Advances in Direct-to-Chip Liquid Cooling Integration

L.W. Mirkarimi	R. Zhang	G.G. Fountain	K.M Bang
Adeia	*Adeia*	*Adeia*	*Adeia*
San Jose, CA-USA	San Jose, CA- USA	Morrisville, NC-USA	San Jose, CA- USA
laura.mirkarimi@adeia.com	ron.zhang@adeia.com	gill.fountain@adeia.com	km.bang@adeia.com
S. Sadiq	A.Avellan Jaramillo	B. Lee	H. Katske
Adeia	*Adeia*	*Adeia*	*Adeia*
San Jose, CA-USA	San Jose, CA- USA	San Jose, CA-USA	San Jose, CA-USA
suhail.sadiq@adeia.com	arianna.avellan@adeia.com	bongsub.lee@adeia.com	helen.katske@adeia.com

Abstract— The rapid rise of power densities in high performance computing (HPC) modules is expected to create thermal challenges at the datacenter, where the incumbent technology cannot cool the module to the required temperature. We report a new approach to silicon microchannel direct-to-chip cooling and share the design performance map that is possible. Thermal system considerations of cooling performance, pressure drop and power usage effectiveness (PUE) tradeoffs are discussed for four different cooling designs. These solutions provide design guidelines for the non-uniform heatmaps we anticipate in real world processor module cooling. We share our approach to thermal design which will address the future needs of the HPC industry. Finally, we discuss the process flow integration concepts of how our modular, custom, integrated cooling solution can be manufactured in today's advanced packaging supply chain.

Keywords— liquid cooling, silicon microchannel, direct bond cold plate, advanced package module, package assembly, thermal resistance, cold plate, manifold, pressure drop,

I. Introduction

Thermal challenges for high performance computing modules are among the most significant issues for the advanced packaging industry to address over the next five years. As the power densities in processor chips rise from 0.8 to 3W/mm^2 in the next 3 years and the possibility of hot spots reaching 5W/mm^2 in 6 years, the industry is seeking an alternative cooling solution. Silicon microchannel cooling eliminates thermal interface material (TIM) thereby reducing the thermal resistance and increasing efficiency of heat removal. The original concept was published as early as 1981 [1] and is now being revisited in the industry [2, 3]. However, integration concerns with large modules, die etching and liquid leakage from the package have created skepticism toward silicon microchannel cooling.

Taking a different approach, we fabricate the cold plate in a separate Si wafer and create channels in the bottom surface. The cold plate is direct bonded to the backside of the integrated circuit chip, creating face-down, closed channels, which reduce the chance of leakage and simplify the overall integration and test process [4]. The cold plate is direct bonded to silicon, following previously reported methods to form strong dielectric-dielectric and hybrid bonds so that the heat transfers effortlessly across the bond line [5]. This differs from other approaches where the microchannels are facing upward away from the integrated circuit chip [2,3].

II. Thermal Performance of Integrated Cooling Solutions

We have previously reported that our thermally optimized design achieves 15% lower thermal resistance [4] than the face-up microchannel that Tuckerman and Pease [1] proposed, offering significant energy efficiency. We demonstrated that the thermal resistance of our silicon microchannel cold plates bonded to a resistive heater chip has a negligible thermal resistance. By carefully measuring the thermal resistance of the components and comparing it to our CFD simulation, we found excellent agreement (~3%) and no need to add an interfacial thermal resistance to the models. Additionally, we have reported experimental comparisons for our integrated cooling solution, RapidCool^{TM}, which cooled to 3x higher power densities (~2 W/mm^2) compared to the typical 0.3-0.5 W/mm^2 power densities with phase change thermal interface materials and off-the-shelf high efficiency cold plates, all at very low flow rates of about 0.2-.3GPM. Flow rates in that study were kept low to maintain a low-pressure drop of 4psi and compare the thermal performance of the TIM-based cold plate designs [2].

More recently, we performed thermal resistance measurements for new cooling system designs. The system includes a 26mm x 32mm resistive heater chip which is bonded to a cold plate and an integrated manifold that is used to deliver the coolant to the chip. The complimentary manifolds were fabricated with a 3D printer using a resin material compatible with the coolant. In Fig.1, thermal resistance is shown as a

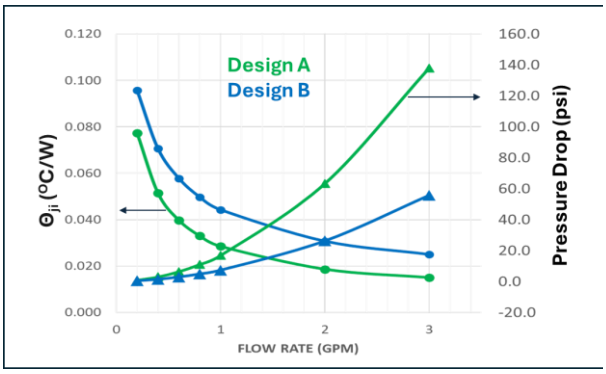


Fig. 1 Thermal resistance vs flow rate for two integrated cooling solution designs.

function of flow rate for two system designs. The measured thermal performance of the chip packages consistently showed a 3-5% variance with our computational fluid dynamics (CFD) simulations over a wide range of power densities and flow rates.

Leveraging our precision fabrication and accurate computational fluid dynamics simulation approach, we mapped out the design/performance for high power densities, and necessary coolant flow rates to maintain a temperature less than 100°C (Fig. 2). The data will be used to understand the scalability of this cooling technology to address higher power densities and the dependency on the cold plate design.

It is imperative to understand the impact of system design constraints on pressure drop and thermal resistance, as it is key to creating a compatible cooling solution for the existing datacenter infrastructure. For example, the pressure drop profile for Design A increases rapidly with flow rate, while Designs B-D are less sensitive and more appealing from an overall system performance perspective. Fig. 2 shows that uniform power densities of 3-5 W/mm^2 applied across a 26mm x 32mm chip with an embedded resistive heater can be cooled and maintain a pressure drop below ~60psi. Design A shows that we can cool up to 6W/mm^2; however, with this design, the pressure drop increases to ~130 psi at a coolant flow rate of 3GPM, which is generally not compatible with datacenter infrastructure. In comparison, Designs B-D easily cool uniform power densities up to 3-4W/mm^2 while maintaining a pressure drop ranging from 10-58psi. The thermal performance results, resistance and pressure drop, provide guidance for selection of design features for our vision of a custom cooling system based on IC heatmap and application system constraints.

The results shared in Fig. 2 are for uniform power density chips, and the design features in the cooling solutions can be mixed and matched for non-uniform heatmaps to optimize the performance. The cold plate design modifications represented here are focused on the channel shape and layout with a complementary integrated manifold for each design. Another approach to optimize thermal performance in non-uniform power density chips is flow segmentation in the cold

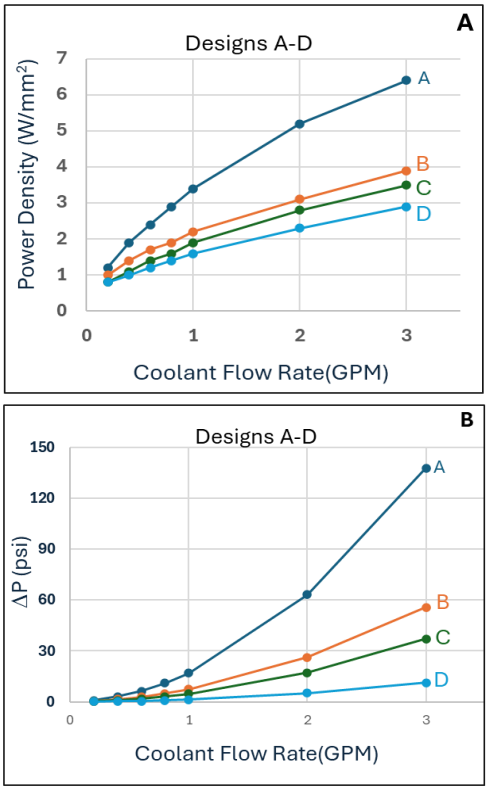


Fig. 2 Design features in four different integrated cooling solution designs. A) Power density across chip vs coolant flow rate to maintain a temperature <100C. B) The pressure-drop in the system, as a function of coolant flow rates.

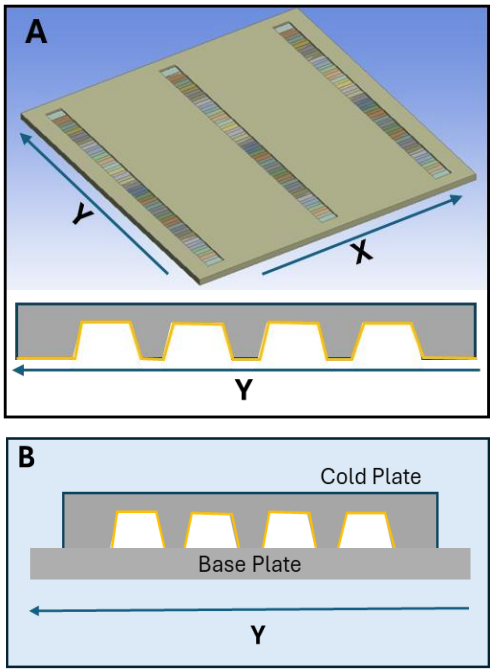


Fig. 3 Schematic images of A) die level silicon microchannel cold plate with one inlet and two outlets, with trapezoid shaped channels, B) cross section of 3A attached to a base plate.

plate with the integrated manifold, which is not discussed in this study [6].

III. Process Development, Integration and Assembly

A. Silicon Cold Plates

The silicon cold plate used in this study has channels facing downward toward the chip (Fig.3a). A low-cost, streamlined wet etch process was developed to create the channels on the bottom side and the coolant inlets on the top side. The inlet/outlet openings were lithographically defined on the top side of the silicon wafer, followed by the coolant channel definition on the bottom side of the silicon wafer. Both sides were etched simultaneously to create the cold plate geometry. We can control the shape of the Si channel depending upon the chosen etch process so that the side wall angle is as high as 90 degrees with a dry etch process to shallow angles such as 54 degrees with the wet etch process.

The upper portion of the plate contains the inlet and outlet openings whose bottom portions connect to the top of the microchannels below. Fig. 3 shows a cold plate with three openings, where one center inlet and two outlets are easily coupled to the integrated manifold which will have a single inlet and two outlets coupled to the server rack coolant lines.

The cold plate wafer can be bonded to a base plate wafer as shown in Fig. 3b, which allows for independent pressure testing of the fixture prior to attaching to an integrated circuit die. The base plate is an optional feature which may be helpful for integration with the supply chain.

B. Advanced Package Assembly Process Flow

The likelihood and speed of adoption of a technology depend on the magnitude of the problem addressed, the manufacturability and alignment with the supply chain. The assembly of integrated cooling solution technology described above falls into the advanced packaging supply chain in the semiconductor sector. The advanced packaging supply chain is known for the multitude of technologies which have evolved over time, each meeting the specific requirements of cost and performance for the target application. The assembly integration concepts that will be needed to enable this technology are discussed below.

The application is a high-performance compute module, so the CoWoS package is used as an example process flow [7]. The fabrication and assembly process that we developed is aligned with a 2.5D advanced package flow for HPC modules. Our extensive process knowledge for direct and hybrid bonding of die to wafer and die to die stacking gave us a head start at considering integration of this cooling technology for the final product modules [2,8]. Fig. 4 shows the envisioned assembly process flow to dice the cold plate for attachment to a HBM memory module. The first step is to dice the cold plate or the cold plate-base plate wafer pair into pieces for bonding to the HBM stacked die.

The dicing process is critical. The bonding surface of the cold plate is covered with a protective layer which is

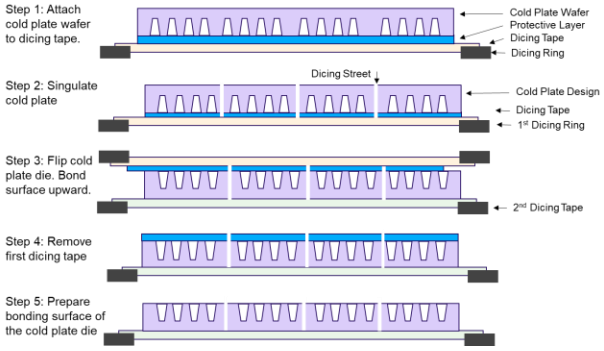


Fig. 4 Dicing process flow for silicon microchannel cold plate preparation in advanced packaging.

attached to a standard dicing tape film placed on a tape frame [8]. Next, the cold plate is diced using either a mechanical saw or stealth laser technique. The diced wafer is flipped onto the second film frame, and the first dicing film is removed. Now the bond surface is facing up. This ensures that the die surface can be properly prepared for a direct bond - a molecular layer bond - between the backside of the IC processor and the channel side of the cold plate.

The surface of the cold plate die must now be prepared for direct/hybrid bonding. The surface preparation sequence in Fig. 5 includes die cleaning with a combination of wet and dry processes (Step 1), plasma activation (Step 2), followed by a hydration termination process sequence. Standard high-volume equipment exists today for 300mm wafer dicing, wafer and die on film frame (>400mm) cleaning and plasma activation.

The film frame with the clean cold plate die, is transferred to the die bonder. There are two package assembly flows in which these cold plate die will be integrated. The first is the HBM DRAM die stack where the DRAM die stacks range from 8 to 16 high die stacks. The cold plate die is attached to the top die in the HBM die stack. Currently HBM die stacks are bonded to a logic die, then the entire wafer is molded and processed to ensure that the top die is exposed for thermal solution treatment. Using this integrated cooling technology, the top die would be the cold plate die not the HBM die or dummy die (Fig. 6).

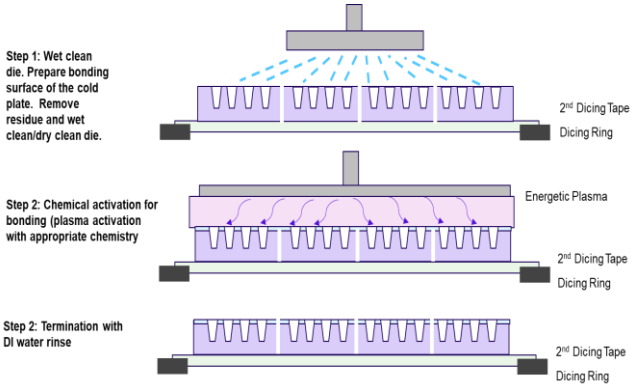
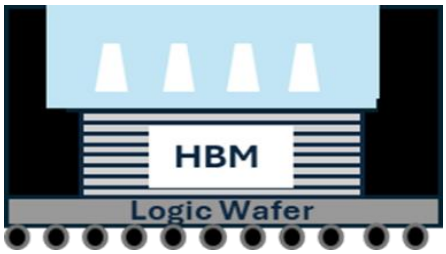


Fig. 5 Cold plate die surface preparation schematic for bonding.

Fig. 6 Schematic of HBM-package with integrated cold plate with base plate. No manifold is shown.



Today most 2.5D modules are silicon interposer based and that assembly flow is selected here as an example in Fig. 7. In Step 1, at least one logic die and one or more HBM DRAM packages are attached to the interposer. While flip chip attachment is shown, hybrid bonding can alternatively be used. The logic/processor die will have been prepared in a similar assembly flow for the HBM packages shown in Fig. 5. The primary difference is that some companies will have a single thinned logic die, as shown in Fig. 7, while others may have chiplet architectures with multiple chiplet die serving portions of the processor/logic function. All variants (logic die, logic chiplet assembly, and HBM packages) will have either solder micro-bumps or hybrid bond interconnects on the bottom processor die with a cold plate attached to the upper die. If flip chip interconnect is used as shown, then underfill is completed to secure the flip chip packages to the interposer (Fig. 7- Step 2). Wafer level molding is performed so that upper surfaces of the silicon cold plates are exposed at the top surface of the molding compound, permitting later attachment of a manifold to the inlet/outlet openings at the top of the cold plate (Fig. 7- Step 3). Once the molding process is complete, the carrier wafer is removed, and the module is diced and attached to a standard substrate material used to form the BGA package that is ultimately attached to the board in a server rack (Fig. 7-Step 4).

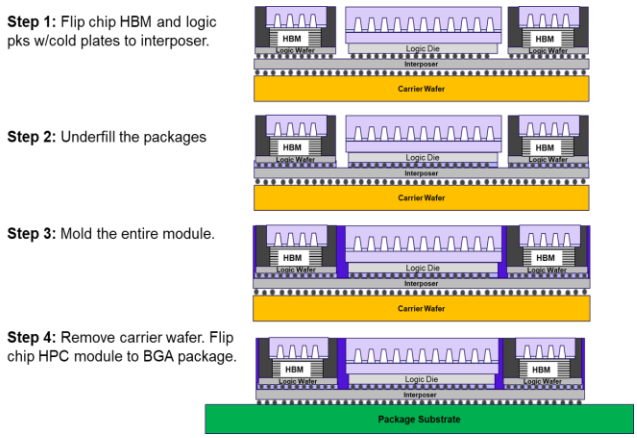


Fig. 7 Schematic HPC module assembly showing package attach, underfill, mold and attach to substrate package.

The integrated manifold is then attached to the package using an adhesive. The coolant fixtures supplied to the server rack are seamlessly coupled to the manifold of our direct-to-chip liquid cooling solution. We have tested our cold plates and manifold for a 2x reticle package up to 100 psi pressure of coolant for various times for >2 days at ~50psi, half day at 70psi and > an hour at 100psi with no leaking. There is sufficient compliance in this package design to tolerate the warpage across the 2x reticle size for the manifold sealing. We anticipate scalability to an overall 4-5x package size and associated warpage can be addressed with the manifold attachment process given our early experimental results.

IV. Discussion

The strength of this new microchannel direct-to-chip integrated cooling solution is the ability to create a custom design that addresses hot spots in the modules by optimizing the physical design of the cold plate and the complimentary integrated manifold. Our cold plate and integrated manifold have demonstrated a lower thermal resistance (~15%). However, the pressure drop performance of the legacy microchannel cold plate has proven too high for practical implementation. Let us explore the pressure drop in the designs shared in this study. The cold plate must perform within an acceptable coolant flow rate and pressure drop range that is consistent with the datacenter facility specifications. The pressure drop in various designs discussed in Fig. 2 ranged from 10-50psi for most designs cooling uniform power densities as high as 3-4.7 W/mm^2, which is about 3-4.7 times higher than what can be done with TIM based cold plates today.

The pressure drop, coolant flow rates and power density, are all important parameters used to calculate the pumping power and the power usage effectiveness (PUE) for the four designs discussed earlier in Fig. 2. The relevant

equations are shown below for thermal resistance (Θ_{ji}), pressure drop (ΔP), pump power (P_{pump}) and PUE. Where T_{jmax} is the maximum temperature, T_i is inlet temperature. In Eq. 3, $\dot{Q}$ is the flow rate/pump efficiency and in Eq. 4, P_{IT} is the infrastructure power.

$$\Theta_{ji} = \frac{Tjmax - Ti}{Power} \quad (1)$$

$$\Delta P = P_{in} - P_{out} \quad (2)$$

$$P_{pump} = \dot{Q} * \Delta P \quad (3)$$

$$\mathbf{PUE = (P_{IT} + P_{pump})/P_{IT}} \quad (4)$$

The measured results from the designs explored in this study were applied to the above equations. IC chips with a 3-4 W/mm^2 power density are cooled with a PUE ranging from 1.007 to 1.026 for all designs which is a highly efficient PUE (Table 1). Design A, with the lowest thermal resistance, allows processor chips to be cooled efficiently up to 4.7 W/mm^2 with a PUE of only 1.025. These excellent PUEs are quite promising for energy savings at the data center given that a typical PUE for air cooled data centers is 1.5-1.58. The results imply that the integrated cooling solutions discussed here, offer higher efficiencies at the datacenter even though the power densities of the chips are significantly higher than those used today with an average power density of 0.8-1W/mm2.

TABLE 1: Power Usage Effectiveness

	Power Usage Efficiency			
Flow Rate (GPM)	Design A	Design B	Design C	Design D
1.6	1.025	1.026	1.02	1.007
2	-	1.033	1.021	1.008
3	-	1.104	1.07	1.02

Our validated fabrication, modelling and simulation protocol can enable custom thermal design optimization for the anticipated higher power density computing modules on the semiconductor roadmap. As the CoWoS packages grow to 4-5x reticle size in the future, the corresponding warpage in both silicon interposer and in silicon bridge die packages will grow [9]. This challenge will be extremely difficult for the incumbent TIM-based cold plate to meet the thermal performance requirements due to the physical constraints of the packages. Our modular approach to integrate the cold plate assembly eliminates the overall package influence of warpage on the thermal performance. This solves the root cause of many performance challenges with the incumbent TIM-based technology.

V. Summary and Next Steps

A drop-in replacement for TIM-based cold plates compatible with the infrastructure that is scalable to high power densities was proposed. The influence of the four different cooling solution designs on the thermal performance was shared. We outlined how the assembly of our integrated cooling solution, could be integrated into current supply chain for HPC modules, including collaboration between the IDMs, wafer foundries, and OSATs. Our wet etch cold plate fabrication process is well suited for high-volume, low-cost manufacturing. However, the manufacturing ramp may take some time as there will be a need to adjust some process modules on the specific equipment sets in each segment of the supply chain.

These results suggest a path to cool HPC modules for several generations with the proof point of 3-4.7 W/mm^2 with acceptably low PUEs (1.007-1.026), which address the challenges in the semiconductor roadmap in the next 5-7 years. Nevertheless, we see pathways to accommodate even higher power density chips while maintaining low PUE. We combine features of Design A with some of our new concepts that allow us to reduce the impact of the pressure drop while achieving the lower thermal resistance to go to 6W/mm^2 with low PUE. Finally, we are studying non-uniform heatmap solutions that mimic the HPC module roadmap and anticipate combinations of the designs reported here, flow segmentation [6] and our hot spot management solutions, which will provide a highly flexible, thermal solution platform for several generations.

References

[1] D.B. Tuckerman and R.F.W. Pease, IEEE Electronic Device Letters, EDL-2, No. 5, May 1981.

[2] Yu-Jen Lien et al., "Direct-to-Silicon Liquid Cooling Integrated on CoWoS® Platform," 2025, IEEE 75th Electronic Components and Technology Conference (ECTC), Dallas, TX, USA,)

[3] "Cooling apparatus, semiconductor device including the apparatus, and manufacturing method therof", US20230298969A.

[4] R. Zhang et al, Revolutionary Thermal Solutions for Hot Chips, 2025 ITherm Proceedings Dallas TX, USA 2025.

[5] G. Gao et al., "Development of Low Temperature Direct Bond Interconnect Technology for Die-To-Wafer and Die-To-Die Applications-Stacking, Yield Improvement, Reliability Assessment," 2018 International Wafer Level Packaging Conference (IWLPC), San Jose, CA, USA, 2018, pp. 1-7.

[6] Zhang et al., "Design and Experimental Validation of Partitioned Flow Strategies for Optimized Microchannel Cooling", ITherm 2026- "in press".

[7] Chip on Wafer on Substrate, CoWoS: https://3dfabric.tsmc.com/english/dedicatedFoundry/technology/cowos.htm

[8] L. Mirkarimi et al., "Fine Pitch Die-to-Wafer Hybrid Bonding," 2023 IEEE 73rdElectronic Components and Technology Conference (ECTC), Orlando, FL, USA, 2023, pp. 83-90.

[9] Hu, et. al., "CoWoS Architecture Evolution for Next Generation HPC on 2.5D System in Package", ECTC 2023.