Independent interlayer microfluidic cooling for heterogeneous 3D IC applications

Yue Zhang and M. S. Bakir

Presented for the first time is the implementation of independent microfluidic cooling of different tiers in a 3D stack based on their power dissipation. The impact of this approach on heterogeneous 3D IC stacks, such as memory-on-processor and processor-on-processor with different power dissipations, has been experimentally explored. The junction temperature difference between tiers with different power dissipation is decreased from 12 to 7°C. Significant junction temperature reduction and thermal decoupling are achieved by this approach compared to air-cooling.

Introduction: 3D 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. Cooling is a key challenge for 3D ICs since both the power dissipation per unit area and the thermal resistance for the dice within the stack, which have no direct access to a heatsink, increase with the increasing number of tiers [1].

Demonstration of heat removal (790 W) by a silicon microfluidic heatsink (MFHS) [2] was first shown by Tuckerman and Pease in 1981. The results have inspired researchers in both academia and industry to explore interlayer microfluidic heatsinks for 3D ICs. Khan *et al.* and Brunschwiler *et al.* demonstrated cooling of a two-tier and a four-tier stack with total power dissipation of 200 and 390 W, respectively [3, 4]. In the previous studies, the coolant was injected into the stack through one common inlet and was distributed into each tier. Thus, one cannot control the distribution flow rate of the coolant in each tier. However, in a realistic 3D stack with heterogeneous elements, one needs to control the coolant flow rate in each tier independently. For example, a coolant may be supplied into the processor tier in a memory-on-processor stack ,or coolants with different flow rates may be supplied to each tier in a two-processor stack with different workloads (thus different power dissipations). For the first time, this work proposes and implements independent interlayer microfluidic cooling in different tiers for heterogeneous 3D IC applications. This approach helps reduce the thermal gradient in a heterogeneous 3D stack, lowering thermomechanical stress as well as minimising thermal induced variations in the stack. Additionally, adjusting the flow rate according to the power dissipation saves pumping power by preventing overcooling of the system.

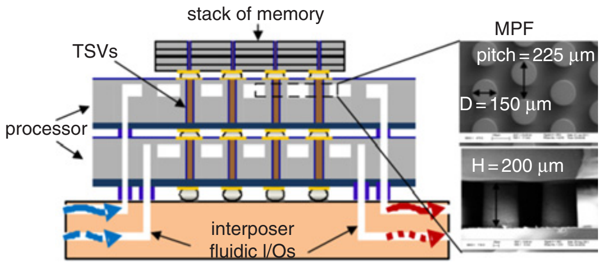

Fig. 1 *Illustration of tier independent microfluidic cooling in heterogeneous 3D ICs*

3D system with independent tier cooling: Fig. 1 illustrates our vision of a heterogeneous high-performance and high-power 3D IC system featuring a flip-chip compatible inlet/outlet system. The proposed 3D IC system features a silicon interposer with embedded fluidic delivery microchannels and a 3D stack of processor and memory tiers. Each high-performance processor tier contains an embedded MFHS. Through silicon vias (TSVs) are routed through the integrated MFHS. The fluid is delivered from the interposer to each tier independently, possibly through microscale fluidic I/Os formed using either solder or polymer [5]. This approach allows independent cooling of each tier. Compared to integrated microfluidic cooling, an air-cooled heatsink (ACHS) is simpler to implement, but has limited cooling capability. Considering a memory and processor stack under air-cooling, the processor chip would have to be placed next to the heatsink in order to have the lowest thermal resistance. However, placing the processor away from the package substrate requires a large amount of power and ground TSVs that need to go through the memory tier(s). This impacts memory design, density and performance. The latter is also impacted by the thermal crosstalk between two tiers. Moreover, an ACHS requires a large lateral footprint, and thus limiting how close two chips or two chip stacks can be placed laterally if each has its own ACHS. This clearly would impact off-chip interconnect length, and thus energy-per-bit and aggregate data rates.

Thermal testbeds and experimental procedure: A 3D stacked testbed featuring MFHS has been developed. The specific micropin-fin (MPF) design provides improved heat transfer capability compared to microchannels [6, 7]. The MPFs are fabricated using DRIE and capped using silicon-to-silicon bonding [6]. Fig. 1 shows SEM images of the MPF used in this work. An on-chip thin-film platinum heater is deposited on each testbed in an area of 0.6 by 0.6 cm. The Pt heater also serves as an resistance thermal detector (RTD) (with < 1% error). Since a single RTD is used per tier, the measured junction temperature represents the average junction temperature in each tier. A thermal interface material is used between the two tiers. The two tiers are stacked orthogonally (Fig. 2) such that the inlets and outlets are accessible. This is an attempt to simplify the fabrication needed to thermally prototype the system shown in Fig. 1. In the MFHS test setup (Fig. 2), two individual pumps were connected to the two inlets in the stack. A flow meter was connected to each outlet serially. The fluid temperatures were measured by K-type thermocouples. An Agilent power analyser was used to source current to the on chip Pt heaters to emulate chip power dissipation. A similar 3D thermal testbed with no embedded MFHS was also constructed. A high-performance ACHS containing aluminium fins and heat pipes designed for the Intel i7 processor was interfaced on top of the stack. The fan operated at 2500 RPM during the testing. The heating area of the ACHS testbed is 1 by 1 cm. Since the heating areas are different in the two testbeds, power density is used in the Figures and for comparison.

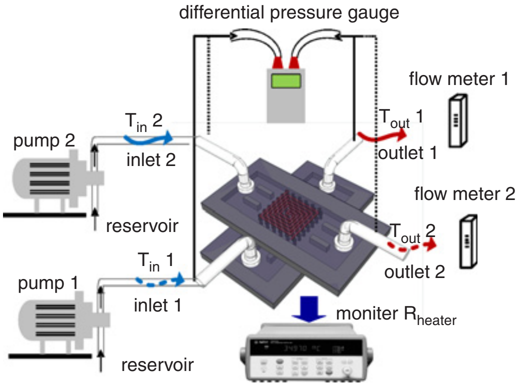

Fig. 2 *Schematic of test setup of MFHS cooling*

Independent cooling of a memory-on-processor stack: In this experiment, the power density of the memory chip is held at ~5 W/cm^2 while that of the processor chip is varied. In Fig. 3*a*, the fluid is only pumped through the processor layer with a flow rate of 60 ± 5 ml/min. Since the memory chip stacked on the processor layer does not have an embedded heatsink, the MFHS in the processor tier serves as a path for cooling of the entire stack. The memory chip temperature increases by 5.4°C when the heat flux of the (bottom) processor increases from 22.9 W/cm^2 to 90 W/cm^2. In Figs. 3*b* and *c*, memory-on-processor and processor-on-memory stacks are cooled using an ACHS. The MFHS cooled 3D stack shows significant junction temperature reduction compared to the ACHS. The advantages of maintaining ICs at low junction temperature are numerous, including lower leakage power, longer device lifetime, reduced electromigration and potentially higher system reliability. The modelling in [8] shows that the power dissipation of a microprocessor decreases from 102 to 83 W (for the same clock frequency) due to reduction of leakage current as the chip temperature decreases from 88 to 47°C. As preliminary validation of the experimental results for MFHS, the heat rejected from the DI water (P_r) is calculated and compared with the power injected into the stack (P_i). In the memory-on-processor stack, P_i is 34.2 W while P_r is

30.7 W (error < 10%). The difference may be induced in part due to the heat exchange with the ambient error.

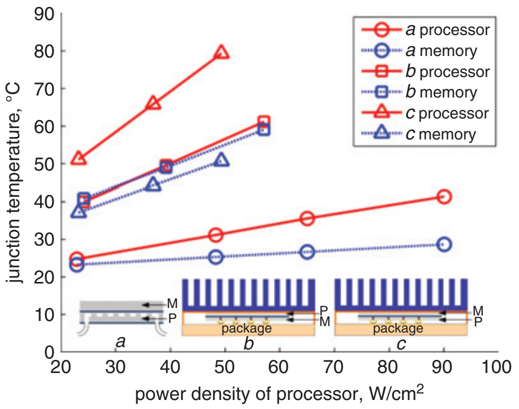

Fig. 3 *Junction temperature for memory and processor tiers using different cooling technologies in stack containing memory and processor chips (P = processor, M = memory)*

Independent cooling of two-processor stack with different power dissipation: This test case emulates two processors (P1 and P2) with different power densities: 100 and 55 W/cm^2, respectively. The proposed scheme of tier independent tailored flow rates was implemented. In one of the shown cases in Fig. 4, the flow rates (Q1 and Q2) for chips P1 and P2 were 70 and 40 ml/min, respectively. Compared to the case where they are both cooled at 45 ml/min, the temperature difference between the two chips decreases from 12 to 7°C. Further increasing Q1 may result in a smaller temperature gradient in the stack.

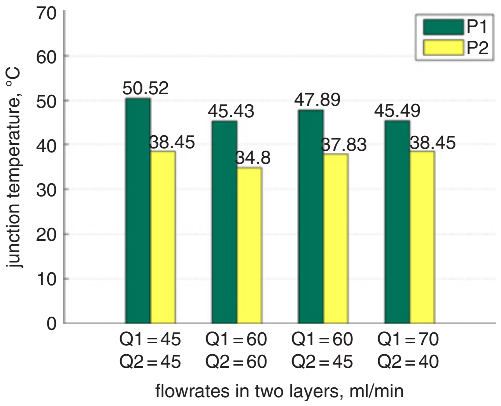

Fig. 4 *Junction temperature of two processors with different power dissipation where independent microfluidic cooling is implemented [9]*

Conclusion: Compared to air cooling, we have demonstrated that MFHS reduces the junction temperature by ∼25°C for a memory-on-processor stack. The independent cooling approach proposed in this Letter is shown to reduce the temperature difference between two tiers with different powers.

Acknowledgments: This work was supported by Microelectronics Advanced Research Corporation (MARCO), its participating companies, and DARPA through the Interconnect Focus Center and funding from DoD.

20 September 2012
doi: 10.1049/el.2012.3313
One or more of the Figures in this Letter are available in colour online.

Yue Zhang and M. S. Bakir (*School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA*)

E-mail: yzhang324@gatech.edu

References

1 Davis, W. R., Wilson, J., Mick, S., Xu, J., Hua, H., Mineo, C., Sule, A. M., Steer, M., and Franzon, P. D.: 'Demystifying 3D ICs: the pros and cons of going vertical', *IEEE Des. Test Comput.*, 2005, **22**, (6), pp. 498–510

2 Tuckerman, D. B., and Pease, R. F. W.: 'High-performance heat sinking for VLSI', *IEEE Electron Device Lett.*, 1981, **2**, pp. 126–129

3 Khan, N., Hong, Y., Tan Siow, P., Wee, H. Soon, Nandar, S., Yin, H. Wai, Kripesh, V., Lau, J. H., and Kok, C. Toh: '3D packaging with through silicon via (TSV) for electrical and fluidic interconnections', Proc. Electronic Components and Technology Conf., San Diego, CA, USA, 2009, pp. 1153–1158.

4 Brunschwiler, T., Paredes, S., Drechsler, U., Michel, B., Cesar, W., Toral, G., Temiz, Y., and Leblebici, Y.: 'Validation of the porous-medium approach to model interlayer-cooled 3D-chip stacks', Proc. 3D System Integration, San Fransisco, CA, USA, 2009, pp. 1–10

5 King, C. R.Jr., Zaveri, J., Bakir, M. S., and Meindl, J. D.: 'Electrical and fluidic c4 interconnections for inter-layer liquid cooling of 3D ICs', Proc. Electronic Components and Technology Conf., Las Vegas, NV, USA, 2010, pp. 1674–1681.

6 Zhang, Y., King, C. R., Zaveri, J., Jo, K. Yoon, Sahu, V., Joshi, Y., and Bakir, M. S.: 'Coupled electrical and thermal 3D IC centric microfluidic heat sink design and technology', Proc. Electronic Components and Technology Conf., Lake Buena Vista, FL, USA, 2011, pp. 2037–2044.

7 Peles, Y., Kosar, A., Mishra, C., Kuo, C., and Schneider, B.: 'Forced convective heat transfer across a pin fin micro heat sink', *Int. J. Heat Mass Trans.*, 2005, **48**, pp. 3615–3627

8 Sekar, D., King, C., Dang, B., Spencer, T., Thacker, H., Joseph, P., Bakir, M., and Meindl, J.: 'A 3D-IC technology with integrated microchannel cooling', Proc. Interconnect Technology Conf., Burlingame, CA, USA, 2008, pp. 13–15

9 Zhang, Y., Dembla, A., Joshi, Y., and Bakir, M. S.: '3D stacked microfluidic cooling for high-performance 3D ICs', Proc. Electronic Components and Technology Conf., San Diego, CA, USA, 2012, pp. 1644–1650