Architecting the Thermal Dissipation and Power Delivery for Chiplet-Based Wafer-Scale Systems and Experimental Demonstration of Segmented Vertical Cooling for Silicon Dielets at 1 W/mm^2 Power Density

Haoxiang Ren[1]*, Ben Yang[1]*, Naarendharan Meenakshi Sundaram[1]*, Dongkai Shangguan[2], Timothy S. Fisher[1], and Subramanian S. Iyer[1]
[1]UCLA Center for Heterogeneous Integration and Performance Scaling (CHIPS), Los Angeles, CA 90095
[2]Thermal Engineering Associates, Inc., Santa Clara, CA 95051 *Equal Contribution

***Abstract*—We have analyzed several options for thermal dissipation and power delivery of wafer-scale systems. A two-phase cooling unit was vertically integrated onto a silicon dielet. With a power density of 1 W/mm^2 to mimic the high-power compute dielets on wafer-scale systems, we demonstrate a vertical heat extraction system scalable up to 300-mm wafer diameters with the potential to go to higher power densities.**

Keywords—wafer-scale systems, two-phase cooling, silicon thermal dielet

I. Introduction

Wafer-scale engines and systems built on chiplet-based large substrates at fine pitch, promise heterogeneously integration at scale to realize the vision of "Scale-down and Scale-out" but heat dissipation and the converse problem of power delivery have been referred to as the "Achilles Heel" of advanced packaging in the National Advanced Packaging Manufacturing Program (NAPMP) vision paper [1]. A wafer-scale system built on an advanced packaging platform like Silicon-Interconnect Fabric (Si-IF) [2] may consume around 50-100 kW of power and dissipate a similar amount of heat assuming a power density of 1-2 W/mm^2. In contrast to the thermal management of chip-scale systems, dissipating such large amounts of heat through 1) lateral heat spreading is not feasible to implement and, 2) embedded micro-channel cooling is not practical due to high-pressure, plumbing complexity and reliability considerations. Thus, vertical heat extraction (Fig.1) is required.

With such a high heat density and no lateral heat spreading, a heat transfer coefficient of >15000 W/m^{2}K is necessary. The traditional air cooling and liquid cooling technologies have an upper limit on heat transfer coefficients of ~200 W/m^{2}K and ~10,000 W/m^{2}K [3]. Two-phase cooling is needed to achieve high heat transfer coefficients required in such advanced packaging systems. Flash cooling – a combination of pressure and temperature-driven boiling – can achieve high heat transfer coefficients [4], [5] and has been demonstrated as a solution for high heat flux cooling without lateral heat spreading [6], [7]. Additionally, a chiplet-based spatially segmented and a temporally dynamic thermal management that anticipates heat loads is needed to optimize the efficiency and utilization of the thermal dissipation unit, to account for (1) the presence of hot spots and other spatially non-uniform thermal profiles across the chiplets, and (2) the effects of per-chiplet dynamic voltage and frequency scaling (DVFS). DVFS is particularly important in heterogeneous chiplet systems where different chiplets may dissipate varying amounts of power. Thus, transient peaks of up to 3-6 W/mm^2 are expected. Effective cooling of such advanced packaging systems is challenging due to the high density of heat fluxes, reducing the viability of systems which rely on maximal heat spreading.

In this study, flash cooling with a methanol working fluid is used to cool 1 W/mm^2 produced by the thermal test chip (TTC). We analyze various power and thermal architectures for chiplet-based systems and experimentally demonstrate a case featuring vertical heat extraction. Section II provides a comprehensive overview of power delivery and thermal dissipation architectures for silicon substrate wafer-scale systems, outlining cooling requirements and emphasizing the importance of thermal interface materials (TIMs). Section III details the experimental setup as well as the implementation of two-phase cooling and various TIMs, including indium, to replicate the assembly structure of wafer-scale systems. Section IV presents the measurement results, which is followed by the conclusion in Section V.

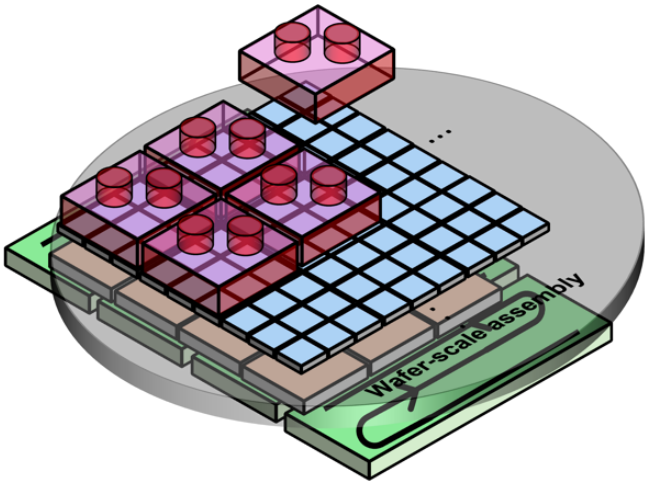

Fig. 1. 3D schematic view of one of the thermal management and power delivery architectures. Two-phase thermal dissipation units (TDUs) are placed at the dielet side to cool the high-power compute chiplets and forced water cooling units are assembled at the substrate side to cool the power modules. Drawing is not to scale.

2377-5726/25/$31.00 ©2025 IEEE
DOI 10.1109/ECTC51687.2025.00258

II. Thermal Dissipation and Power Delivery Architectures for Chiplet-Based Wafer-Scale Systems

Thermal dissipation and power delivery exist hand-in-hand in advanced packaging and wafer-scale systems, necessitating a unified approach when defining system architecture. The architectural discussion begins with power distribution along the z-axis in chiplet-based wafer-scale systems. In chiplet-based scale-out systems, power delivery presents unique challenges compared to traditional monolithic integration. In monolithic designs, power I/Os are directly accessible and can be easily connected to power modules. However, in chiplet-based architectures, chiplets are flip-chip bonded onto a thick substrate, requiring alternative power delivery pathways. Power can either be supplied from the substrate side via through-wafer vias (TWVs) or from the dielet side. When power is supplied from the substrate side, the power delivery network (PDN) spans both the substrate and dielet sides. Conversely, if power is delivered from the dielet side, it is assumed that no power modules or mission chiplets are placed on the bottom of the substrate. This eliminates the need for TWVs, thereby reducing integration costs while allowing the substrate to remain a full-thickness silicon layer.

Beyond the z-axis, power distribution across the x-y plane follows two primary strategies: lateral (i.e., edge) supply from the substrate edges or uniform vertical supply through an array of connectors. These approaches parallel the distinction between peripheral pad I/Os with wire bonding and uniform bump I/Os in flip-chip C4 bonding. While vertical power delivery offers benefits such as enhanced uniformity and lower PDN impedance, it also introduces challenges such as increased I/O complexity, difficult power connector integration, and extended assembly time. More critically, vertical power delivery obstructs access to thermal dissipation units (TDUs). For instance, if power is uniformly supplied from the dielet side, heat can only be extracted through the substrate side. In contrast, if power is delivered from the system edges, both the top and bottom of the system remain available for thermal management, allowing for more efficient integration of thermal modules.

By considering uniform versus edge and dielet-side (DS) versus substrate-side (SS) power delivery, four distinct power architectures emerge, as illustrated in Fig. 2. Specifically, implementing uniform power distribution on the dielet side requires through-polymer vias (TPVs), where a polymer mold is positioned between chiplets [8]. The TPV terminals then establish connections to power modules located on the dielet side. In this configuration, the entire system must be flipped upside down to accommodate the gravitational constraints associated with two-phase cooling within the TDUs.

In all four cases, voltage regulation modules (VRMs) can be positioned in different configurations to optimize power delivery and efficiency [9], [10]. They may be placed off-package, where the system receives direct point-of-load (PoL) voltage. However, this results in high current distribution. Alternatively, VRMs can be fully on-package, supplying a high input voltage (e.g., 48V, as used in current data center conventions) to minimize distribution current and reduce delivery losses. A hybrid approach is also possible, where high voltage is first converted to an intermediate level off-package before being further stepped down on-package. This balances distribution current and mitigates I^2R losses in the package routing. All these configurations are accounted for in the four architectures and are averaged in the analysis and calculations.

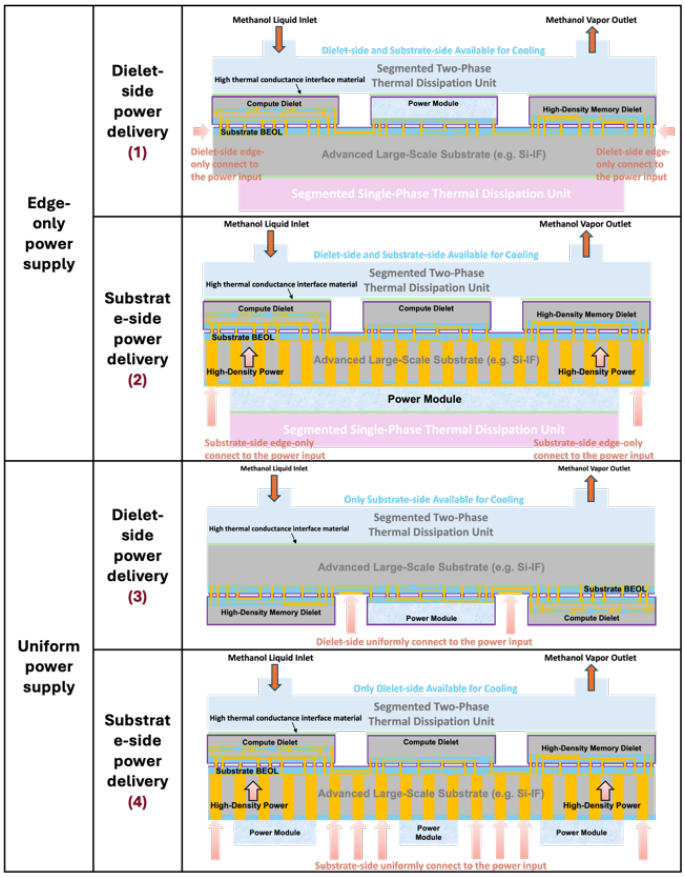

Fig. 2. Schematic view of various thermal dissipation and power delivery architectures. Drawing is not to scale.

The analysis is conducted based on mathematical models, where thermal resistances are estimated using material thickness and thermal conductivity values summarized in Table 1. Furthermore, no lateral heat spreading is assumed. Heat generation naturally occurs across multiple layers in the system, and we consider six distinct heat generation sources: (1) load dissipation at the FEOL layer in dielets, (2) routing loss at the BEOL layer in dielets, (3) routing loss at the dielet-side BEOL layer in the substrate, (4) routing loss at the TWV layer in the substrate, (5) routing loss at the substrate-side BEOL layer in the substrate, and (6) voltage conversion loss in power modules on the package. Any heat generated by off-package components is not captured in this model, as we assume that it is managed separately using a dedicated thermal module, such as a lid or forced air cooling.

TABLE I. MATERIALS THERMAL PROPERTIES USED IN THE MODEL

	0.02" Thermal grease/pad TIM	200 µm Indium TIM	700 um silicon	700 um silicon + Cu TWVs (20% density)	100 um silicon	20 um ABF BEOL (60% Cu density)	10 um oxide BEOL (60% Cu density)	2 mm power module substrate
Thermal conductivity (W/mK)	7	86	150	197	150	231.68	232.12	0.5
Thermal resistance (K/W)	71	2.3	4.67	3.55	0.67	0.086	0.043	4000

Three TIM scenarios are considered: an ideal zero-temperature-drop TIM, a 200 µm uniform indium TIM, and a 500 µm layer of thermal grease/gap pad TIM. The load power density is varied across 0.1, 0.5, and 1 W/mm². For computational simplicity, we assume a fixed heat transfer coefficient (h) of 5000 W/m²K for single-phase water-cooling at the bottom side, representing forced water cooling with a high flow rate. By varying power density and thermal/power architectures, we determine the minimum required cooling capability at the side labeled as the "segmented two-phase thermal dissipation unit" in Fig.2. This does not necessarily imply a two-phase cooling setup. The goal is to maintain the chip temperature below 85°C.

It is important to note that the assumed 5000 W/m²K water-cooling condition may be excessive for low heat flux cases, resulting in a zero-cooling requirement on the TDU side. In practical applications, the water flow rate can be reduced, or the cooling unit repositioned as needed. The results are presented in Fig. 3. Additionally, another interesting conclusion is that when legacy TIMs are used, the chip temperature cannot be maintained at 85°C, regardless of how advanced the TDU is. Even with an infinitely high h-value, the temperature drop across the legacy TIM is significant enough that the chip cannot be cooled to 85ºC In these cases, instead of plotting the required h-value, we predict the maximum chip temperature under optimal two-phase TDU cooling, with h set to 16,000 W/m²K, while maintaining the substrate side water-cooling coefficient at 5000 W/m²K.

As shown in Fig. 3(a) and (b), when using an ideal or a thin, uniform metal TIM, two-phase cooling is required at a power density of 1 W/mm². If cooling is available on only one side, as in architectures 3 and 4, two-phase cooling becomes necessary even at the lower power density of 0.5 W/mm². In contrast, for architectures 1 and 2, two-sided liquid cooling is sufficient to maintain the chip temperature below 85°C at 0.5 W/mm².

In Fig. 3(c), when conventional grease or gap pad TIMs are used, all architectures can be effectively cooled with chip temperatures below 85°C under maximum cooling conditions and lower power densities. However, once the load power density reaches 1 W/mm², the chip temperature exceeds the 85°C limit, highlighting the need for advancements in TIM technology to enable effective heat dissipation at higher power densities.

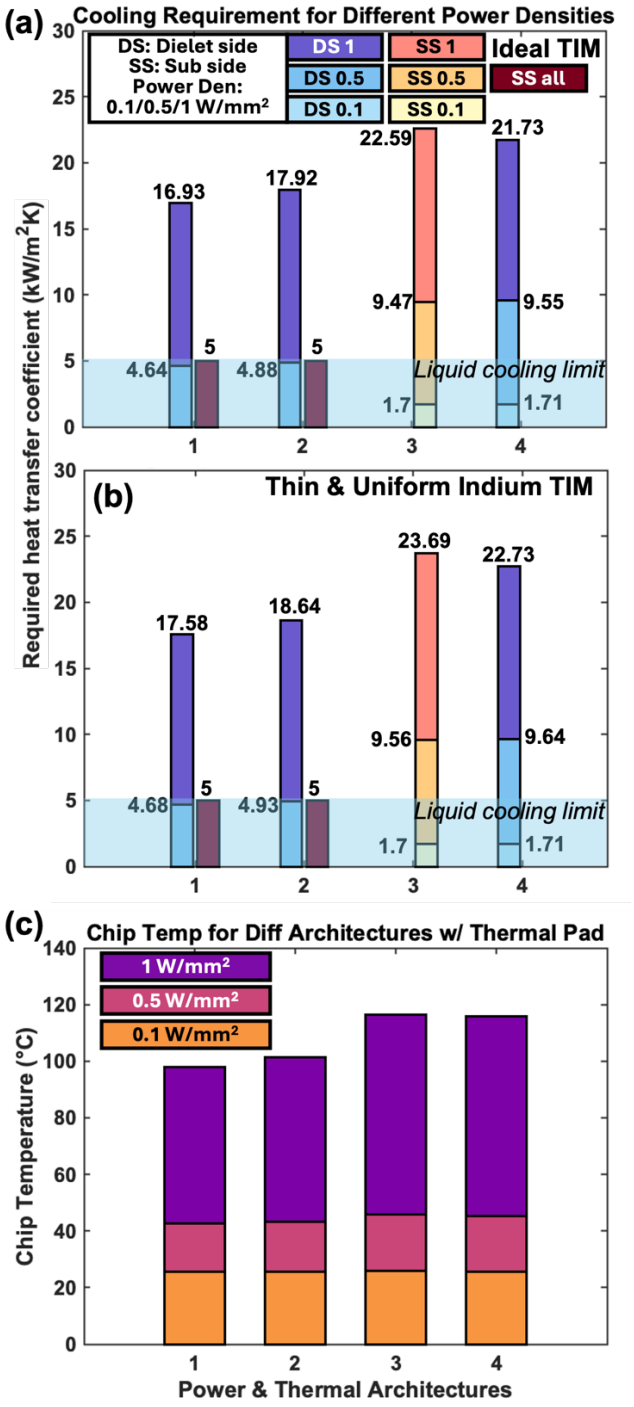

Fig. 3. (a) and (b) Predicted cooling requirements at the dielet side and substrate side for different power/thermal architectures across various power densities maintaining chip temperature at 85ºC. (c) Predicted chip temperature under maximum cooling conditions for different power/thermal architectures at various power densities. The secnarios 1, 2, 3 and 4 are explained in Fig. 2.

III. Experimental Setup for the Demonstration of Vertical Heat Extraction

In Fig. 4(a), the experimental loop and the associated components are illustrated. Liquid methanol is stored in a storage tank (not pictured, below setup) and pumped by a peristaltic pump to the TDU (evaporator assembly). The TDU is maintained at low pressure with the help of a vacuum pump (behind pegboard). The liquid-vapor mixture at the exit of the TDU goes to an accumulator tank where the liquid and vapor phases are separated. The vacuum pump separates the vapor from the accumulator, sending it to the condenser. Both the accumulator and the condenser tanks are behind the pegboard. Completing the loop, the methanol vapor condenses and flows to the storage tank. To enable pulsing of methanol flow, a solenoid valve is placed at the inlet of the TDU; this type of pulsed fluid flow is referred to as flash cooling. In Fig. 4(b) the TTC is pictured, and due to the size mismatch with the TDU, there is a minor amount of heat spreading. The TTC is a silicon chip that employs diodes distributed across the chip area to extract average chip temperature in 1 mm^2 areas.

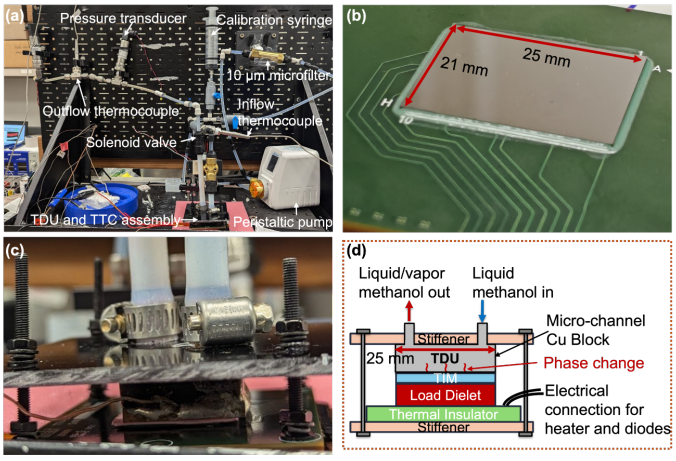

Fig. 4. (a) experimental setup; (b) silicon test chip used in the experiment; (c) closer view of the TDU assembly; (d) schematic view of the TDU assembly.

In Fig. 4(c), the physical evaporator assembly is pictured. Fig. 4(d) illustrates the assembly with a schematic view. The TDU is placed on top of the TTC with a layer of thermal grease (OmegaTherm 201) acting as thermal interface material. For high heat flux experiments, an additional thermal interface material of indium is melted and replaces the thermal grease. The TDU is a microchannel cold plate (Mikros Technologies) with a hydraulic diameter of 160 µm and is made of copper. The assembly is clamped down to a contact pressure of 30 psi.

For system temperature measurements, multiple K-type thermocouples (inlet of the flow, outlet of the flow, bottom of the TDU) are used to record the temperature. The pressure at the exit of the TDU is measured with the help of a pressure transducer. The power supplied to the TTC and the dynamic monitoring and recording of the temperature and pressure data are enabled through LabVIEW. The uncertainties in the measurement of temperature and pressure transducers are ±2.1°C and 1%, respectively.

TABLE II. Design of Experiments

DOE #	Heat Density (W/mm^2)	Methanol Flow Rate (mL/s) (calculated from equation 2)	Pulse Cycle Time (s) and Duty Cycle (on/on+off)	TIM Material
1	0.1	0.5	∞, 1 (continuous)	Thermal Grease (500 µm)
2	0.15	0.5	1, Adaptive (pulse 4)	
3	0.25	0.5	1, Adaptive (pulse 4)	
4	0.5	0.5	∞, 1	
5			0.5, 0.06 (pulse 1)	
6			1, 0.08 (pulse 2)	
7	1	1	∞, 1	
8			0.33, 0.09 (pulse 3)	
9	1	1	∞, 1	Indium (200 µm)
10			0.33, 0.09 (pulse 3)	

Table II. shows the design of experiments (DOE) for the current study. The flow rate for the corresponding heat flux is fixed based on the equation:

$$\frac{P}{\rho_l v h_{fg}} = 0.7 \quad (1)$$

Here, P denotes the power supplied, ρ_l denotes the density of liquid methanol, v denotes the volumetric flow rate, and h_{fg} denotes the latent heat of vaporization. Flow rates higher than this do not affect the system's performance as measured by the TDU's bottom temperature and the chip temperature. The pulse cycle time is fixed so that there is a metered supply of liquid methanol to the TDU as controlled by the peristaltic pump [6].

IV. Results and Discussions

Fig. 5 (a) presents a case study of the quasi-steady-state (QSS) region for the measured TDU temperature, chip temperature, and TDU outlet pressure at a heat flux of 0.5 W/mm^2. QSS is defined where the change in TDU bottom temperature does not exceed 1°C when averaged over the subsequent 200 seconds. The left y-axis represents the temperature data, while the right y-axis shows the outlet pressure. The observed oscillations are attributed to frequent formation and departure from the heated sides of TDU which is a characteristic of the two-phase cooling process and flow pulsing. Fig. 5 (b) illustrates another case study of QSS behavior at 0.25 W/mm^2. In this experiment, the closed-loop system was configured such that liquid flow through the solenoid valve was enabled only when the TDU temperature reached 40°C. At such low heat flux levels, the sawtooth pattern in temperature is due to the periodic nature of liquid methanol injection. When the TDU temperature exceeds 40°C, the valve opens, allowing liquid to enter and cool the system. The valve is then closed until the TDU temperature again surpasses 40°C. Thus, the pulse width and duty cycle were adaptively assigned based on the real-time TDU temperature.

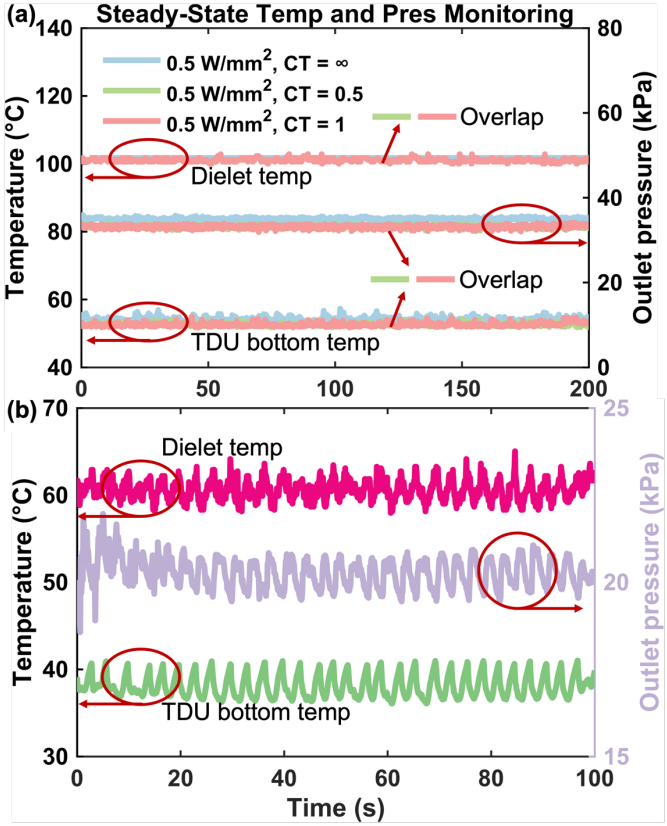

Fig. 5. QSS TDU temperature, chip temperature, and TDU outlet pressure measrement results.

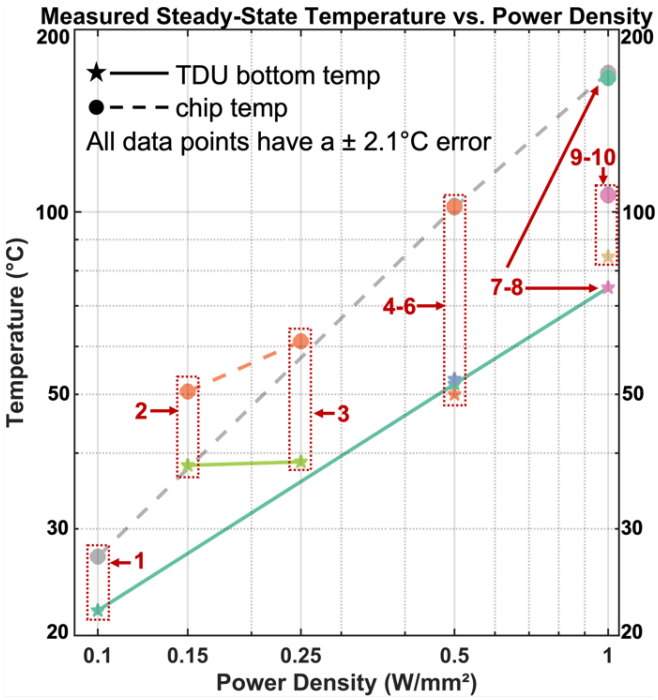

Fig. 6. Measured chip and TDU temperatures at various power densities and DOE conditions (numbers are the DOE #s in Table II).

Fig. 6 compiles all QSS data points across different DOE conditions (detailed in Table II). As expected, both TDU and chip temperatures increase with rising heat flux. Notably, at 1 W/mm², the TDU temperature remains below 85°C, confirming that effective cooling is achieved with minimal lateral heat spreading. However, the temperature drop across the TIM is approximately 80°C at 1 W/mm², leading to a chip temperature as high as 160°C. This highlights a significant TIM limitation which was also observed in section II, which can be mitigated by employing high-performance TIM materials.

To address this issue, experiments were conducted using a ceramic heater (Watlow) and indium TIM, which has a thermal conductivity 40× higher than grease. The results are shown in Fig. 6 (DOE 9 and 10). With indium as the TIM, the heater (replicates the chiplet in a wafer-scale system) temperature dropped to 106°C, and the measured temperature drop across the TIM is 20°C. However, this value deviates from the theoretical 2°C prediction used in Section II, suggesting non-uniform indium distribution and possible air gaps affecting thermal performance.

All experiments were conducted under a fixed flow rate, determined using Equation 1. Under these conditions, the system operates within the nucleate boiling regime, explaining the minimal difference between continuous and pulsed (flash) boiling cases. However, as the system approaches the dry-out limit on the boiling curve, flash boiling demonstrates superior performance over continuous boiling [11].

The TDU side wall temperature was measured using a calibrated infrared (IR) camera (FLIR Boson, 12 µm resolution) for the thermal couple measurement data cross validation. The results, shown in Fig. 7, indicate that the side wall temperature closely follows the TDU bottom temperature, demonstrating temperature uniformity within the TDU due to efficient boiling inside the microchannels.

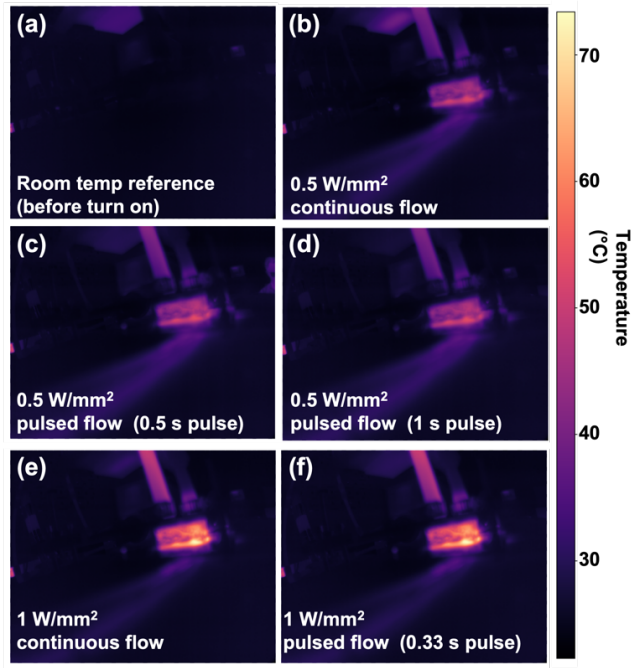

Fig. 7. IR images showing the TDU sidewall temp at various conditions.

V. Conclusion

Wafer-scale systems are driving the future of high-performance computing and large language models, offering data center-level performance on a single wafer. However, effective solutions for heat dissipation and power delivery remain an open challenge. In this paper, we analyzed various power and thermal architectures and experimentally demonstrated two-phase segmented cooling using a silicon heater dielet, representing a unit cell of a wafer-scale prototype. Our results show that flash boiling can cool a heat density of 1 W/ mm² while maintaining the TDU temperature below 85°C. However, the temperature drop across the TIM results in a higher dielet temperature. Shifting from thermal grease to an indium TIM helped narrow down this drop (~160°C of chip temperature with grease and ~106°C with indium), but for future advanced packaging systems, we need an 'all of the above approach' in coming up with high density coolers and better interface materials for good thermal contact between the cooler and the dielet. To the best of our knowledge, this is the first demonstration of high heat flux vertical thermal management for silicon dielets in wafer-scale systems.

Acknowledgment

This work was supported in part by the SRC JUMP CHIMES and the UCLA CHIPS consortium. We thank all UCLA CHIPS and UCLA Nano Transport Research Group (NTRG) students and members.

References

[1] https://www.nist.gov/system/files/documents/2023/11/19/NAPMP-Vision-Paper-20231120.pdf

[2] S. Jangam and S. S. Iyer, "Silicon-Interconnect Fabric for Fine-Pitch (≤10 μm) Heterogeneous Integration," *IEEE Trans. Compon. Packag. Manuf. Technol.*, vol. 11, no. 5, Art. no. 5, 2021, doi: 10.1109/TCPMT.2021.3075219.

[3] T. L. Bergman, *Fundamentals of Heat and Mass Transfer*. John Wiley & Sons, 2011.

[4] J. D. Engerer and T. S. Fisher, "Flash boiling from carbon foams for high-heat-flux transient cooling," *Appl. Phys. Lett.*, vol. 109, no. 2, p. 024102, Jul. 2016, doi: 10.1063/1.4958117.

[5] J. D. Engerer, J. H. Doty, and T. S. Fisher, "Transient thermal analysis of flash-boiling cooling in the presence of high-heat-flux loads," *Int. J. Heat Mass Transf.*, vol. 123, pp. 678–692, Aug. 2018, doi: 10.1016/j.ijheatmasstransfer.2018.02.109.

[6] U. Shah, S. S. Iyer, and T. S. Fisher, "Segmented Thermal Management with Flash Cooling for Heterogeneous Wafer-Scale Systems," in *2021 20th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm)*, Jun. 2021, pp. 589–594. doi: 10.1109/ITherm51669.2021.9503210.

[7] R. Pugazhendhi, T. S. Fisher, and S. S. Iyer, "Pulsed flash boiling for high heat flux electronics cooling," in *2024 23rd IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, May 2024, pp. 1–7. doi: 10.1109/ITherm55375.2024.10709408.

[8] H. Ren, S. Pal, G. Ouyang, R. Irwin, Y.-T. Yang, and S. S. Iyer, "TSV-less Power Delivery for Wafer-scale Assemblies and Interposers," in *2022 IEEE 72nd Electronic Components and Technology Conference (ECTC)*, May 2022, pp. 1934–1939. doi: 10.1109/ECTC51906.2022.00303.

[9] H. Ren *et al.*, "Heterogeneous Power Delivery for Large Chiplet-based Systems using Integrated GaN/Si-Interconnect Fabric with sub-10 μm Bond Pitch," in *2023 International Electron Devices Meeting (IEDM)*, Dec. 2023, pp. 1–4. doi: 10.1109/IEDM45741.2023.10413759.

[10] H. Ren, K. Sahoo, T. Xiang, G. Ouyang, and S. S. Iyer, "Demonstration of a Power-efficient and Cost-effective Power Delivery Architecture for Heterogeneously Integrated Wafer-scale Systems," in *2023 IEEE 73rd Electronic Components and Technology Conference (ECTC)*, May 2023, pp. 1614–1621. doi: 10.1109/ECTC51909.2023.00274.

[11] N. Meenakshi Sundaram, R. Pugazhendhi, S. Iyer, and T. Fisher, "Flash boiling of methanol/water mixtures in a microchannel cooler," in *2025 24th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, May 2025. (under review)