high-temperature hotspots. For the multi-layer stack of high-performance processors, advanced microfluidic cooling is crucial to maintain the operating temperature of the chiplets in future applications. However, in energy-efficient applications like processing-in-memory architecture, the cost of the advanced

cooling method is not appropriate for integration in the system package. For this reason, the hybrid cooling methods are suggested as an affordable option for the processing-in-memory device. An additional cooling layer on top of the chiplet can significantly reduce the temperature. As shown in figures 73 and 74, the hybrid cooling method reduces the peak temperature from 149.17°C to 109.19°C.

In conclusion, the thermal management methods of 3D chiplet can be chosen by the designers according to their application and system limitations. The new microfluidic cooling feature in HotSpot 7.0 is fast, accurate, and design flexibility to design and optimize thermal management of modern chiplets.

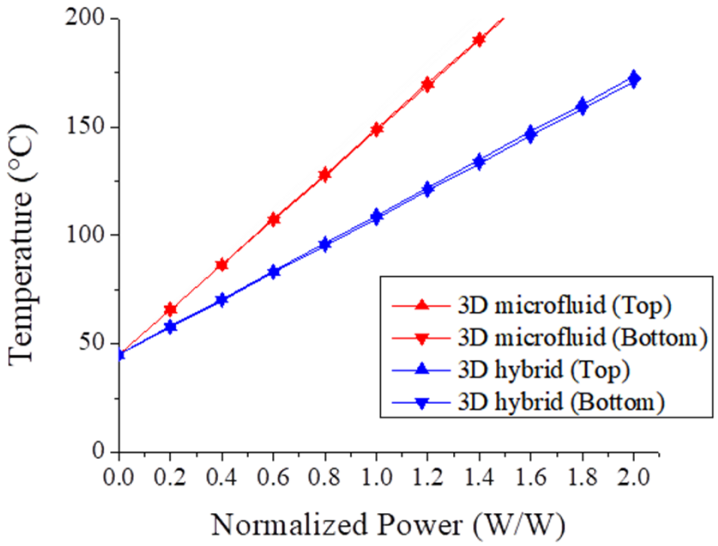


Figure 74. . Temperature comparison of 3D chiplet with microfluidic cooling and hybrid of heat sink and microfluidic cooling.

7. Conclusion and Future directions

7.1 Microchamber cooling method

We proposed the microchamber cooling method with low-cost production and superior cooling capacity. The proposed method exploits an existing die cavity between dies to build a fluidic chamber hence being able to remove heat directly from the dies with coolant flow. The microchamber structure is advantageous in fluidic control and multi-layer scalability. Also, the microchamber method achieves low-cost production by utilizing the die cavity and 3D printing package. In contrast to the conventional microchannel cooling system, the microchamber system does not have to fabricate the additional fluidic structures on the silicon wafer causing an increase in fabrication cost and chip dimensions.

The validity of microchamber device design is experimentally proven with a thermal test system. We designed and fabricated a microchamber with a height of 68 μm using the thermal test chip TTC-1002. The system senses the temperature with four sensing diodes while generating heat flux with resistors. The temperature drops to 21°C with coolant flowing through the microchamber using a paper pump.

In addition to proving the basic concept, we investigated the thermal behavior of the microfluidic cooling device. We modeled the system as a microfluidic circuit and a thermal circuit to predict the temperature of the chip with coolant flow through the microchamber of a specific flow rate under a given pressure supply. The fluidic circuit is calculated by using Hagen-Poiseuille equations [43] with the supply of the hydraulic pump with the reservoir. The thermal circuit is calculated by using thermal resistance of conduction in layers and convention between silicon and coolant.

We validated the proposed microchamber cooling system with experimental results and the conceptual prototype. The 3D printing package and encapsulation process can be improved from the experimental aspect using UV transparent and microfluidic compatible material and an adhesive dispenser system. The multiplexing flow method is another promising thermal management method to manage the peak temperature problem of arbitrary hotspots.

7.2 Thermal management with HotSpot

We implemented the thermal model to calculate the temperature of the chiplet under the specific thermal and power budgets. Processing-in-3D-memory is a commonly used example of 3D chiplets for thermal aspects. As the logic layer at the bottom generates heat flux from the increased power consumption of the integrated processing unit, the heat diffuses through the memory layers to the heatsink on the top. This vertical thermal gradient causes the thermal issue in the 3D chiplet. We calculated the temperature in the processing-in-3D-memory system for the different heatsinks and high-rise 3D-stacked IC. Because processing-in-memory computing presupposes parallel computing, the cooling budget is limited in consumer applications. Under the limited cooling capacity, the processing-in-3D-memory is limited to under a 16-layer stack due to the tight thermal and power density budget.

We investigated the thermal behavior of the chiplets with HotSpot 7.0. The recent update of the 20 years old simulation tool includes a microfluidic cooling feature for future thermal management. We simulated the thermal behavior of the microchamber embedded in the processing-in-3D-memory device and compared it with the microchannel cooling method. We found that the thermal boundary layer at the surface affects the cooling capacity of the microchamber system. However, with a detailed thermal model including interconnection micropillars in the chamber, the thermal boundary layer effects have been reduced.

We validated the result of the HotSpot simulator with the multiphysics simulator, COMSOL. The thermal behavior with an increasing flow rate in the two simulators shows identical trends in average temperature and qualitative thermal behaviors shown in thermal distribution.

The thermal management design using HotSpot can achieve a more detailed system description by integrating a performance simulator into a toolchain. McPAT generates and delivers to the HotSpot power and thermal information based on performance simulation results. PIMSim[64] configured a toolchain based on the gem5 simulator, and CoMeT[65] configured one based on the sniper simulator. The microfluidic feature of HotSpot 7.0 can provide a detailed thermal management design combined with the toolchains.

7.3 Thermal behavior in modern chiplet

We investigated the thermal characteristic of a high-performance processor in 2.5D and 3D chiplet integration. Current processors have already reached the thermal limit with passive heatsinks and are spending resources on active cooling methods. In 2.5D integration, microfluidic cooling has a higher peak temperature of 106.29°C than 102.34°C of the server heatsink system. While the average temperature in the microfluidic cooling system is 97.78°C, it is lower than that of the server heatsink which is 102.34°C. The current hotspot issue is caused by the thermal gradient along with the coolant flow direction. By alternating coolant flow direction in the adjacent channels, the heatsink problem can be solved with higher uniformity of temperature distribution. In 3D integration, conventional heatsink cooling cannot manage the high power density of 3D stacked processors. Similarly, the peak temperature of the microfluidic cooling system increased to 166.57°C. Even with two times larger pressure of 400 mbar supply, the peak temperature drops only to 123.1°C, which is higher than the operating temperature of ICs. By adding an additional fluidic cooling layer between the processors, the temperature was reduced to 90.13°C. The novel thermal management methods such as alternating flow directions and multi-layer scalability are the outstanding features of HotSpot 7.0.

We also investigated a heterogeneous processor-memory chiplet system in 2.5D and 3D integration. By stacking the 3D memories on top of the processor, the chiplet can shorten data movement and enhance the processing-in-memory capability. From a thermal perspective, 3D stacking of processor and memory causes a significant temperature accretion from 106.78°C to 125.09°C. A microfluidic cooling with a 200 mbar pressure supply can reduce the peak temperature by 29.17°C. The 200 mbar pressure is chosen because it is a nominal value from a commercially available liquid cooler and hydraulic pump system. The novel thermal management method can be applied to secure a sufficient thermal budget for the 3D chiplet. While multi-layer cooling is the most effective cooling solution for 3D chiplet, a hybrid cooling system of microfluidic cooling and heatsink is an affordable solution for commercial applications. When the microfluidic cooling is integrated at the interposer with the top surface heat sink, the operating temperature decreases to 47.2°C. The heatsink on the top can be replaced with a liquid cold plate by sharing the resources with the microfluidic cooling system.

Appendix A. List of Publications

1. Publications
    1. **Jun-Han Han**, Karina Torres-Castro, Robert E. West, Walter Varhue, Nathan Swami, and Mircea Stan, "Microfluidic Cooling for 3D-IC with 3D Printing Package," *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S),* 2019.
    2. **Jun-Han Han**, Karina Torres-Castro, Robert E. West, Nathan Swami, and Mircea Stan, "Power and Thermal Modeling of In-3D-Memory Computing," *International Symposium on Devices, Circuits and Systems (ISDCS)*, 2021.
    3. **Jun-Han Han**, Karina Torres-Castro, Robert E. West, Nathan Swami, and Mircea Stan, "Thermal Analysis of Microchamber Cooling for Processing-in-Memory Integrated 3D-Memory," *22nd International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems (EuroSimE)*, 2021.
    4. **Jun-Han Han**, Robert E. West, Kevin Skardon, and, Mircea Stan, "Thermal Simulation of Processing-in-Memory Devices using HotSpot 7.0", *27th International Workshop on Thermal Investigations of ICs and Systems (THERMINIC),* 2021.
    5. **Jun-Han Han**, Xinfei Guo, Kevin Skardon, and, Mircea Stan, "From 2.5D to 3D Chiplet Systems: Investigation of Thermal Implications with HotSpot 7.0", *Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm*), 2022.
    6. Mircea Stan, Kevin Skadron, Xinfei Guo, **Junhan Han**, "HotSpot Through the Ages", *International Symposium on Computer Architecture (ISCA)*, 2022.
    7. Sankatali Venkateswarlu, Subrat Mishra, Herman Oprins, Bjorn Vermeersch, Moritz Brunion, **Jun-Han Han**, Mircea R. Stan, Pieter Weckx, and Francky Catthoor, "Thermal Performance Analysis of Mempool RISC-V Multi-core SOC", *IEEE Transactions on Very Large Scale Integration Systems (TVLSI)*, Under review.
    8. MWSCAS, Aug 2022.

2. Workshops
    1. **Jun-Han Han**, Robert E. West, Karina Torres-Castro, Walter Varhue, Nathan Swami, and Mircea Stan, “Modeling and Experimentation of Microfluidic Cooling Using the Existing Die Cavity as a Fluid Chamber,” *SRC Techcon*, Austin, TX, 2019.
    2. Robert E. West, **Jun-Han Han**, Karina Torres-Castro, Mircea Stan, and Kevin Skardon, “Extending HotSpot to Model 3D Cooling “, *SRC Techcon*, 2020.
    3. **Jun-Han Han**, Karina Torres-Castro, Robert E. West, Nathan Swami, and Mircea Stan, “Thermal Analysis of Microfluidic cooling in Processing-in-3D-Stacked Memory,” *SRC Techcon*, 2020.
    4. **Jun-Han Han**, Karina Torres-Castro, Robert E. West, Nathan Swami, and Mircea Stan, ‘Microfluidic Cooling for PIM integrated 3D-Stacked Memory’ DAC Young fellow, Design Automation Conference 2020.

3. Poster Presentations
    1. UVA ECE department presentation, 2018, 2019.
    2. CRISP annual reviews 2018, 2019, 2020, 2021
    3. MIST center proposal ‘Microfluidic Cooling for 3D-IC’ 2018.

4. Award
    1. University Demo Best Demonstration, Honorable Mention Issued by THE 31ST ACM SIGDA UNIVERSITY DEMONSTRATION AT DAC 2021 (UBOOTH@DAC), 2021.

5. Industry relationship
    1. Samsung Research America - Hyunsik Chae
    2. ETRI - Jeongik Lee
    3. ARM - Supreet Jeloka, Mudit Bhargava

4. imec - Francky Catthoor, Sankatali Venkateswarlu, Herman Oprins, Bjorn Vermeersch

6. Publications of past projects (2017 ~ 2020)
    1. Sukyung Choi, Chan-mo Kang, Chun-Won Byun, Hyunsu Cho, Byoung-Hwa Kwon, **Jun-Han Han**, Jong-Heon Yang, Jin-Wook Shin, Chi-Sun Hwang, Nam Sung Cho, Kang Me Lee, Hee-Ok Kim, Eungjun Kim, Seunghyup Yoo, Hyunkoo Lee, “Thin-film transistor-driven vertically stacked full-color organic light-emitting diodes for high-resolution active-matrix displays” *Nature communications 11 (1), 1-9, 2020*
    2. O Eun Kwon, Jin-Wook Shin, Himchan Oh, Chan-mo Kang, Hyunsu Cho, Byoung-Hwa Kwon, Chun-Won Byun, Jong-Heon Yang, Kang Me Lee, **Jun-Han Han**, Nam Sung Cho, Jong Hyuk Yoon, Seung Jin Chae, Jin Sung Park, Hyunkoo Lee, Chi-Sun Hwang, Jaehyun Moon, Jeong-Ik Lee, ”A prototype active-matrix OLED using graphene anode for flexible display application” *Journal of Information Display 21 (1), 49-56, 2020*
    3. Jaehyun Moon, Jin-Wook Shin, Hyunsu Cho, **Jun-Han Han**, Byoung-Hwa Kwon, Jeong-Ik Lee, Nam Sung Cho, “Technical issues and integration scheme for graphene electrode OLED panels” *Graphene for Flexible Lighting and Displays, 73-98, 2020*
    4. **Jun-Han Han**, Jaehyun Moon, Doo-Hee Cho, Jin-Wook Shin, Hye Yong Chu, Jeong-Ik Lee, Nam Sung Cho, Jonghee Lee, “Luminescence enhancement of OLED lighting panels using a microlens array film”, *Journal of Information Display* 19 (4), 179-184, 2018.
    5. Hyunkoo Lee, Hyunsu Cho, Chun-Won Byun, Chan-Mo Kang, **Jun-Han Han,** Jeong-Ik Lee, Hokwon Kim, Jeong Hwan Lee, Minseok Kim, Nam Sung Cho “Device Characteristics of Top-Emitting Organic Light-Emitting Diodes Depending on Anode Materials for CMOS-Based OLED Microdisplays”, *IEEE Photonics Journal* 10 (6), 1-9

6. Jaehyun Moon, Hyunsu Cho, Min-Jae Maeng, Kwangmin Choi, Đăng Thành Nguyen, **Jun-Han Han**, Jin-Wook Shin, Byoung-Hwa Kwon, Jonghee Lee, Seungmin Cho, Jeong-Ik Lee, Yongsup Park, Jong-Sook Lee, Nam Sung Cho, ”Mechanistic understanding of improved performance of graphene cathode inverted organic light-emitting diodes by photoemission and impedance spectroscopy” *ACS applied materials & interfaces 10 (31), 26456-26464, 2018*
7. Hyunkoo Lee, Hyunsu Cho, Chun-Won Byun, Chan-Mo Kang, **Jun-Han Han**, Jeong-Ik Lee, Hokwon Kim, Jeong Hwan Lee, Minseok Kim, Nam Sung Cho, “Color-tunable organic light-emitting diodes with vertically stacked blue, green, and red colors for lighting and display applications” *Optics express 26 (14), 18351-18361, 2018.*
8. Jin-Wook Shin, Hyunsu Cho, Byoung-Hwa Kwon, **Jun-Han Han**, Kang Me Lee, Jong-Heon Yang, Chun-Won Byun, Jeong-Ik Lee, Jaehyun Moon, Nam Sung Cho, ”Flexible OLED Panels with Pixelated Graphene Anode” *SID Symposium Digest of Technical Papers 49 (1), 415-417, 2018*
9. Jin-Wook Shin, Hyunsu Cho, Jonghee Lee, Jaehyun Moon, **Jun-Han Han**, Kisoo Kim, Seungmin Cho, Jeong-Ik Lee, Byoung-Hwa Kwon, Doo-Hee Cho, Kang Me Lee, Maki Suemitsu, Nam Sung Cho, “Overcoming the efficiency limit of organic light-emitting diodes using ultra-thin and transparent graphene electrodes” *Optics express 26 (2), 617-626, 2018*
10. Jin-Wook Shin, **Jun-Han Han*(Co-first)**, Hyunsu Cho, Jaehyun Moon, Byoung-Hwa Kwon, Seungmin Cho, Taeshik Yoon, Taek-Soo Kim, Maki Suemitsu, Jeong-Ik Lee, Nam Sung Cho, “Display process compatible accurate graphene patterning for OLED applications”, *2D Materials* 5 (1), 014003, 2017.

Bibliography

[1] "2015 International Technology Roadmap for Semiconductors (ITRS)," *Semiconductor Industry Association*, Jun. 05, 2015. https://www.semiconductors.org/resources/2015-international-technology-roadmap-for-semiconductors-itrs/ (accessed Jul. 01, 2022).

[2] A. Boroumand *et al.*, "Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks," *ACM SIGPLAN Not.*, vol. 53, no. 2, pp. 316–331, Mar. 2018, doi: 10.1145/3296957.3173177.

[3] S. Naffziger, K. Lepak, M. Paraschou, and M. Subramony, "2.2 AMD Chiplet Architecture for High-Performance Server and Desktop Products," in *2020 IEEE International Solid-State Circuits Conference - (ISSCC)*, Feb. 2020, pp. 44–45. doi: 10.1109/ISSCC19947.2020.9063103.

[4] H. Braunisch, A. Aleksov, S. Lotz, and J. Swan, "High-speed performance of Silicon Bridge die-to-die interconnects," in *2011 IEEE 20th Conference on Electrical Performance of Electronic Packaging and Systems*, Oct. 2011, pp. 95–98. doi: 10.1109/EPEPS.2011.6100196.

[5] F. Sheikh, R. Nagisetty, T. Karnik, and D. Kehlet, "2.5D and 3D Heterogeneous Integration: Emerging applications," *IEEE Solid-State Circuits Mag.*, vol. 13, no. 4, pp. 77–87, 2021, doi: 10.1109/MSSC.2021.3111386.

[6] C.-C. Lee *et al.*, "An Overview of the Development of a GPU with Integrated HBM on Silicon Interposer," in *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)*, May 2016, pp. 1439–1444. doi: 10.1109/ECTC.2016.348.

[7] H. Tsugawa *et al.*, "Pixel/DRAM/logic 3-layer stacked CMOS image sensor technology," in *2017 IEEE International Electron Devices Meeting (IEDM)*, Dec. 2017, p. 3.2.1-3.2.4. doi: 10.1109/IEDM.2017.8268317.

[8] M. M. Shulaker *et al.*, "Three-dimensional integration of nanotechnologies for computing and data storage on a single chip," *Nature*, vol. 547, no. 7661, pp. 74–78, Jul. 2017, doi: 10.1038/nature22994.

[9] Y. Ma, L. Delshadtehrani, C. Demirkiran, J. L. Abellan, and A. Joshi, "TAP-2.5D: A Thermally-Aware Chiplet Placement Methodology for 2.5D Systems," in *2021 Design, Automation Test in Europe Conference Exhibition (DATE)*, Feb. 2021, pp. 1246–1251. doi: 10.23919/DATE51398.2021.9474011.

[10] A. Kaul, S. K. Rajan, M. Obaidul Hossen, G. S. May, and M. S. Bakir, "BEOL-Embedded 3D Polylithic Integration: Thermal and Interconnection Considerations," in *2020 IEEE 70th Electronic Components and Technology Conference (ECTC)*, Jun. 2020, pp. 1459–1467. doi: 10.1109/ECTC32862.2020.00231.

[11] B. Rakesh, K. Mahindra, M. Sai Venkat Goud, N. Arun Vignesh, T. Padma, and A. Kumar Panigrahy, "Facile approach to mitigate thermal issues in 3D IC integration using effective FIN orientation," *Mater. Today Proc.*, vol. 33, pp. 3085–3088, Jan. 2020, doi: 10.1016/j.matpr.2020.03.663.

[12] P. Leduc *et al.*, "Challenges for 3D IC integration: bonding quality and thermal management," in *2007 IEEE International Interconnect Technology Conferencee*, Jun. 2007, pp. 210–212. doi: 10.1109/IITC.2007.382392.

[13] "Heterogeneous Integration Roadmap: Chapter 20. Thermal." 2021.

[14] A. Bar-Cohen, “Gen-3 Thermal Management Technology: Role of Microchannels and Nanostructures in an Embedded Cooling Paradigm,” *J. Nanotechnol. Eng. Med.*, vol. 4, no. 2, p. 020907, May 2013, doi: 10.1115/1.4023898.
[15] T. Acikalin and C. Schroeder, “Direct liquid cooling of bare die packages using a microchannel cold plate,” in *Fourteenth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, Orlando, FL, USA, May 2014, pp. 673–679. doi: 10.1109/ITHERM.2014.6892346.
[16] Jun-Han Han, Karina Torres-Castro, Robert E. West, Walter Varhue, Nathan Swami, Mircea Stan, “Microfluidic Cooling for 3D-IC with 3D Printing Package,” in *IEEE SOI-3D-SUBTHRESHOLD MICROELECTRONICS TECHNOLOGY UNIFIED (S3S) CONFERENCE*, 2019, vol. 45, p. 15.04.
[17] G. Deshpande and D. K. Bhatia, “Microchannels for thermal management in FPGAs,” in *2017 International Conference on ReConFigurable Computing and FPGAs (ReConFig)*, Cancun, Dec. 2017, pp. 1–5. doi: 10.1109/RECONFIG.2017.8279803.
[18] Y. Eckert, “Thermal Feasibility of Die-Stacked Processing in Memory,” *2nd Workshop -Data Process. WoNDP*, 2014.
[19] S. Li, D. Reddy, and B. Jacob, “A performance & power comparison of modern high-speed DRAM architectures,” in *Proceedings of the International Symposium on Memory Systems*, Alexandria Virginia USA, Oct. 2018, pp. 341–353. doi: 10.1145/3240302.3240315.
[20] J. Lin, “Thermal modeling and management of DRAM memory systems,” p. 94.
[21] M. J. Khurshid and M. Lipasti, “Data compression for thermal mitigation in the Hybrid Memory Cube,” in *2013 IEEE 31st International Conference on Computer Design (ICCD)*, Asheville, NC, USA, Oct. 2013, pp. 185–192. doi: 10.1109/ICCD.2013.6657041.
[22] D. Lee *et al.*, “Adaptive-latency DRAM: Optimizing DRAM timing for the common-case,” in *2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*, Burlingame, CA, USA, Feb. 2015, pp. 489–501. doi: 10.1109/HPCA.2015.7056057.
[23] T. Wei, H. Oprins, V. Cherman, E. Beyne, and M. Baelmans, “Conjugate Heat Transfer and Fluid Flow Modeling for Liquid Microjet Impingement Cooling with Alternating Feeding and Draining Channels,” *Fluids*, vol. 4, no. 3, p. 145, Aug. 2019, doi: 10.3390/fluids4030145.
[24] C. R. King, D. Sekar, M. S. Bakir, B. Dang, J. Pikarsky, and J. D. Meindl, “3D stacking of chips with electrical and microfluidic I/O interconnects,” in *2008 58th Electronic Components and Technology Conference*, Lake Buena Vista, FL, USA, May 2008, pp. 1–7. doi: 10.1109/ECTC.2008.4549941.
[25] Y. Zhang, L. Zheng, and M. S. Bakir, “3-D Stacked Tier-Specific Microfluidic Cooling for Heterogeneous 3-D ICs,” *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 3, no. 11, pp. 1811–1819, Nov. 2013, doi: 10.1109/TCPMT.2013.2281605.
[26] A. Bar-Cohen, J.J Maurer, and J.G. Felbinger, “DARPA’s Intra/Interchip Enhanced Cooling (ICECool) Program,” *CS MANTECH Conf*, pp. 171–174, 2013.
[27] W. Yueh, Z. Wan, Y. Joshi, and S. Mukhopadhyay, “Design, Characterization, and Application of a Field-Programmable Thermal Emulation Platform,” *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 6, no. 9, pp. 1330–1339, Sep. 2016, doi: 10.1109/TCPMT.2016.2578347.
[28] T. E. Sarvey *et al.*, “Monolithic Integration of a Micropin-Fin Heat Sink in a 28-nm

FPGA," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 7, no. 10, pp. 1617–1624, Oct. 2017, doi: 10.1109/TCPMT.2017.2740721.
[29] A. Sridhar, A. Vincenzi, M. Ruggiero, T. Brunschwiler, and D. Atienza, "3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling," in *2010 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, San Jose, CA, USA, Nov. 2010, pp. 463–470. doi: 10.1109/ICCAD.2010.5653749.
[30] C. R. King, J. Zaveri, M. S. Bakir, and J. D. Meindl, "Electrical and fluidic C4 interconnections for inter-layer liquid cooling of 3D ICs," in *2010 Proceedings 60th Electronic Components and Technology Conference (ECTC)*, Las Vegas, NV, USA, 2010, pp. 1674–1681. doi: 10.1109/ECTC.2010.5490755.
[31] X. Y. Chen, K. C. Toh, J. C. Chai, and D. Pinjala, "Direct liquid cooling of a stacked multichip module," in *4th Electronics Packaging Technology Conference, 2002.*, Singapore, 2002, pp. 380–384. doi: 10.1109/EPTC.2002.1185702.
[32] A. Bar-Cohen, J. R. Sheehan, and E. Rahim, "Two-Phase Thermal Transport in Microgap Channels—Theory, Experimental Results, and Predictive Relations," *Microgravity Sci Technol*, p. 15, 2012.
[33] T. Brunschwiler *et al.*, "Heat-removal performance scaling of interlayer cooled chip stacks," in *2010 12th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, Las Vegas, NV, USA, Jun. 2010, pp. 1–12. doi: 10.1109/ITHERM.2010.5501254.
[34] M. S. Bakir *et al.*, "3D heterogeneous integrated systems: Liquid cooling, power delivery, and implementation," in *2008 IEEE Custom Integrated Circuits Conference*, San Jose, CA, USA, Sep. 2008, pp. 663–670. doi: 10.1109/CICC.2008.4672173.
[35] Y. Zhang, A. Dembla, Y. Joshi, and M. S. Bakir, "3D stacked microfluidic cooling for high-performance 3D ICs," in *2012 IEEE 62nd Electronic Components and Technology Conference*, San Diego, CA, USA, May 2012, pp. 1644–1650. doi: 10.1109/ECTC.2012.6249058.
[36] D. Sekar *et al.*, "A 3D-IC Technology with Integrated Microchannel Cooling," in *2008 International Interconnect Technology Conference*, Burlingame, CA, USA, Jun. 2008, pp. 13–15. doi: 10.1109/IITC.2008.4546911.
[37] L. Zheng, Y. Zhang, and M. S. Bakir, "A Silicon Interposer Platform Utilizing Microfluidic Cooling for High-Performance Computing Systems," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 5, no. 10, pp. 1379–1386, Oct. 2015, doi: 10.1109/TCPMT.2015.2470544.
[38] V. Sahu, Y. K. Joshi, A. G. Fedorov, Je-Hyeong Bahk, Xi Wang, and A. Shakouri, "Experimental Characterization of Hybrid Solid-State and Fluidic Cooling for Thermal Management of Localized Hotspots," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 5, no. 1, pp. 57–64, Jan. 2015, doi: 10.1109/TCPMT.2014.2332516.
[39] B. Shi and A. Srivastava, "TSV-constrained micro-channel infrastructure design for cooling stacked 3D-ICs," in *Proceedings of the 2012 ACM international symposium on International Symposium on Physical Design - ISPD '12*, Napa, California, USA, 2012, p. 113. doi: 10.1145/2160916.2160941.
[40] B. Shi, A. Srivastava, and P. Wang, "Non-uniform micro-channel design for stacked 3D-ICs," in *Proceedings of the 48th Design Automation Conference on - DAC '11*, San Diego, California, 2011, p. 658. doi: 10.1145/2024724.2024874.

[41] A. Sridhar, A. Vincenzi, D. Atienza, and T. Brunschwiler, “3D-ICE: A Compact Thermal Model for Early-Stage Design of Liquid-Cooled ICs,” *IEEE Trans. Comput.*, vol. 63, no. 10, pp. 2576–2589, Oct. 2014, doi: 10.1109/TC.2013.127.
[42] G. Chen, J. Kuang, Z. Zeng, H. Zhang, E. F. Y. Young, and B. Yu, “Minimizing Thermal Gradient and Pumping Power in 3D IC Liquid Cooling Network Design,” in *Proceedings of the 54th Annual Design Automation Conference 2017*, Austin TX USA, Jun. 2017, pp. 1–6. doi: 10.1145/3061639.3062285.
[43] Bruus, “Theoretical microfluidics,” 2006.
[44] F. Kreith, Ed., *The CRC handbook of thermal engineering*. Boca Raton, Fla: CRC Press, 2000.
[45] J. T. Pawlowski, “Hybrid memory cube (HMC),” in *2011 IEEE Hot Chips 23 Symposium (HCS)*, Stanford, CA, USA, Aug. 2011, pp. 1–24. doi: 10.1109/HOTCHIPS.2011.7477494.
[46] D. Milojevic *et al.*, “Thermal characterization of cloud workloads on a power-efficient server-on-chip,” in *2012 IEEE 30th International Conference on Computer Design (ICCD)*, Montreal, QC, Canada, Sep. 2012, pp. 175–182. doi: 10.1109/ICCD.2012.6378637.
[47] J. Ahn, S. Hong, S. Yoo, O. Mutlu, and K. Choi, “A scalable processing-in-memory accelerator for parallel graph processing,” in *Proceedings of the 42nd Annual International Symposium on Computer Architecture*, Portland Oregon, Jun. 2015, pp. 105–117. doi: 10.1145/2749469.2750386.
[48] D. Zhang, N. Jayasena, A. Lyashevsky, J. L. Greathouse, L. Xu, and M. Ignatowski, “TOP-PIM: throughput-oriented programmable processing in memory,” in *Proceedings of the 23rd international symposium on High-performance parallel and distributed computing - HPDC ’14*, Vancouver, BC, Canada, 2014, pp. 85–98. doi: 10.1145/2600212.2600213.
[49] H. A. D. Nguyen, J. Yu, M. A. Lebdeh, M. Taouil, S. Hamdioui, and F. Catthoor, “A Classification of Memory-Centric Computing,” *ACM J. Emerg. Technol. Comput. Syst.*, vol. 16, no. 2, pp. 1–26, Apr. 2020, doi: 10.1145/3365837.
[50] L. Nai, R. Hadidi, H. Xiao, H. Kim, J. Sim, and H. Kim, “Thermal-aware processing-in-memory instruction offloading,” *J. Parallel Distrib. Comput.*, vol. 130, pp. 193–207, Aug. 2019, doi: 10.1016/j.jpdc.2019.03.005.
[51] R. Zhang, M. R. Stan, and K. Skadron, “HotSpot 6.0: Validation, Acceleration and Extension,” p. 8.
[52] W. Huang, M. R. Stan, and K. Skadron, “Physically-based compact thermal modeling—achieving parametrization and boundary condition independence,” Oct. 2004.
[53] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, “Temperature-Aware Microarchitecture,” 2003.
[54] H. Zhu, F. Hu, H. Zhou, D. Z. Pan, and D. Zhou, “Interlayer Cooling Network Design for High-Performance 3D ICs Using Channel Patterning and Pruning,” vol. 37, no. 4, p. 12, 2018.
[55] “UVA HotSpot 7.0,” *GitHub*. https://github.com/uvahotspot (accessed Dec. 18, 2021).
[56] J.-H. Han, R. E. West, K. Skadron, and M. R. Stan, “Thermal Simulation of Processing-in-Memory Devices using HotSpot 7.0,” in *2021 27th International Workshop on Thermal Investigations of ICs and Systems (THERMINIC)*, Sep. 2021, pp. 1–5. doi: 10.1109/THERMINIC52472.2021.9626520.
[57] J. W. Demmel, J. R. Gilbert, and X. S. Li, “SuperLU users’ guide,” LBNL--44289, 751785, Nov. 1999. doi: 10.2172/751785.

[58] J. W. Demmel, S. C. Eisenstat, J. R. Gilbert, X. S. Li, and J. W. H. Liu, “A Supernodal Approach to Sparse Partial Pivoting,” *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 3, pp. 720–755, Jan. 1999, doi: 10.1137/S0895479895291765.
[59] R. Wu, X. Zhang, Y. Fan, R. Hu, and X. Luo, “A Bi-Layer compact thermal model for uniform chip temperature control with non-uniform heat sources by genetic-algorithm optimized microchannel cooling,” *Int. J. Therm. Sci.*, vol. 136, pp. 337–346, Feb. 2019, doi: 10.1016/j.ijthermalsci.2018.10.047.
[60] S. Berry, “Increasing Heat Transfer in Microchannels with Surface Acoustic Waves,” 2014, p. 7.
[61] “Thermal boundary layer thickness and shape,” *Wikipedia*. 2018. Accessed: Jun. 28, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Thermal_boundary_layer_thickness_and_shape&oldid=1031772112
[62] A. L. Shimpi, “Intel Core i7 3960X (Sandy Bridge E) Review: Keeping the High End Alive.” https://www.anandtech.com/show/5091/intel-core-i7-3960x-sandy-bridge-e-review-keeping-the-high-end-alive (accessed Dec. 18, 2021).
[63] Intel®, “Core™ i7-5960X product specification.” www.intel.com
[64] S. Xu, X. Chen, Y. Wang, Y. Han, X. Qian, and X. Li, “PIMSim: A Flexible and Detailed Processing-in-Memory Simulator,” *IEEE Comput. Archit. Lett.*, vol. 18, no. 1, pp. 6–9, Jan. 2019, doi: 10.1109/LCA.2018.2885752.
[65] L. Siddhu *et al.*, “CoMeT: An Integrated Interval Thermal Simulation Toolchain for 2D, 2.5D, and 3D Processor-Memory Systems,” *ACM Trans. Archit. Code Optim.*, Apr. 2022, doi: 10.1145/3532185.

Appendix B. HotSpot 7.0 Tutorial

HotSpot V7.0 Tutorial

Jun-Han Han, Robert E. West, Kevin Skadron, and Mircea Stan

University of Virginia

jh2vs@Virginia.edu

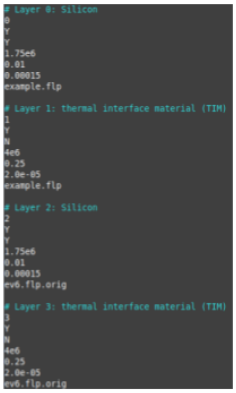
Contents

- Examples
  - Example 1: Configuration file review
  - Example 2: Block vs Grid model
  - Example 3: 3D-stack
  - Example 5: mircrofluidic cooling
- ArchFP floor-plan

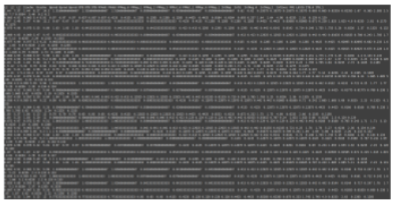
Introduction

- User input
  - The user provides pre-RTL information and HotSpot interprets it as a 3D IC

Floorplan(s)

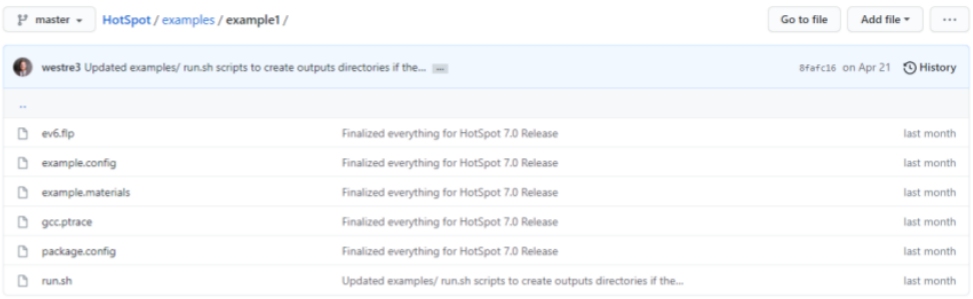


Layer Configuration



Power Traces

Example 1 of HotSpot 7.0




```
./run.sh
```

http://github.com/uvahotspot/

Example1 run.sh

UNIVERSITY of VIRGINIA

```
../../hotspot -c example.config -f ev6.flp -p gcc.ptrace -materials_file
example.materials -model_type block -steady_file outputs/gcc.steady
-o outputs/gcc.ttrace
```

Config, floor plan file, Power trace file,

- Temperature trace file

-c <config_file>

-p <ptrace_file>

-flp <floorplan_file>

-materials_file

-model_type

-steady_file
(path to a Steady Results output file.)

-o
(path to a Transient Results output file)

example1.config

UNIVERSITY of VIRGINIA

Chip

- thickness, thermal conductivity, volumetric heat capacity

Heat sink/Heat spreader/Interface

Secondary path (C4/underfill, package, solder, etc)

Microfluidic cooling parameters

- pump (pressure, R_{out}), coolant (material, temperature, heat capacity, resistivity, viscosity), channel walls, heat transfer coefficient

Floor plan parameters

example1.config

```
    # heat sink specs
# heat sink material
-material_sink none
            # convection capacitance in J/K
            -c_convec                   140.4
            # convection resistance in K/W
            -r_convec                   0.1
            # heatsink side in meters
            -s_sink                     0.06
            # heatsink thickness  in meters
            -t_sink                     0.0069
            # heatsink thermal conductivity in W/(m-K) (overrided by heat sink material)
            #-k_sink                    400.0
            # heatsink volumetric heat capacity in J/(m^3-K) (overrided by heat sink materia
            #-p_sink                    3.55e6

    # heat spreader specs
# spreader material
-material_spreader none
            # spreader side in meters
            -s_spreader                 0.03
            # spreader thickness in meters
            -t_spreader                 0.001
            # heat spreader thermal conductivity in W/(m-K) (overrided by spreader material)
            #-k_spreader                400.0
            # heat spreader volumetric heat capacity in J/(m^3-K) (overrided by spreader mat
            # -p_spreader                3.55e6

    # interface material specs
            # interface material thickness in meters
            -t_interface          2.0e-05
            # interface material thermal conductivity in W/(m-K) (overrided by interface mat
             -k_interface               4.0
            # interface material volumetric heat capacity in J/(m^3-K) (overrided by interfa
             -p_interface               4.0e6
```

```
# others
        # ambient temperature in kelvin
        -ambient                    318.15
        # initial temperatures from file
        -init_file                  (null)
        # initial temperature (kelvin) if not from file
        -init_temp                  318.15
        # steady state temperatures to file
        -steady_file            (null)
        # interval between power traces in seconds
        -sampling_intvl         0.01
        # base processor frequency in Hz
        -base_proc_freq         3e+09
        # is DTM employed?
        -dtm_used                   0
        # model type - block or grid
        -model_type                 block

        # consider temperature-leakage loop within HotSpot?
        -leakage_used 0

        # leakage calculation modes: (only valid when -leakage_used=1)
        # 0 user-defined leakage power model, do temp-leakage loop within HotSpot
        #       1 use HotLeakage -- !NOT implemented in this release!, coming later.
        -leakage_mode   0

        # use detailed package model?
        -package_model_used                 0
        -package_config_file                package.config
```


example1.materials

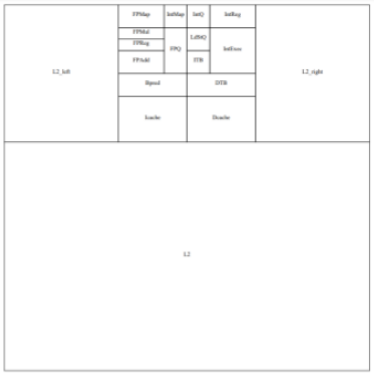
# material name # material type (solid or fluid) # thermal conductivity in W/(m-K) # volumetric heat capacity in J/(m^3-K) # dynamic viscosity in Pa-s (fluid only)	silicon solid 130.0 1630300	aluminum solid 237.0 2.422e6

example1.Floor Plan

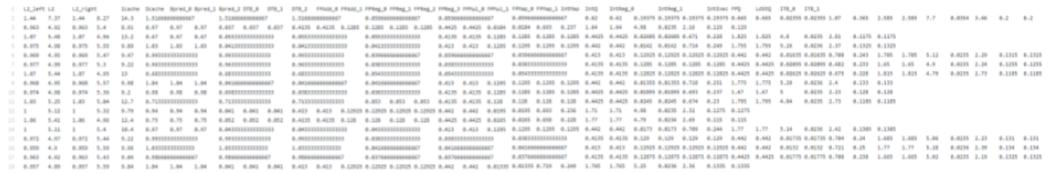
#Line Format:

<unit-name> <width><height><left-x><bottom-y>~~[<specific-heat>] [<resistivity>]~~

```
L2_left 0.004900        0.006200        0.000000        0.009800
L2      0.016000        0.009800        0.000000        0.000000
L2_right        0.004900        0.006200        0.011100        0.009800
Icache  0.003100        0.002600        0.004900        0.009800
Dcache  0.003100        0.002600        0.008000        0.009800
Bpred_0 0.001033        0.000700        0.004900        0.012400
Bpred_1 0.001033        0.000700        0.005933        0.012400
Bpred_2 0.001033        0.000700        0.006967        0.012400
DTB_0   0.001033        0.000700        0.008000        0.012400
DTB_1   0.001033        0.000700        0.009033        0.012400
DTB_2   0.001033        0.000700        0.010067        0.012400
FPAdd_0 0.001100        0.000900        0.004900        0.013100
FPAdd_1 0.001100        0.000900        0.006000        0.013100
FPReg_0 0.000550        0.000380        0.004900        0.014000
FPReg_1 0.000550        0.000380        0.005450        0.014000
FPReg_2 0.000550        0.000380        0.006000        0.014000
FPReg_3 0.000550        0.000380        0.006550        0.014000
FPMul_0 0.001100        0.000950        0.004900        0.014380
FPMul_1 0.001100        0.000950        0.006000        0.014380
FPMap_0 0.001100        0.000670        0.004900        0.015330
FPMap_1 0.001100        0.000670        0.006000        0.015330
IntMap  0.000900        0.001350        0.007100        0.014650
IntQ    0.001300        0.001350        0.008000        0.014650
IntReg_0        0.000900        0.000670        0.009300        0.015330
IntReg_1        0.000900        0.000670        0.010200        0.015330
IntExec 0.001800        0.002230        0.009300        0.013100
FPQ     0.000900        0.001550        0.007100        0.013100
LdStQ   0.001300        0.000950        0.008000        0.013700
ITB_0   0.000650        0.000600        0.008000        0.013100
ITB_1   0.000650        0.000600        0.008650        0.013100
```



example1.Power trace



```
L2_left L2      L2_right        Icache  Dcache  Bpred_0 Bpred_1 Bpred_2 DTB_0   DTB_1   DTB_2   FPAdd_0 FPAdd_1 FPReg_0 FPReg_1 FPReg_2 FPReg_3 FPMul_0 FPMul_1
FPMap_0 FPMap_1 IntMap  IntQ    IntReg_0        IntReg_1        IntExec FPQ     LdStQ   ITB_0   ITB_1

0.963   4.92    0.963   5.4     8.91    0.97    0.97    0.97    0.037   0.037   0.037   0.4135  0.4135  0.1285  0.1285  0.1285  0.1285  0.4425  0.4425  0.0184
0.0184  0.693   0.237   1.64    1.64    4.98    0.0235  2.16    0.129   0.129
```

Example1 Result

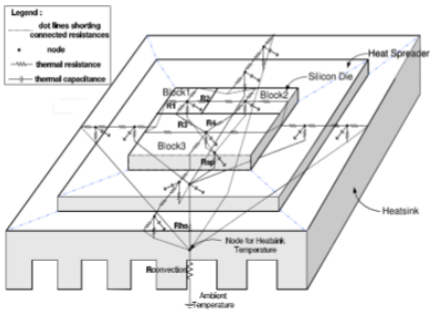
L2_left	324.35	FPReg_2	329.3
L2	323.85	FPReg_3	329.23
L2_right	324.69	FPMul_0	327.8
Icache	330.84	FPMul_1	328.23
Dcache	335.27	FPMap_0	325.52
Bpred_0	333.16	FPMap_1	326.08
Bpred_1	333.6	IntMap	329.09
Bpred_2	333.43	IntQ	328.59
DTB_0	328.92	IntReg_0	342.43
DTB_1	328.81	IntReg_1	343.01
DTB_2	328.48	IntExec	334.95
FPAdd_0	329.13	FPQ	327.96
FPAdd_1	329.43	LdStQ	336.58
FPReg_0	328.86	ITB_0	330.04
FPReg_1	329.07	ITB_1	330.64

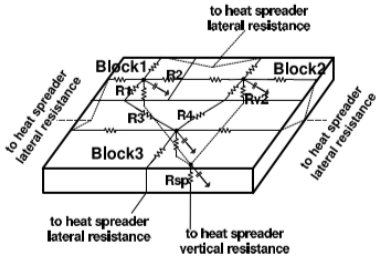
	DTB_1	DTB_2	FPAdd_0	FPAdd_1	FPReg_0	FPReg_1
2	330.39	329.98	331.41	331.76	330.99	331.27
3	328.69	328.35	329.25	329.55	328.94	329.15
4	329.33	328.98	328.84	329.17	328.77	328.98
5	328.90	328.56	329.29	329.59	328.94	329.15
6	328.75	328.42	329.23	329.53	328.92	329.13
7	328.61	328.28	329.18	329.47	328.88	329.09
8	329.23	328.89	328.82	329.14	328.74	328.95
9	328.88	328.54	329.27	329.58	328.92	329.14
10	328.63	328.30	329.19	329.48	328.88	329.08
11	329.18	328.85	328.85	329.16	328.75	328.95
12	328.74	328.41	329.13	329.42	328.86	329.07
13	329.13	328.79	328.89	329.19	328.75	328.96
14	328.90	328.56	329.17	329.47	328.88	329.09
15	328.65	328.32	329.20	329.49	328.88	329.09
16	328.83	328.49	329.26	329.56	328.91	329.12
17	328.58	328.26	329.19	329.47	328.88	329.08
18	328.78	328.44	329.26	329.56	328.91	329.12
19	328.61	328.29	329.19	329.48	328.87	329.08
20	328.60	328.27	329.20	329.48	328.87	329.08
21	329.04	328.70	328.93	329.24	328.77	328.98
22	328.95	328.61	329.08	329.38	328.83	329.04
23	328.70	328.37	329.20	329.49	328.88	329.08
24	328.78	328.45	329.26	329.56	328.90	329.12
25	328.63	328.30	329.20	329.49	328.88	329.09
26	328.72	328.39	329.24	329.53	328.90	329.11
27	328.62	328.29	329.20	329.49	328.88	329.08

gcc.steady gcc.ttrace

Block model

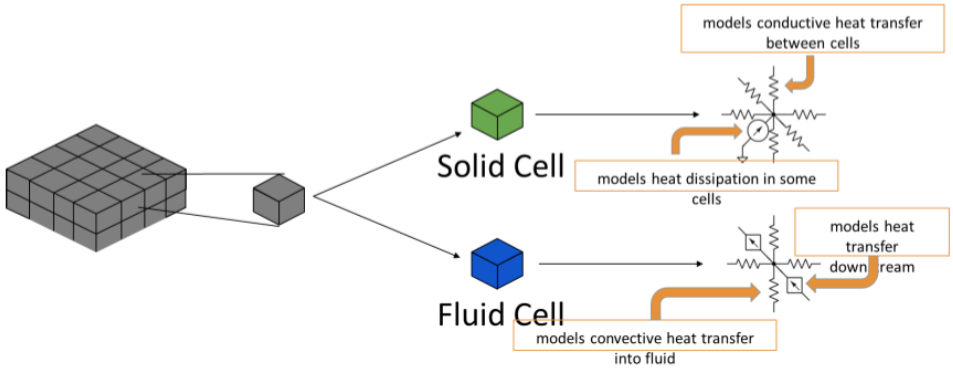
- HotSpot 1.0-6.0 Model
  - Major contribution: pre-RTL design space exploration
  - Allows dynamic power model in space and time
  - 2018 Influential ISCA Paper Award (for ISCA 2003 paper)





Grid Model

- HotSpot 7.0 Thermal Model
  - Both Thermal-electrical analogy and Fluid dynamics-electrical analogy
  - Also allows dynamic thermal model



Example2 run script

```
../../hotspot -c example.config -f ev6.flp -p gcc.ptrace
-materials_file example.materials -model_type grid -steady_file
outputs/gcc.steady -grid_steady_file outputs/gcc.grid.steady
```

- trade-off between speed and accuracy

-c <config_file>

-p <ptrace_file>

-flp <floorplan_file>

```
../../scripts/split_grid_steady.py outputs/gcc.grid.steady 4 64 64
../../scripts/grid_thermal_map.pl ev6.flp outputs/gcc_layer0.grid.steady > outputs/gcc.svg
../../scripts/grid_thermal_map.py ev6.flp outputs/gcc_layer0.grid.steady outputs/gcc.png
```

example2.config

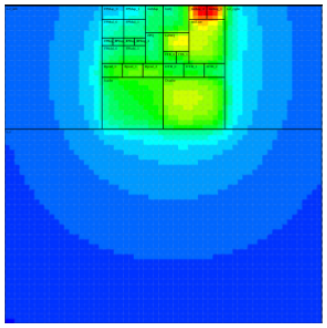
```
# grid model specific parameters
        # grid resolution - no. of rows
        -grid_rows                      64
        # grid resolution - no. of cols
        -grid_cols                      64
        # layer configuration from file
        -grid_layer_file        (null)
        # dump internal grid steady state temperatures
        -grid_steady_file       (null)
        # grid to block mapping mode - (avg|min|max|center)
        # i.e., a block's temperature is the avg, min or max
        # of all the grid cells in it or equal to that of
        # the grid cell in its center
        -grid_map_mode          avg
```

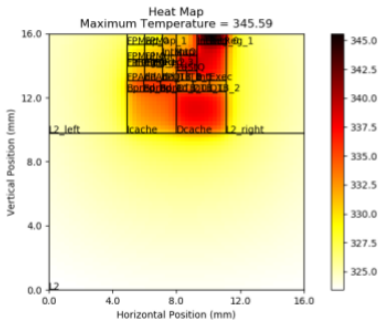
```
../../scripts/split_grid_steady.py outputs/gcc.grid.steady 4 64 64
../../scripts/grid_thermal_map.pl ev6.flp outputs/gcc_layer0.grid.steady > outputs/gcc.svg
../../scripts/grid_thermal_map.py ev6.flp outputs/gcc_layer0.grid.steady outputs/gcc.png
```


Example2 result

```
../../scripts/split_grid_steady.py outputs/gcc.grid.steady 4 64 64
../../scripts/grid_thermal_map.pl ev6.flp outputs/gcc_layer0.grid.steady > outputs/gcc.svg
../../scripts/grid_thermal_map.py ev6.flp outputs/gcc_layer0.grid.steady outputs/gcc.png
```

- two versions of this script
  - in Perl (.pl extension) and in Python (.py extension)
  - PL generates svg image, PY generates png image





Example3 run script

```
../../hotspot -c example.config -p example.ptrace -grid_layer_file
example.lcf -materials_file example.materials -model_type grid
-detailed_3D on -steady_file outputs/example.steady
-grid_steady_file outputs/example.grid.steady
```

-c <config_file>

-p <ptrace_file>

~~-flp <floorplan_file>~~ -grid_layer_file <layer config file>

-detailed_3D <on OR off>

: allow individual floorplan elements to list their own thermal properties.

example3.lcf

#File Format: #<Layer Number> #<Lateral heat flow Y/N?> #<Power Dissipation Y/N?> #<Specific heat capacity in J/(m^3K)> #<Resistivity in (m-K)/W> #<Thickness in m> #<floorplan file>	# Layer 0: 0 Y Y 1.75e6 0.01 0.00015 floorplan1.flp	# Layer 1:(TIM) 1 Y N 4e6 0.25 2.0e-05 floorplan1.flp	# Layer 2: 2 Y Y 1.75e6 0.01 0.00015 floorplan2.flp

example3.floor plan

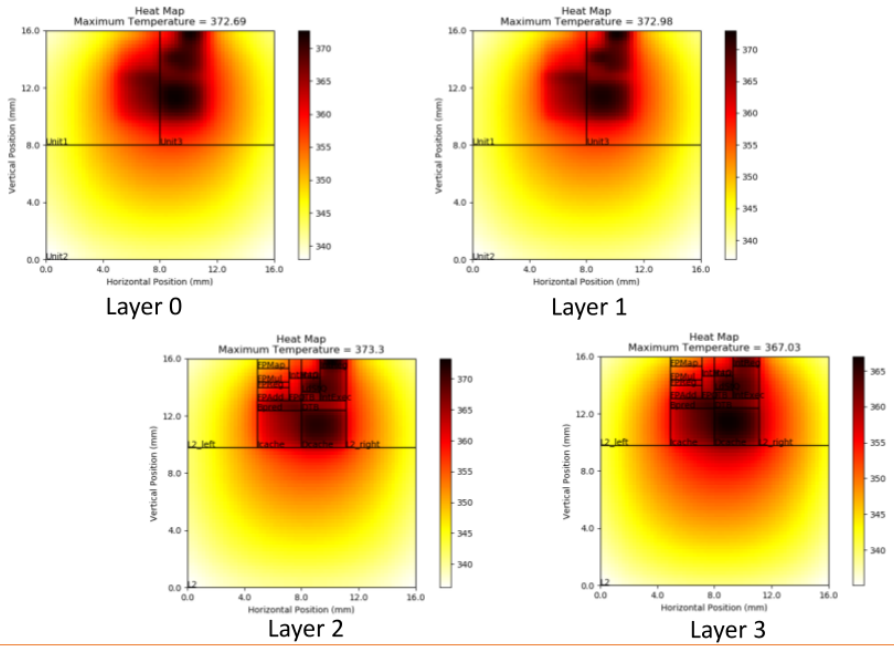
floorplan1.flp

Unit1	0.008000	0.008000	0.000000	0.008000
Unit2	0.016000	0.008000	0.000000	0.000000
Unit3	0.008000	0.008000	0.008000	0.008000

floorplan2.flp

L2_left	0.004900	0.006200	0.000000	0.009800
L2	0.016000	0.009800	0.000000	0.000000
L2_right	0.004900	0.006200	0.011100	0.009800
Icache	0.003100	0.002600	0.004900	0.009800
Dcache	0.003100	0.002600	0.008000	0.009800
Bpred	0.003100	0.000700	0.004900	0.012400
DTB	0.003100	0.000700	0.008000	0.012400
FPAdd	0.002200	0.000900	0.004900	0.013100
FPReg	0.002200	0.000380	0.004900	0.014000
FPMul	0.002200	0.000950	0.004900	0.014380
FPMap	0.002200	0.000670	0.004900	0.015330
IntMap	0.000900	0.001350	0.007100	0.014650
IntQ	0.001300	0.001350	0.008000	0.014650
IntReg	0.001800	0.000670	0.009300	0.015330
IntExec	0.001800	0.002230	0.009300	0.013100
FPQ	0.000900	0.001550	0.007100	0.013100
LdStQ	0.001300	0.000950	0.008000	0.013700
ITB	0.001300	0.000600	0.008000	0.013100

example3.results



Layer 0 Layer 1

Layer 2 Layer 3

example5.run

```
../../hotspot -c example.config -p example.ptrace -materials_file
example.materials -grid_layer_file example.lcf -model_type grid
-detailed_3D on -use_microchannels 1 -grid_steady_file
outputs/example.grid.steady -steady_file
outputs/example.steady
```

-c <config_file>

-p <ptrace_file>

~~-flp <floorplan_file>~~ -grid_layer_file <layer config file>

-detailed_3D <on>

-use_microchannels <0 or 1>

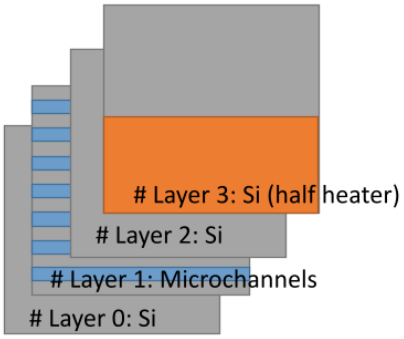
example5.config

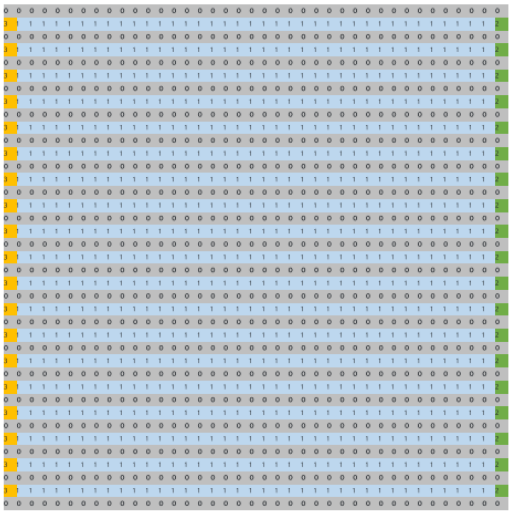
```
# microfluidic cooling parameters
  # using microfluidic cooling?
  -use_microfluidic_cooling 0

  # pumping pressure (Pa)
  -pumping_pressure       52000

  # Pump Internal Resistance (Pa-S/m^3)
  -pump_internal_res 0

  # temperature of coolant at inlet (K)
  -inlet_temperature     298.15

  # coolant material
  -coolant_material water

  # volumetric heat capacity of coolant (J/m^3-K) (overrided by coolant material)
  # -coolant_capac  4172638

  # resistivity of coolant (m-K/W) (overrided by coolant material)
  # -coolant_res   1.647717911

  # dynamic viscosity of coolant (Pa-s) (overrided by coolant material)
  # -coolant_visc  8.89e-4

  # channel wall material
  -wall_material silicon

  # volumetric heat capacity of channel walls (J/m^3-K) (overrided by wall material)
  # -wall_capac   1630300

  # resistivity of channel walls (m-K/W) (overrided by wall material)
  # -wall_res     0.0076923077

  # heat transfer coefficient (W/m^2-K)
   -htc   27132
```

example5.material

# material name # material type (solid or fluid) # thermal conductivity in W/(m-K) # volumetric heat capacity in J/(m^3-K) # dynamic viscosity in Pa-s (fluid only)	silicon solid 130.0 1630300	aluminum solid 237.0 2.422e6	water fluid 0.6069 4172638 8.89e-4

example5.flp



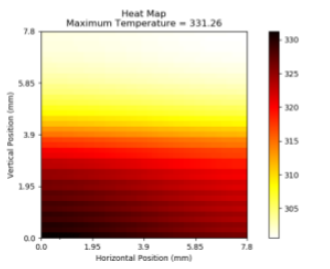


Layer 1: Microchannels horizontal.csv

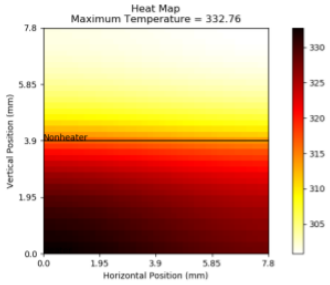
example5.lcf

#File Format: #<Layer Number> #<Lateral heat flow Y/N?> #<Power Dissipation Y/N?> #<Specific heat capacity in J/(m^3K)> #<Resistivity in (m-K)/W> #<Thickness in m> #<floorplan file>	# Layer 0: Si 0 Y N silicon 100e-6 floorplans/interposer.flp	# Layer 1 : fluid 1 Y N silicon 100e-6 microchannel_geometries/horizontal.csv	#Layer 2: Si 2 Y N silicon 90e-6 floorplans/silicon.flp	# Layer 3: Si 3 Y Y 1630300 0.0076923 10e-6 floorplans/heater.flp

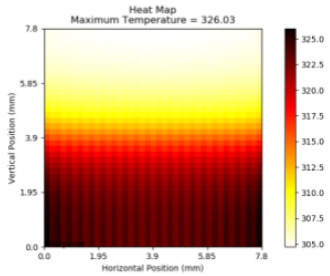
example5.results

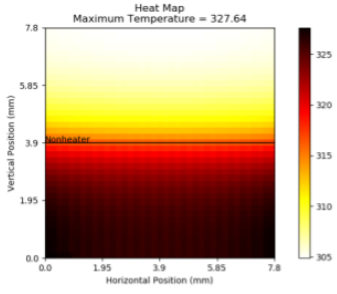


Layer 0



Layer 3





Appendix C. ArchFP Tutorial

This tutorial is based on the HOWTO of ArchFP released on Github.

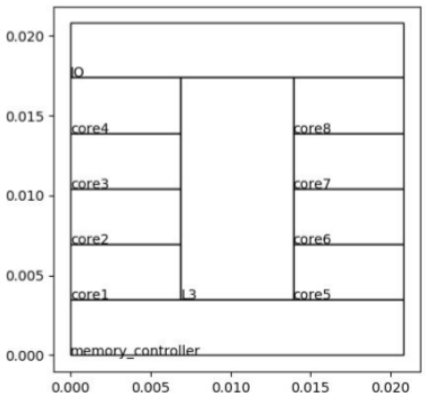
(https://github.com/uvahotspot/ArchFP)

1. Download ArchFP, read Howto.txt
2. Tar -xvf ArchFP into your work folder
3. Edit Main.cc - chip is the top level and mandatory. You can make many subunits.
4. Recompile → make at your work folder
5. Run → ./ArchFP
6. Draw → ../visualize_floorplan.py Sample.txt sample.png

Arch FP

UNIVERSITY of VIRGINIA

```
memory_controller 0.0208  0.00348  0  0

core1             0.0069  0.00348  0  0.00348
core2             0.0069  0.00348  0  0.00696
core3             0.0069  0.00348  0  0.01044
core4             0.0069  0.00348  0  0.01392

L3                0.007   0.01392  0.0069  0.00348

core5             0.0069  0.00348  0.0139  0.00348
core6             0.0069  0.00348  0.0139  0.00696
core7             0.0069  0.00348  0.0139  0.01044
core8             0.0069  0.00348  0.0139  0.01392

IO                0.0208  0.0034   0       0.0174
```



Text design of floor plan file.

Use script visualize_floorplan.py

CRISP hplp

28

Arch FP

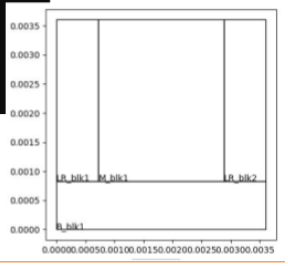
- Change the code and recompile it
- Change main.cc

```
int main(int argc, char* argv[])
{
   geogLayout * Unit = new geogLayout();
   // Type, Count, Area, MaxAspectRatio, MinAspectRatio, Hint
   Unit->addComponentCluster("L_blk", 1, 6, 50., 1., Left);
   Unit->addComponentCluster("R_blk", 1, 2, 50., 1., Right);
   ...
}
```


Arch FP

- example1

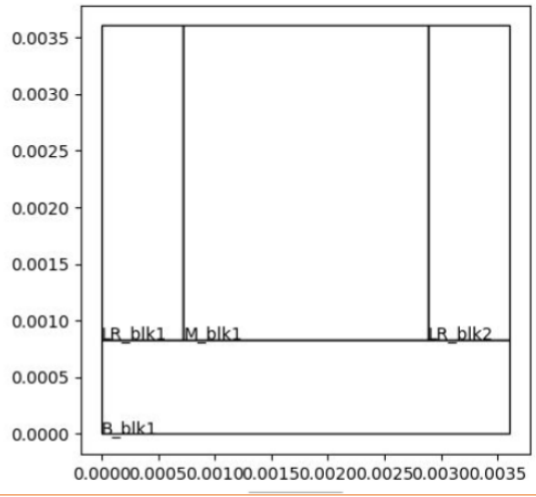
```
int main(int argc, char* argv[])
{
   geogLayout * Unit = new geogLayout();
   Unit->addComponentCluster("LR_blk", 2, 2, 50., 1., LeftRight);
   Unit->addComponentCluster("M_blk", 1, 6, 50., 1., Center);

   geogLayout * chip = new geogLayout();
   chip->addComponent(Unit, 1, Top);
   chip->addComponentCluster("B_blk", 1, 3, 50., 1., Bottom);
}
```



Arch FP

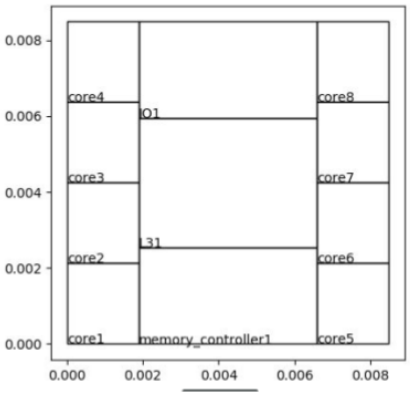
- example1

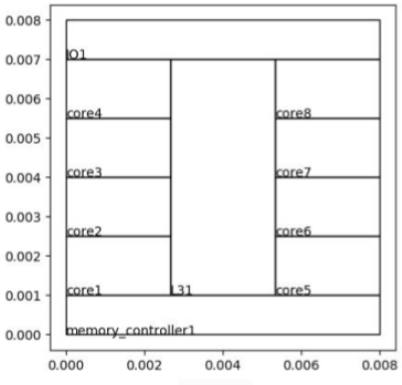
```
int main(int argc, char* argv[])
{
   geogLayout * Unit = new geogLayout();
   Unit->addComponentCluster("LR_blk", 2, 2, 50., 1., LeftRight);
   Unit->addComponentCluster("M_blk", 1, 6, 50., 1., Center);

   geogLayout * chip = new geogLayout();
   chip->addComponent(Unit, 1, Top);
   chip->addComponentCluster("B_blk", 1, 3, 50., 1., Bottom);
}
```



Arch FP

- Hierarchy

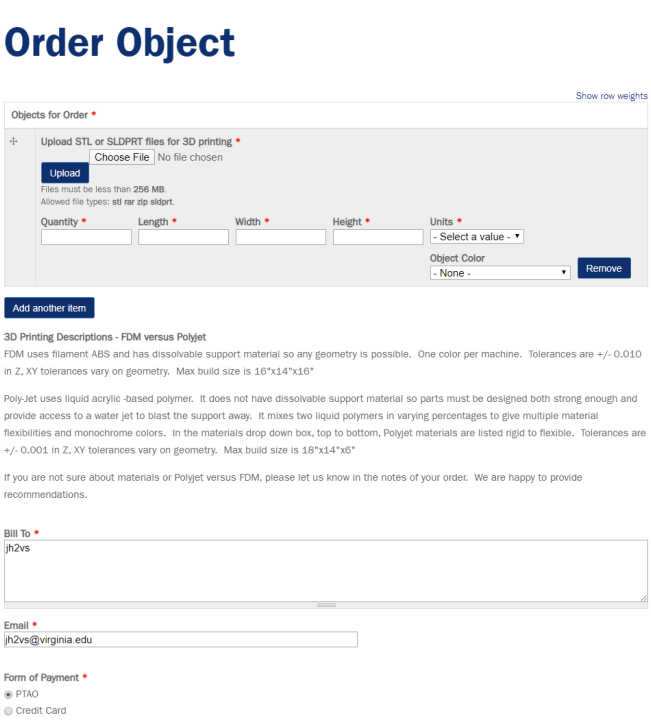




Appendix D. HotSpot 7.0 Tutorial

Before Start: order 3D printing

- We can order 3d printings to the prototype lab using FDM or polyjet.
- https://rpl.mae.virginia.edu/rapid-prototyping-lab
- Difference between FDM and polyjet
    - Polyjet requires support, but more precise
    - http://engatech.com/difference-fdm-polyjet-3d-printing/
- How to order
    - https://rpl.mae.virginia.edu/order-form
    - The design should use inch metrics.



Before Start: SolidWorks

You can download solid works through UVA Software Gateway

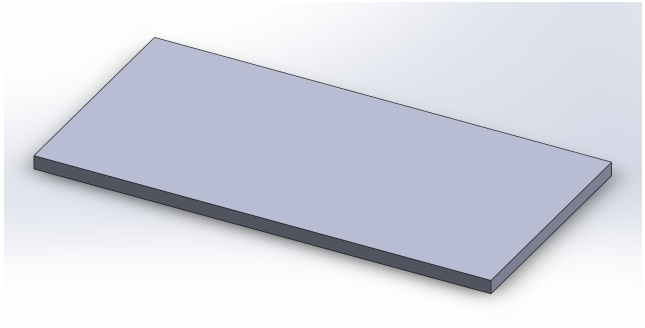
- Change default unit system in SolidWorks to IPS for ‘inch-feet-oz’
- https://grabcad.com/tutorials/how-to-change-default-unit-system-in-solidworks

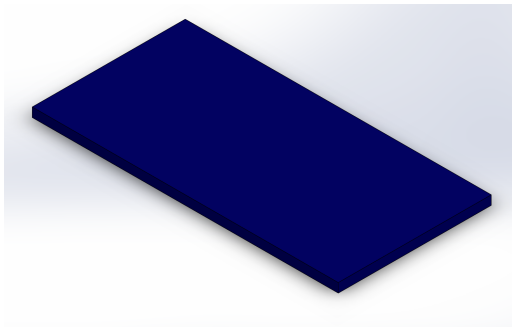
- We should stick to the IPS metrics for the 3D printing order.

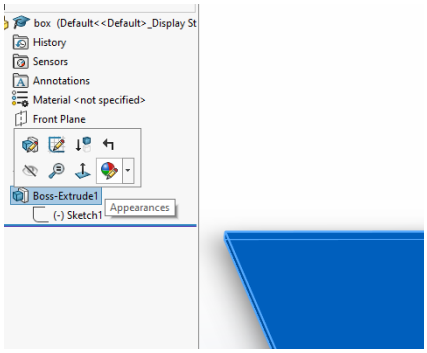
Design a jig

1. Design dummy PCB < 2 x 1 x 0.06 inch >

- Create a simple box
- http://www.solidworkstutorials.com/how-to-create-simple-box/
- Change the color using ‘Appearance’

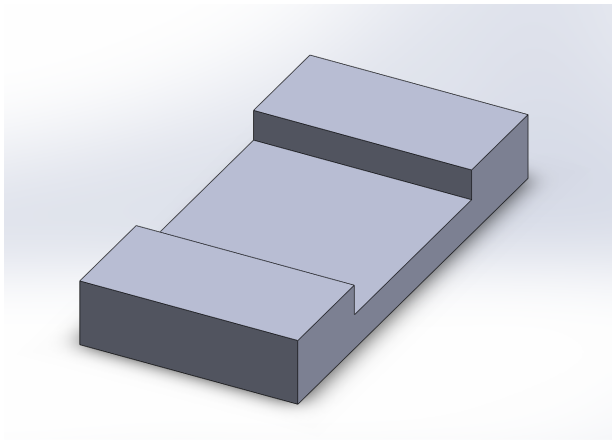


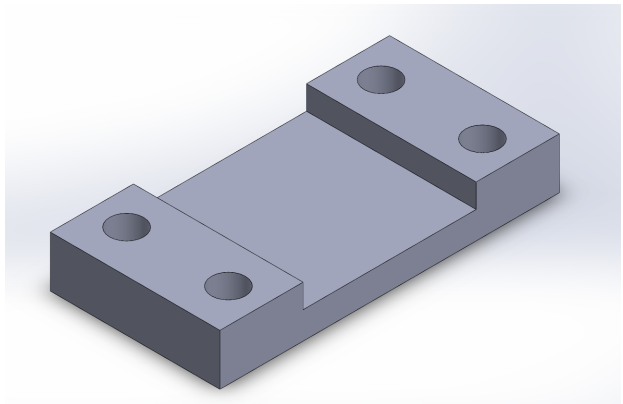




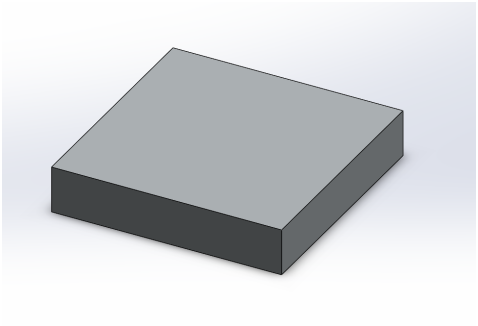
2. Make a base with cut-Extrude

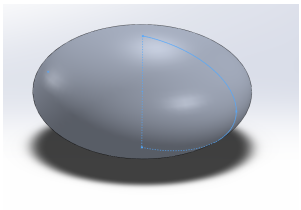
- http://www.solidworkstutorials.com/how-to-create-simple-plate/
- Create a simple box <1x2x 0.3 inch> (0.08 PDMS + 0.06 PCB + 0.16 Base)
- Cut a rectangle with half depth
- Cut four circles with full depth



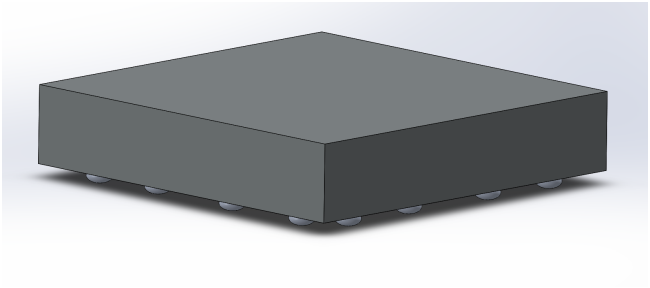


3. Design a dummy chip with bumps
- Create simple box 0.1 x 0.1 x 0.02 box, change appearance to metal
- Create a flat sphere using 0.0018 Resolve.

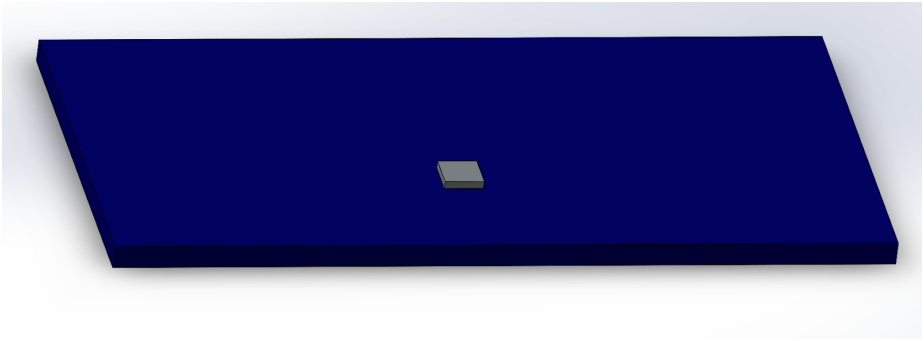




4. Assemble chip
- Toolbar - New - Assembly
- We can assemble parts and other assemblies
- Add instants, add a chip first.
- Add a bump. We can copy it shift + drag

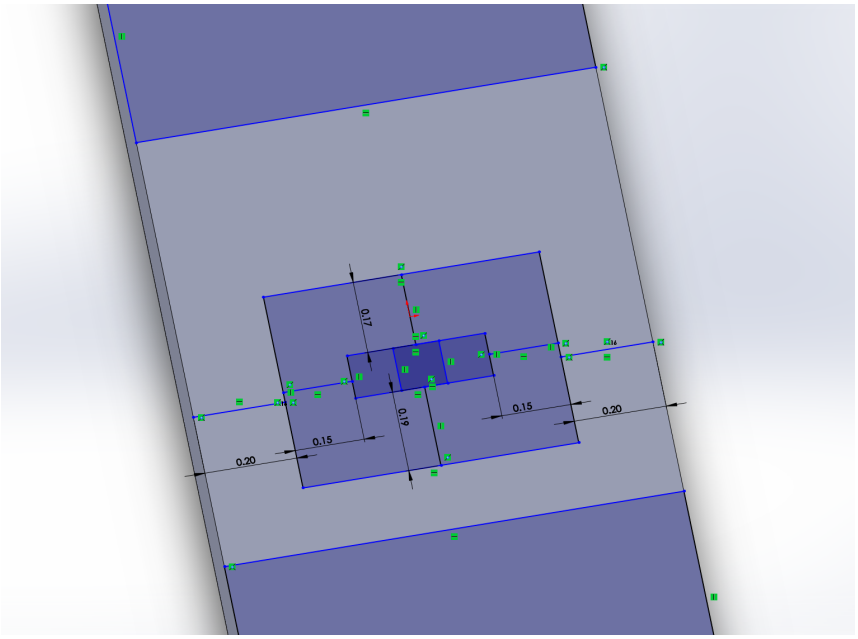


5. Assemble pcb
- Add pcb first, and then add the assemble chip

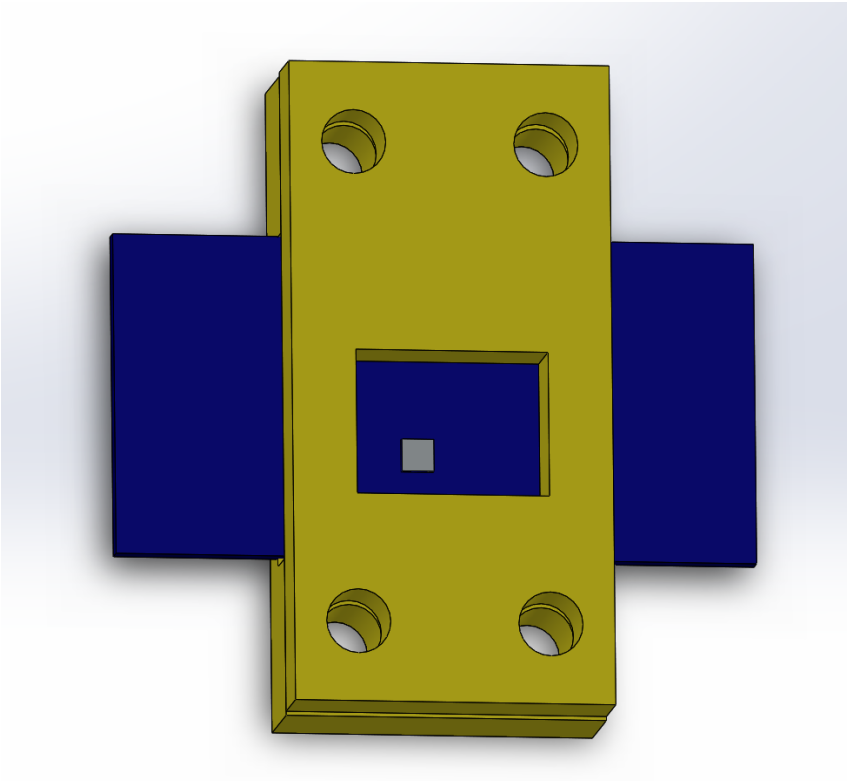


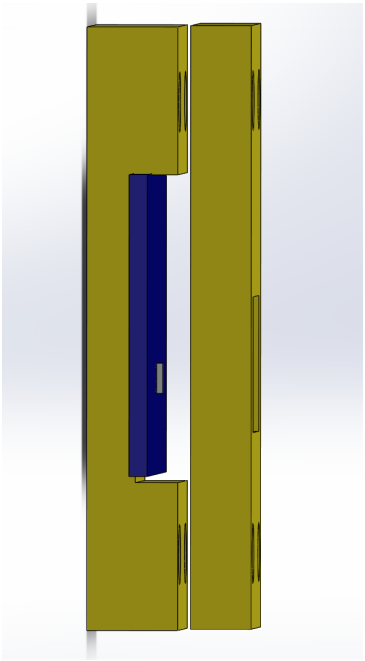
6. Design fluid jacket

7. Design cover



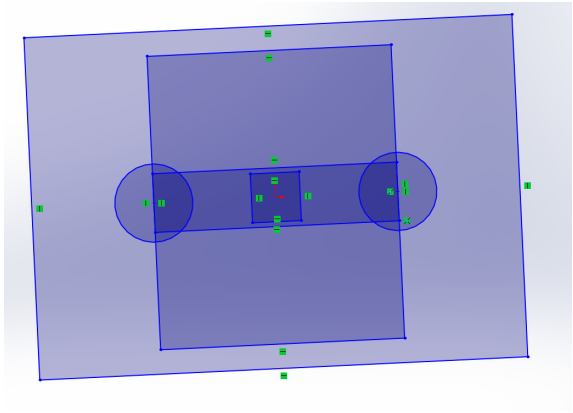
8. Assembly

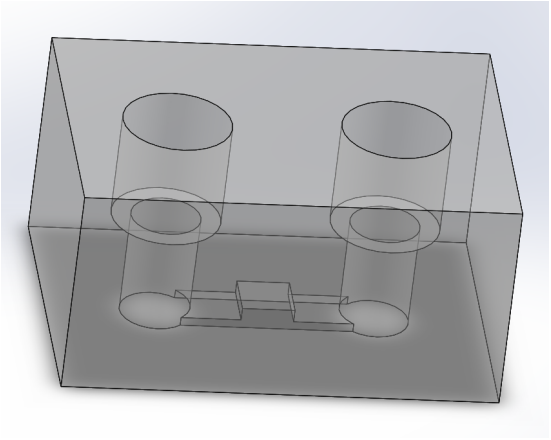




Design a fluid jacket

1. Draw the sketch with pcb dimensions





Sketch and the 3D model
0.1 x 0.1 inch square chip, 0.12 x 0.12 chip box
0.12-inch wide channel
0.16-inch low diameter, 0.25-inch upper diameter for insert

2. Start with a simple box. <1 x 0.7 x 0.5 inch box>

3. Chip box Extruded cut <0.12 x 0.12 x 0.05>

4. Channel Extruded cut <0.5 x 0.12 x 0.02>

5. 0.16-inch diameter low hole Extruded cut

6. 0.25-inch diameter upper hole Extruded cut

7. Assembly