3D Stacked Microfluidic Cooling for High-Performance 3D ICs

Yue Zhang[1], Ashish Dembla[1], Yogendra Joshi[2], Muhannad S. Bakir[1]
[1]School of Electrical & Computer Eng., [2]School of Mechanical Eng., Georgia Institute of Technology, Atlanta, USA
[1]yzhang324@gatech.edu

Abstract—Cooling is a significant challenge for high-performance high-power 3D ICs. In this paper, we describe the experimental evaluation of 3D ICs with embedded microfluidic cooling. Different architectures are experimentally evaluated including: 1) a memory-on-processor stack, 2) a processor-on-processor stack with equal power dissipation, and 3) a processor-on-processor stack with different power dissipation. In all cases, embedded microfluidic cooling shows significant junction temperature reduction compared to air-cooling.

Keywords- High performance 3D IC; on-demand interlayer cooling; single phase cooling; 3D IC centric heat sink design

I. INTRODUCTION

With continued aggressive CMOS scaling, interconnect performance and power dissipation have become a limiting factor for higher-performance integrated circuits [1, 2]. Three-dimensional ICs offer new opportunities for improving chip performance and reducing power dissipation by enabling shorter interconnection length (both on- and off-chip) as well as the possibility of heterogeneous integration. However, a number of challenges must be overcome before 3D ICs can be adopted for high-performance and high-power applications [3-5]. Cooling is a key issue for 3D ICs since both the power dissipation per unit area and the thermal resistance for the dice in the stack to the heat sink increase with the number of tiers. For reference, a few ITRS projections of interest are shown in Figure 1. To address the challenges in heat removal, innovative cooling solutions have been proposed, including single-phase forced microfluidic cooling [6-9], two-phase microfluidic cooling [10, 11], and active thermoelectric coolers to address hotspots [12, 13]. This paper focuses on integrated single-phase microfluidic cooling in 3D ICs.

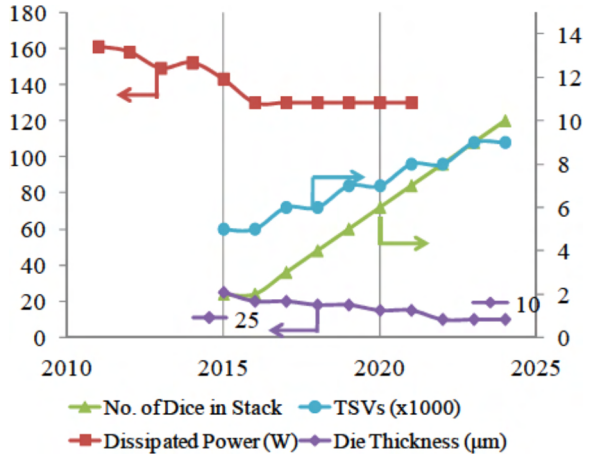

Figure 1. ITRS projections for the number of dice in a stack, number of TSVs, die thickness, and power of a single high-performance chip.

Some advantages and disadvantages of conventional air cooling and microfluidic cooling are summarized below:

(a) Adopting an air-cooled heat sink (ACHS) to reject heat from a 3D stack is simpler to implement (Figure 2(a)). However, it has limited vertical and lateral scalability, as well as limited cooling capability [14]. Even more, considering a memory and processor stack, the processor chip should be placed next to the heat sink in order to have the lowest thermal resistance. However, placing the processor away from the package substrate requires possibly a large number of processor power and ground interconnections (TSVs) through the memory tier, which can present challenges. A single processor requires few thousands of power and signaling interconnections. This large number of I/Os has to be routed through the memory chip which effects memory design, density, and performance. The latter is also impacted by the thermal crosstalk between the two tiers. An additional point of value is the fact that an air-cooled heat sink (and its heat spreader) requires large lateral footprint, which limits how close two chips (whether single-chips or a stack of chips) can be placed laterally if each has its own heat sink. This clearly would impact interconnect length and thus energy and data rate.

(b) Due to the limitations of air-cooled heat sink, many groups have investigated the use of integrated microfluidic heat sink (MFHS) to reject heat. Figure 2(b) depicts a typical system with embedded MFHS where the fluid is supplied through a single inlet [6, 8] from the top of the stack. The authors [6, 8] demonstrate the cooling of a 4-tier and a 2-tier stack with total power dissipation of 200 W and 390 W, respectively.

(c) Figure 2(c) illustrates our vision of a heterogeneous high-performance and high-power 3D IC system featuring a flip-chip compatible inlet/outlet system [15]. The proposed 3D IC system features a silicon interposer with embedded fluidic delivery channels and an array of 3D stacked processor and memory tiers. The processor tiers each contain an embedded microfluidic heat sink. TSVs are routed through the integrated MFHS. The fluid is delivered from the interposer to each tier, possibly, independently through microscale fluidic I/Os formed using either solder or polymer [15]. This approach allows on-demand cooling to each tier and helps minimize the thermal gradient across the stack when power dissipation varies in the stack. Without a bulky air-cooled heat sink, this approach allows high lateral scalability of the electronic components, i.e., placing an array of 3D ICs laterally next to each other. Pumping power may be reduced by adjusting the flow rates

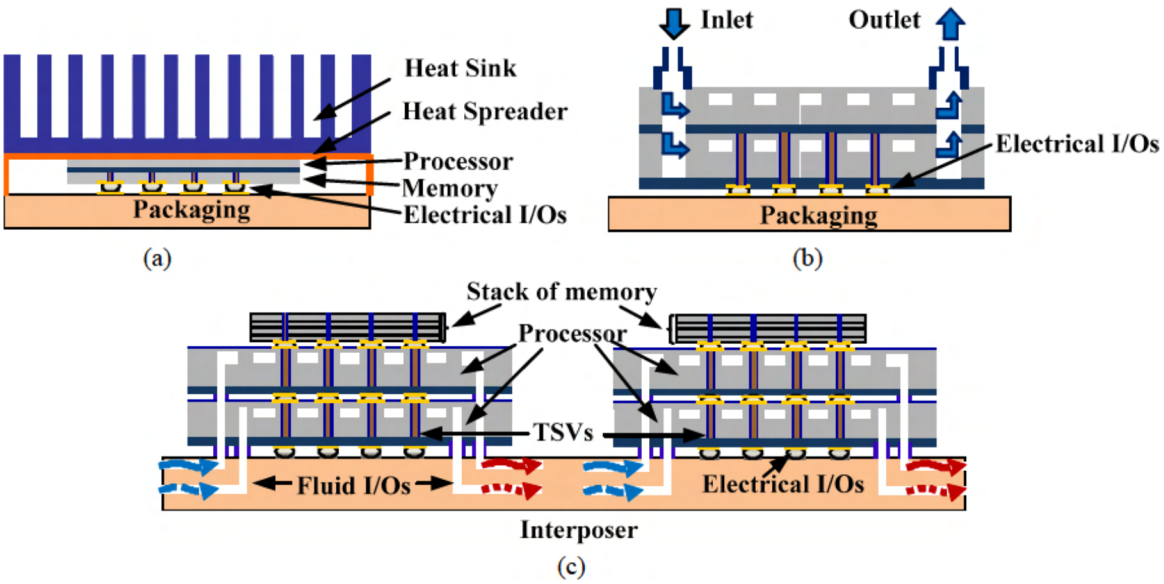

Figure 2. Illustration of (a) conventional air cooling technology, (b) integrated microfluidic cooling technology, and (c) on-demand microfluidic cooling technology.

to the needed value for a given power dissipation per tier.

II. Design Considerations of Micropin-fin Heat Sink

Tuckerman and Pease have evaluated the heat removal capability of microchannel heat sinks [7]. Their results have inspired significant continued research in this field [16-19]. Thermal resistance and pressure drop are two key parameters used to evaluate the microfluidic heat sink performance. However, there is a limited discussion in the literature on the design of embedded microfluidic heat sink while considering electrical interconnect (TSV) placement in 3D ICs. In [20], we reported an initial attempt at the co-design of embedded microfluidic heat sinks and TSVs using silicon micropin-fins. For tier-to-tier communication (as well as power delivery to the stack), TSVs must be routed through the microfluidic heat sink (Figure 3). Thus, the height of the heat sink is a critical consideration in the design of the electrical interconnect network. Increasing micropin-fin height limits the fabrication of TSVs that are typically aspect ratio limited, leading to large diameter TSVs in tall micropin-fins. These TSVs have a relatively large capacitance resulting in energy expensive TSVs and limited vertical interconnect densities (as shown in Figure 4). This ultimately impacts the bandwidth and energy per bit of 3D IC interconnects.

Designing the microfluidic heat sink to be taller reduces the thermal resistance and pressure drop, in general, but increases the capacitance and thus energy dissipation of TSVs. With these new design considerations, an ultra-short staggered micropin-fin structure (Figure 5) was introduced to benefit TSV performance and density while having thermally acceptable performance. In our previous work [20], we reported the performance of the micropin-fin heat sink integrated and benchmarked its thermal resistance with a conventional ACHS. At a power density of 100 W/cm^2, a junction temperature reduction of more than 24 °C was observed for the microfluidic cooled chip relative to the

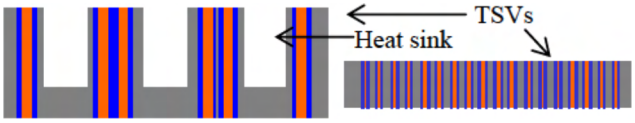

Figure 3. Comparison of TSVs in chips with (left) and without (right) integrated microfluidic heat sink.

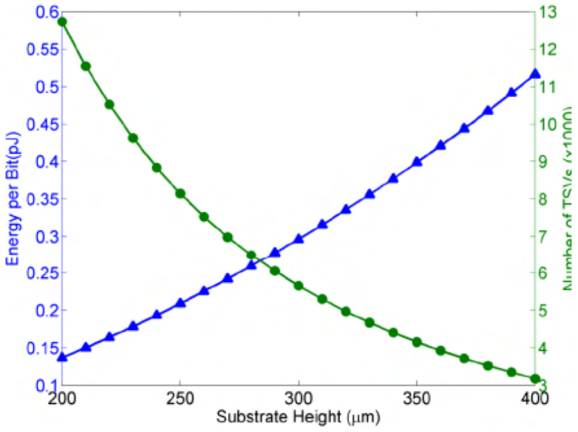

Figure 4. Energy per bit and number of TSVs as a function of substrate height.

air-cooled chip. Even though the heat sink is intentionally designed to have a minimal height, the chip thickness is around 200 μm, which is much larger than typical 3D chips without MFHS (<50 μm). Larger chip thickness results in larger TSV parasitics leading to larger latency and energy dissipation. One solution to reduce TSV capacitance is to integrate polymer clad TSVs instead of oxide liner TSVs [21]. Another method to decrease TSV capacitance is to integrate high aspect ratio TSVs. In [20], we show that increasing TSV aspect ratio from 10:1 to 20:1 leads to a TSV capacitance reduction of ~40%.

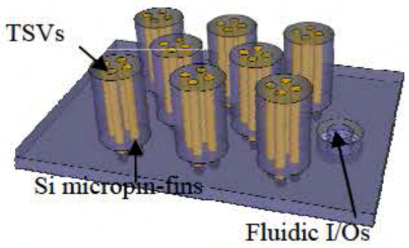

Figure 5. Staggered micropin-fin heat sink concept.

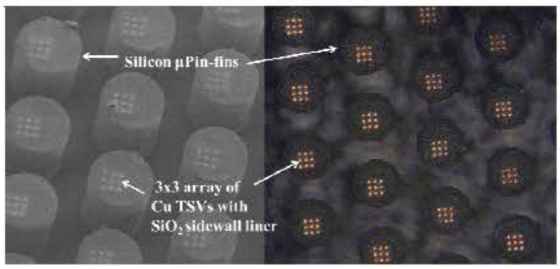

Figure 6. Top view of high aspect ratio TSVs integrated in a silicon micropin-fin heat sink (10 μm TSV diameter) [22].

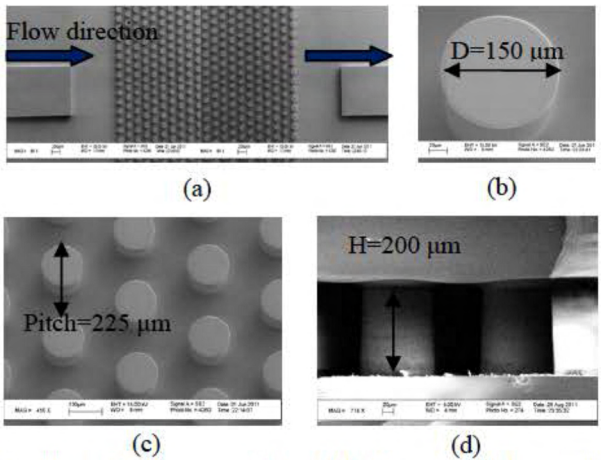

Figure 7. SEM images of (a) global view of the heat sink, (b) single micropin-fin, (c) closer view of micropin-fin array, and (d) cross-section of the heat sink

The fabrication of high aspect ratio (~20:1) TSVs within the micropin-fin heat sink [22] is realized by using the standard Bosch process which alternates between the etch and deposition steps. Thermal oxide liner is grown to insulate TSVs and the substrate. A pulsed plating step is then performed to electroplate the copper into the vias. SEM images show no voids in the TSV (Figure 6 [22]). The high aspect ratio TSVs help reduce the TSV parasitics to improve TSV performance within ICs with integrated MFHS.

III. THERMAL TESTBEDS AND TESTING SETUP

Figure 7 illustrates SEM images of the micropin-fins. Detailed fabrication flow of the micropin-fins, including capping, is described in [20]. The dimensions of the shown staggered micropin-fin heat sink are labeled in the SEM

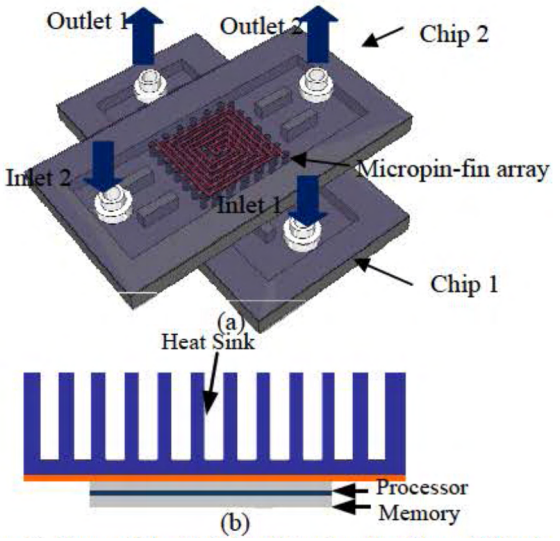

Figure 8. Thermal testbed for (a) microfluidic and (b) air cooling testing.

images. The heating area is 0.6 cm by 0.6 cm. Figure 8(a) shows the 3D stacked testbed for microfluidic testing adopted in this work. This is an attempt to simplify the fabrication needed to thermally prototype the system shown in Figure 2(c). Thermal interface material (TIM) is used between the two tiers. For the sake of simplified port access, the two tiers are stacked orthogonally such that the inlets and outlets are easily accessible. Independent coolants can be delivered to each tier and with differing flow rates. In addition, in a heterogeneous 3D IC stack, the different stacked chips may have different workloads/performance, which results in different power dissipations. Thus, one may not need the same flow rates in each tier (in fact, even the same microfluidic heat sink design). The thermal testbed under consideration provides the ability to explore the benefits of on-demand cooling in each tier, i.e., independently tailored flow rates in each tier. Results for this configuration are discussed in Section IV.

In order to attain an initial insight into the benefits of embedded microfluidic cooling, a 3D thermal testbed with no liquid cooling was constructed. In this case, we used an air-cooled heat sink on the top of the stack. This is shown in Figure 8(b). The same type of TIM is applied between the fan base and the stack as well as between the two tiers. Note that since the TIMs were manually applied, the thickness of TIMs may not be identical in the two testbeds. The heating area of the ACHS testbed is 1 cm by 1 cm. Similar to the MFHS, the two chips are stacked orthogonally. Since heat is removed from the top, the bottom chip has a longer path to the heat sink. Experimental results are shown in Section IV.

In the test setup for the MFHS (Figures 9 and 10), two pumps are connected to the two inlets in the stack (i.e., each tier has its own inlet and pump). De-ionized (DI) water is pumped from a nearby reservoir. Polyester based filters are connected to the outlet of the pump to eliminate any particles (>20 μm) that may potentially block the microfluidic heat sink. An acrylic block flow meter that measures up to 100 mL/min is connected to each inlet serially to measure the flow rate. An Agilent N6705B power analyzer with 4 outputs

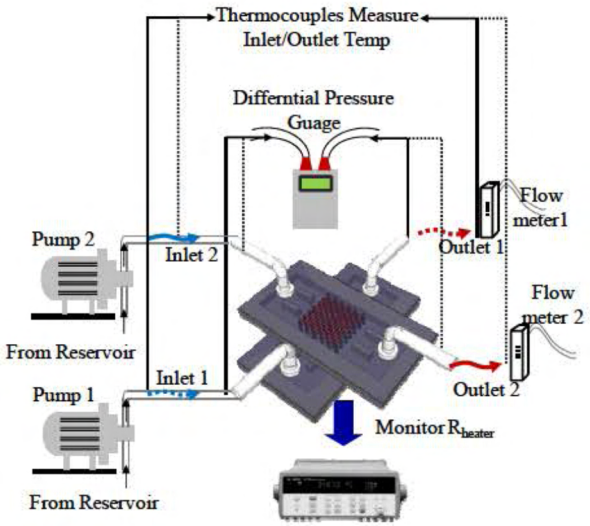

Figure 9. Schematic of the test setup of MFHS cooling.

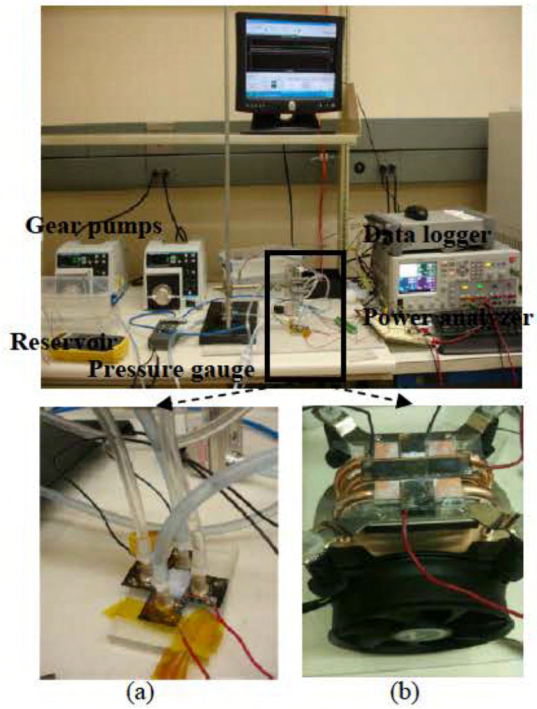

Figure 10. Test setup for (a) microfluidic and (b) air cooled 3D stack.

is used to source current to the thin-film platinum heaters/thermometers on the chip surfaces to emulate chip power dissipation [16]. The Pt heaters also serve as temperature sensors (with <1% error) due to their linear electrical resistance-temperature relationship. The heater resistance in each tier is measured and tracked using an Agilent 34970A data logger. Since we use a single heater per tier, the measured resistance, and thus junction temperature, represents the average junction temperature in each tier.

IV. EXPERIMENTAL RESULTS

MFHS are evaluated in different 3D architectures in this section. The flow rate used in this set of measurements is 60 mL/min unless otherwise specified. The inlet water temperature is at 19 °C with a variance of < 1 °C. The inlet and outlet fluid temperatures are recorded, and are used to calculate the heat removal by the fluid. At high power, the values are in good agreement with the power dissipation of the chips. At low power, however, there is a difference between the two values. Since the heating areas are different in the MFHS and ACHS testbeds, power density is used in the figures and for comparison. This section will provide the raw data, and the following section will compare the results.

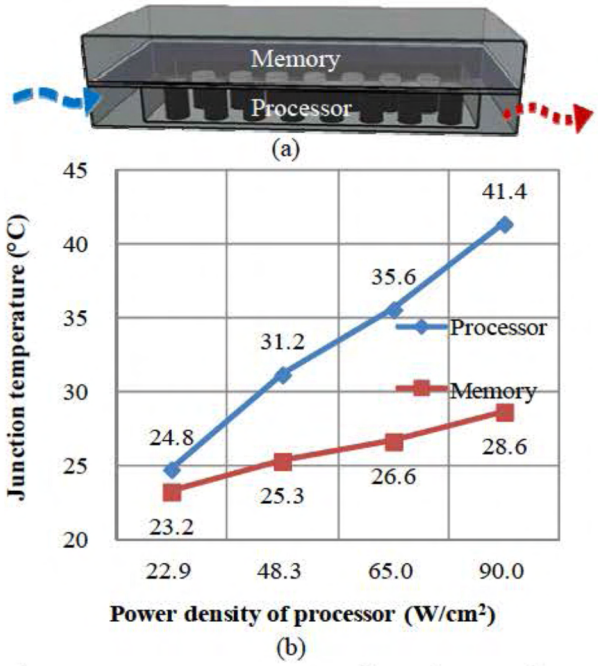

Figure 11. Memory-on-processor thermal test results.

A. Memory-on-processor stack

In Figure 11(a), the fluid is pumped only into the processor layer at a flow rate of 60±5 mL/min. In this experiment, the heat flux of the memory chip was held at ~5 W/cm^2. Since the memory chip is stacked above the processor layer with integrated microfluidic cooled heat sink, the microfluidic heat sink serves as a path for cooling of the memory chip as well. Junction temperature results for this scenario are shown in Figure 11(b). The memory temperature only increases by 5.4 °C when the heat flux of the (bottom) processor increases from 22.9 W/ cm^2 to 90W/ cm^2.

B. A two processor stack with identical power dissipation

In Figure 12(a), the shown two chip stack dissipates up to 100 W/cm^2 per tier to simulate the stacking of processors. A microfluidic heat sink is integrated into both tiers. The flow rate in both tiers is 60 mL/min. Two set of measurements were done for the same stack, and the average junction temperature in each chip is plotted in Figure 12(b). The difference in the two measurements did not exceed 1.1 °C. As seen from the plots, when the power dissipation in each tier is more than 100 W/cm^2, the temperature in either tier is less than 48 °C.

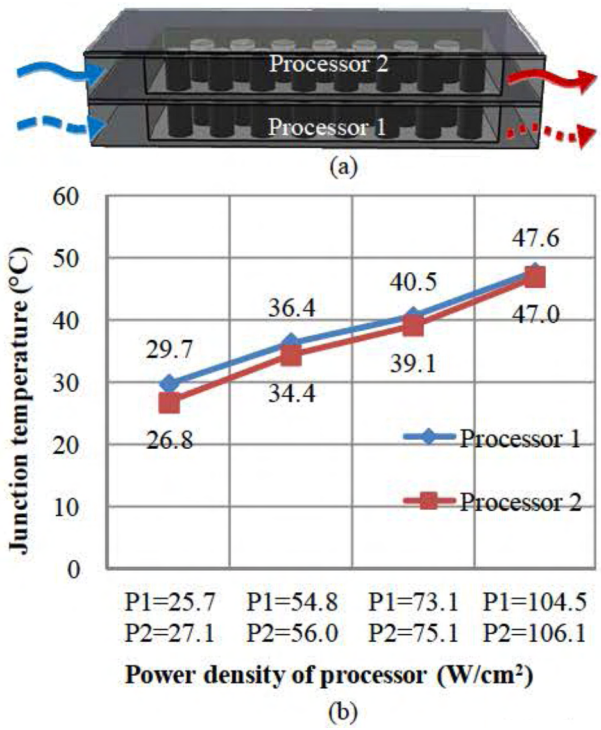

Figure 12. A stack of two processors cooled using MFHS.

C. A two processor stack with differing power dissipation

This test case simulates two processors with different power densities: 55 W/cm^2 and 100 W/cm^2. The on-demand flow-rate (and thus cooling) that we proposed is implemented. The junction temperature in each tier as a function of flow rate (Q) is shown in Figure 13. In one of the shown cases, the flow rates for the 100 W/cm^2 and 55 W/cm^2 chips are 70 mL/min and 40 mL/min, respectively. Compared to the case when they are cooled at the same flow rate, the temperature difference between the two chips decreases from 12 °C to 7 °C. Further increasing the flow rate difference may result in a smaller temperature gradient in the stack.

D. Air-cooled 3D IC stack

In order to gain preliminary insight into how much improvement embedded microfluidic cooling provides, we build a similar 3D thermal testbed with the main difference being that no embedded cooling was integrated. We interfaced an ACHS to the top of the stack. The thickness of the silicon tiers was 300 µm. Moreover, no TSVs were integrated in the testbed, which can improve the thermal measurements we report in this study. Using this simplified testbed, similar measurements were made for the ACHS testbed; here we only show the measured junction temperatures for case when two high power chips are stacked. In this experiment, the two tiers are powered up to ~50 W/cm^2, which is set to prevent the chip temperatures from being too high (100 °C). As shown in Figure 14, the temperatures of the two tiers are much higher than the chips under microfluidic cooling (Figure 14). Moreover, there is a large temperature gradient (19.2 °C) between the two chips at high power. Similar to the microfluidic cooled cases, we also tested processor and memory stacks. The power density of the memory chip was

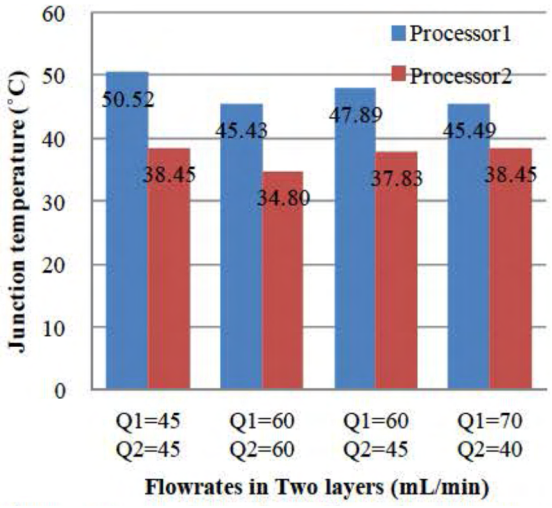

Figure 13. On-demand cooling for two processors dissipating the same value of power.

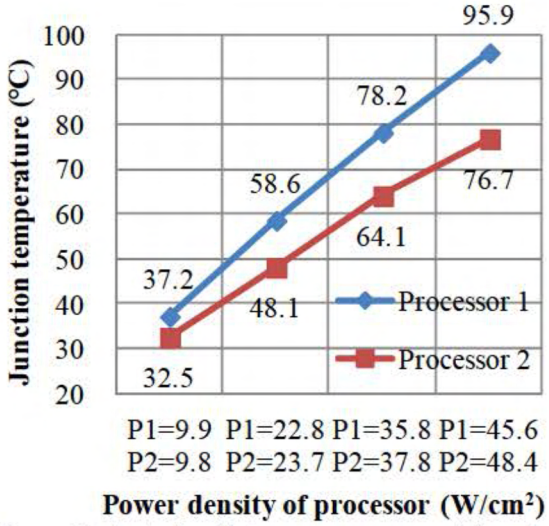

Figure 14. A stack with two processors cooled under an ACHS.

again held at 5 W/cm^2 and the processor power was ramped up to 50 W/cm^2. The two configurations tested were: memory closer to the air-cooled heat sink (ACHS (1)) and processor closer to the air-cooled heat sink (ACHS (2)). These initial experiments help verify the benefits of embedded liquid cooling.

V. DISCUSSION AND IMPLICATIONS

Based on the microfluidic and air-cooled experiments, representative data points are listed in Table 1 for comparison. As shown, in all cases, embedded MFHS shows significant junction temperature reduction compared to ACHS even with the higher power dissipation. The MFHS thermal resistance for a single tier was characterized to be ~0.26 Kcm^2/W, while the ACHS thermal resistance was tested to be 0.55 Kcm^2/W

(this value includes the TIM between the chip and the heat sink). The advantages of maintaining ICs at low junction temperature are numerous, including lower leakage power, higher device lifetime, reduced electromigration, and higher system reliability potentially. It is shown in [16] using a compact physical model that the power dissipation of a microprocessor decreases from 102 W to 83 W as the chip temperature decreases from 88 °C to 47 °C. The thermal coupling that occurs in the 3D stack as a function of cooling technology is interesting. From the measured data (for both the ACHS and MFHS 3D stack), the junction temperature of the low-power memory chip is influenced by the power variation in the processor chip even when the power dissipation of the memory chip is held constant. To quantify this effect, the temperature increase (ΔT) in the memory chip as a function of power in the processor tier is plotted in Figure 15. The junction temperature increases more rapidly for ACHS than for MFHS. The slopes of ACHS (1) and ACHS (2) are 0.55 Kcm^2/W and 0.52 Kcm^2/W, respectively, while that of the MFHS is 0.08 Kcm^2/W. The temperature varies 6 times slower in the MFHS case. The reason it is believed that this occurs is because the DI water in the MFHS serves as a thermal buffer between the memory and processor tiers. Therefore, the processor workload variance does not greatly impact the junction temperature of the memory.

Table 1. Summary of some representative data points for microfluidic cooling (white) and air cooling (blue)

	Power density (W/cm^2)		Flow rate (ml/min)		Junction T (°C)	
	BTM	Top	BTM	Top	BTM	Top
Two Processors	54.8	56.0	60	60	36.4	34.4
	104.5	106.1	60	60	47.6	47.0
	45.6	48.4	-	-	95.9	76.7
Memory + Processor	5	48.3	0	60	25.3	31.2
	5	90.0	0	60	28.6	41.4
	5	57.1	-	-	61.2	59.0
	49.3	5	-	-	79.3	50.7
Identical flow rate	102	55	45	45	50.5	38.5
Different flow rate	105	55	70	40	45.5	38.5

*BTM=Bottom

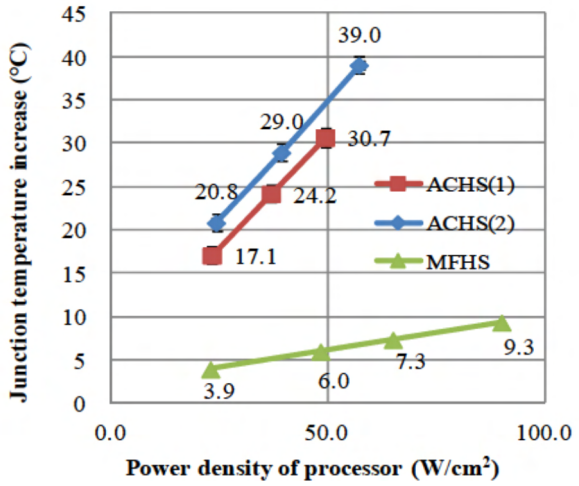

Figure 15. Increase of memory chip junction temperature as a function of different 3D stack configration. ACHS (1) is the architecture where memory is closer to the fan. ACHS (2) is the case when the processor is closer to the fan. For the case of the MFHS, only the processor is liquid cooled, as shown in Figure 11(a).

VI. Conclusion

Experimental results show that a MFHS has superior heat removal capability relative to an ACHS for 3D ICs. The MFHS maintains the stack temperature below 50 °C for a total power density of 200 W/cm^2 in a two-tier stack. Moreover, the thermal coupling effect is reduced when a MFHS is used. Finally, MFHS based on-demand cooling approach is shown to enable a reduction in the thermal gradient within the stack by supplying liquid at different flow rates to tiers with different power dissipation.

Acknowledgment

This work has been carried out in part under Microelectronics Advanced Research Corporation (MARCO), its participating companies, and DARPA through the Interconnect Focus Center and funding from DoD.

References

[1] M. T. Bohr, "Interconnect scaling-the real limiter to high performance ULSI," in *Proc. International Electron Devices Meeting,* 1995, pp. 241-244.

[2] S. Borkar, "Thousand core chips: a technology perspective," in *Proc. 44th annual Design Automation Conference*, San Diego, California, 2007. pp. 746-749

[3] W. R. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. M. Sule, M. Steer, and P. D. Franzon, "Demystifying 3D ICs: the pros and cons of going vertical," *IEEE Des. Test Comput.,* vol. 22, no. 6, pp. 498-510, 2005.

[4] P. G. Emma and E. Kursun, "Is 3D chip technology the next growth engine for performance improvement?," *IBM Journal of Research and Development,* vol. 52, pp. 541-552, 2008.

[5] S. M. Sri-Jayantha, G. McVicker, K. Bernstein, and J. U. Knickerbocker, "Thermomechanical modeling of 3D electronic packages," *IBM Journal of Research and Development,* vol. 52, pp. 623-634, 2008.

[6] T. Brunschwiler, S. Paredes, U. Drechsler, B. Michel, W. Cesar, G. Toral, Y. Temiz, and Y. Leblebici, "Validation of the porous-medium approach to model interlayer-cooled 3D-chip stacks," in *Proc. 3D System Integration,* 2009, pp. 1-10.

[7] D. B. Tuckerman and R. F. W. Pease, "High-performance heat sinking for VLSI," *IEEE Electron Device Lett.,* vol. 2, pp. 126-129, 1981.

[8] N. Khan, Y. Hong, P. Tan Siow, H. Soon Wee, S. Nandar, H. Wai Yin, V. Kripesh, Pinjala, J. H. Lau, and C. Toh Kok, "3D packaging with through ilicon via (TSV) for electrical and fluidic interconnections," in *Proc. Electronic Components and Technology Conference,* 2009, pp. 1153-1158.

[9] Y. Peles, A. Kosar, C. Mishra, C.-J. Kuo, and B. Schneider, "Forced convective heat transfer across a pin fin micro heat sink," *International Journal of Heat and Mass Transfer,* vol. 48, pp. 3615-3627, 2005.

[10] B. Agostini, J. R. Thome, M. Fabbri, B. Michel, D. Calmi, and U. Kloter, "High heat flux flow boiling in silicon multi-microchannels – Part I: Heat transfer characteristics of refrigerant R236fa," *International Journal of Heat and Mass Transfer,* vol. 51, pp. 5400-5414, 2008.

[11] W. Qu and A. Siu-Ho, "Experimental study of saturated flow boiling heat transfer in an array of staggered micro-pin-fins," *International Journal of Heat and Mass Transfer,* vol. 52, pp. 1853-1863, 2009.

[12] V. Sahu, Y. K. Joshi, and A. G. Fedorov, "Hybrid solid state/fluidic cooling for hotspot removal," in *Proc. Thermal and Thermomechanical Phenomena in Electronic Systems,* 2008, pp. 626-631.

[13] Y. Bao, W. Peng, and A. Bar-Cohen, "Thermoelectric mini-contact cooler for hot-spot removal in high power devices," in *Proc. Electronic Components and Technology Conference, 2006*

[14] L. Sheng-Chih and K. Banerjee, "Cool Chips: Opportunities and Implications for Power and Thermal Management," *IEEE Trans. Electron Devices,* vol. 55, pp. 245-255, 2008.

[15] 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 *Proc. Electronic Components and Technology Conference,* 2008, pp. 1-7.

[16] D. Sekar, C. King, B. Dang, T. Spencer, H. Thacker, P. Joseph, M. Bakir, and J. Meindl, "A 3D-IC Technology with Integrated Microchannel Cooling," in *Proc. Interconnect Technology Conference,* 2008, pp. 13-15.

[17] A. Kosar, C. Mishra, and Y. Peles, "Laminar Flow Across a Bank of Low Aspect Ratio Micro Pin Fins," *Journal of Fluids Engineering,* vol. 127, pp. 419-430, 2005.

[18] T. Brunschwiler, B. Michel, H. Rothuizen, U. Kloter, B. Wunderle, H. Oppermann, and H. Reichl, "Interlayer cooling potential in vertically integrated packages," *Microsystem Technologies,* vol. 15, pp. 57-74, 2009.

[19] M. S. Bakir, C. King, D. Sekar, H. Thacker, D. Bing, H. Gang, A. Naeemi, and J. D. Meindl, "3D heterogeneous integrated systems: Liquid cooling, power delivery, and implementation," in *Proc. Custom Integrated Circuits Conference,* 2008, pp. 663-670.

[20] Y. Zhang, C. R. King, J. Zaveri, K. Yoon Jo, V. Sahu, Y. Joshi, and M. S. Bakir, "Coupled electrical and thermal 3D IC centric microfluidic heat sink design and technology," in *Proc. Electronic Components and Technology Conference,* 2011, pp. 2037-2044.

[21] Y. Civale, D. S. Tezcan, H. G. G. Philipsen, F. F. C. Duval, P. Jaenen, Y. Travaly, P. Soussan, B. Swinnen, and E. Beyne, "3-D Wafer-Level Packaging Die Stacking Using Spin-on-Dielectric Polymer Liner Through-Silicon Vias," *IEEE Trans. Compon, Packag and Manuf. Technol.,* vol. 1, pp. 833-840, 2011.

[22] A. Dembla, Y. Zhang, and M. S. Bakir, "Fine Pitch TSV Integration in Silicon Micropin-Fin Heat Sinks for 3D ICs," in *Proc. Intenational Interconnect Technology Conference*, 2012.