Tier-Independent Microfluidic Cooling for Heterogeneous 3D ICs with Nonuniform Power Dissipation

Yue Zhang, Li Zheng and Muhannad S. Bakir

School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA
yzhang324@gatech.edu

Abstract

Embedded microfluidic cooling is considered a promising solution for heat removal in 3D ICs. This paper presents tier-independent microfluidic cooling in a 2-tier chip thermal testbed. Each tier has 4 segmented heaters emulating a simplified multicore processor. Tier-independent cooling is shown to reduce the pumping power by 37.5% by preventing over-cooling when an operating temperature is specified. Thermal coupling for 3D chips with liquid cooling is also discussed.

Introduction

Three-dimensional ICs offer opportunities for improving performance and reducing power dissipation by enabling shorter on- and off-chip interconnects and heterogeneous integration [1][2]. However, due to the increased power densities and the high thermal resistance of tiers furthest from the heat sink, heat removal in 3D ICs is challenging when air-cooled heat sink is used [3]. The challenge becomes exacerbated when multiple logic tiers are stacked. Embedded microfluidic cooling (MFC) is considered a potential solution to solve this thermal problem due to its effective cooling capability, microscale form factor, and CMOS microfabrication compatibility [4][5].

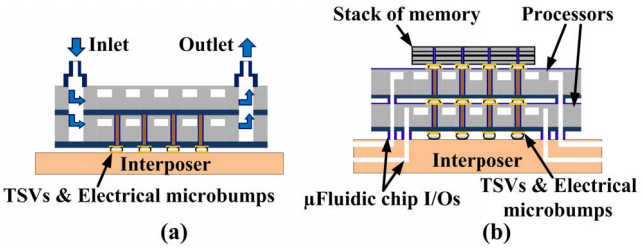

Fig. 1. Topology of (a) a general embedded microfluidic heat sink and (b) our tier-independent microfluidic heat sink within a 3D stack.

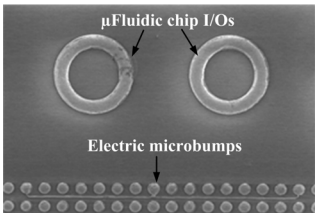

Fig. 2. SEM of solder µFluidic chip I/Os and Electric microbumps.

Prior work in 3D MFC has focused on pumping a coolant into a stack through a single inlet, as shown in Fig. 1 (a) [5][6]. With this approach, it is not possible to control or tailor the flow rate in each tier. However, in a realistic 3D stack, especially in a heterogeneous stack, the power dissipation in each tier may be different (workload dependent). Thus, one needs the capability to control the coolant flow rate in each tier independently. Even more, there is likely a need to control the flow rate locally within a single tier, as discussed later in the paper. To address this need, wafer-level batch fabricated solder microfluidic chip I/Os and fine pitch electrical microbump I/Os have been recently demonstrated, as shown in Fig. 2 [7]. Based on this innovative chip I/O technology, this paper proposes and experimentally implements tier-independent embedded MFC in a 2-tier stack with each tier containing four independent thin-film heaters/thermometers (Fig. 1 (b)). We demonstrate 1) temperature profile of a single chip with uniform heating, 2) thermal coupling between the two stacked tiers with nonuniform heating as a function of cooling in each tier, and 3) reduced pumping power in two tiers with different power dissipations using tier-independent flow rates.

Thermal Testbed and Experimental Setup

The 2-tier chip thermal testbed under consideration utilizes an embedded silicon micropin-fin heat sink in each tier (Fig. 1 (b)). The details of the fabrication process are reported in [8]. The dimensions of the micropin-fins are listed in Fig. 3 [8]. Each tier has 4 independent Pt-based thin-film segmented heaters/thermometers in order to emulate the heating of a simplified multicore microprocessor (Fig. 4). The dimension of each heater is 0.22 cm × 1 cm with a spacing of 0.03 cm between them. The total heating area in each tier is 1 cm × 1 cm. The heaters are independently controlled. Moreover, each tier has its own set of inlet and outlet ports allowing tier-independent control of the flow rate. The two tiers are bonded using a thermal interface material with a thermal resistance of 0.28 K/W.

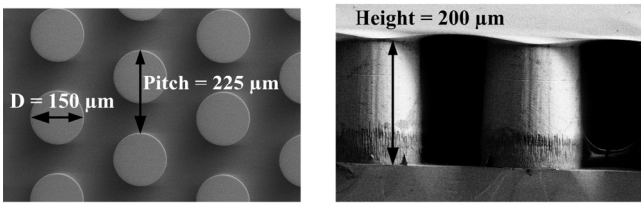

Fig. 3. (a) Top and (b) cross-section views of the silicon micropin fin heat sink in each tier.

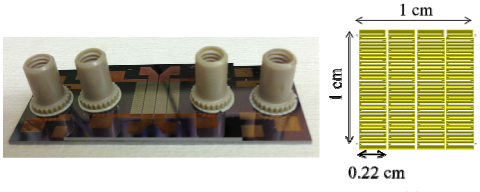

Fig. 4. (a) Micrograph of a 2-tier chip thermal testbed with segmented heaters. (b) Illustration of the segmented heaters on each tier.

978-1-4799-0440-2/13/$31.00 © 2013IEEE

Two gear pumps are connected to each tier to supply DI water. The temperature of the inlet DI water is 20 ± 1 °C. The temperature of each segmented heater is sensed by the on-chip Pt heater/thermometer and recorded at a rate of 1 Hz using an Agilent data logger.

Experiment Procedures and Thermal Results

A. Single tier with uniform heating

To capture the lateral temperature increase as a coolant flows from the inlet to the outlet, a single tier measurement is performed. Fig. 5 illustrates the temperature of each heater on the chip as the total chip power density ramps from 25 W/cm^2 to 100 W/cm^2. The DI water flow rate is 80 mL/min in all of the measurements unless specified otherwise. In the high power density case (100 W/cm^2), the junction temperature of heater 4 (i.e., the heater closest to the outlet) increases by 33 °C while that of location 1 increases only by 17 °C. This result is expected since as the coolant flows from the inlet to the outlet, its temperature increases, and thus, the chip junction temperature [4]. The lateral thermal gradient across the chip becomes exacerbated for higher power densities. One way to mitigate the thermal gradient is to increase the flow rate. But the pressure drop, and thus, the pumping power will increase. Later in the paper, we propose another possible solution based on local coolant delivery.

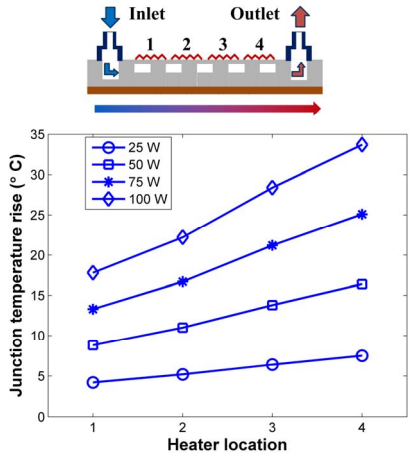

Fig. 5. Junction temperature rise at different locations on the chip as a function of power dissipation.

B. Vertical thermal coupling

Vertical thermal coupling between two tiers with embedded microfluidic heat sink is investigated in this section. In case A and B (Fig. 6), DI water is only pumped into the top tier such that the two tiers share the same microfluidic heat sink. In case A, heaters 1 and 4 in the top tier are each powered up to 25 W. In case B, heaters 2 and 3 in the lower tier are each powered up to 25 W in addition to those heaters in the upper tier. Once the heaters in the lower tier are turned on, as shown in Fig. 7, the junction temperature of heaters 1, 2, 3 and 4 in the upper tier are elevated by 3.3 °C, 6.8 °C, 9.7 °C and 10.1 °C, respectively. In case C, the power dissipation profile in the two tiers is the same as that in case B. The difference is that DI water is pumped into both tiers. Clearly, the temperature of the upper tier (Fig. 7) in case A and case C overlap indicating the impact of the lower tier is minimal. In Fig. 8, the temperature of the lower tier in the three cases is plotted. For case A, the lower tier is idle. However, due to the temperature increase of the coolant, the temperature of heaters 1, 2, 3 and 4 of the lower tier are elevated by 4.9 °C, 5.7 °C, 7.8 °C and 9.7 °C, respectively. Vertical thermal coupling may cause idle tiers to get warmer, leading to unwanted leakage power [9]. To reduce the thermal coupling between tiers in MFC, each tier can have its own microfluidic heat sink (case C) instead of sharing one heat sink (case B).

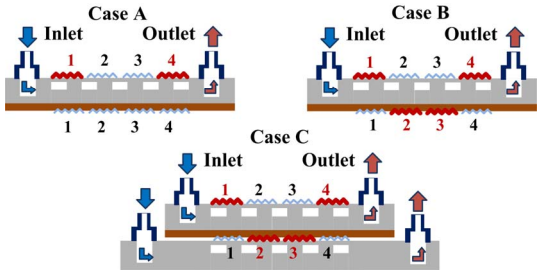

Fig. 6. Vertical thermal coupling test cases. (A) Heaters 1 and 4 in upper tier are powered. (B) Heaters 1 and 4 in upper tier and heaters 2 and 3 in lower tier are powered. (C) Heaters 1 and 4 in upper tier and heaters 2 and 3 in lower tier are powered. DI water is pumped into both tiers in case C.

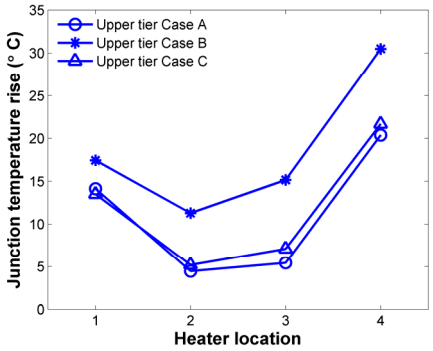

Fig. 7. The junction temperature rise of the upper tier at different heater locations on the chip for the cases shown in Fig. 6.

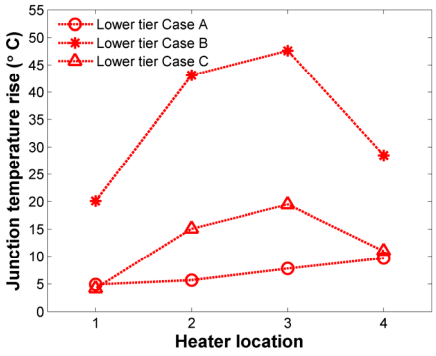

Fig. 8. The junction temperature rise of the lower tier at different locations on the chip for the cases shown in Fig. 6.

C. Tier-independent flow rates in ICs with different power dissipations

In this section, we experimentally evaluate a 3D stack of two high-power chips with different power densities: 50 W/cm^2 (P1) and 100 W/cm^2 (P2). The tier-independent flow rate (and thus cooling) that we proposed is implemented. As shown in Fig. 9, when each tier in the

stack is initially cooled under the same flow rate (Q1 = Q2 = 45 mL/min), the average junction temperatures of P1 and P2 are 49.5 °C and 60.6 °C, respectively. Next, the flow rate of each tier is varied independently so that the junction temperatures of the two tiers are equalized at the higher and lower ends. In the case where flow rates Q1 and Q2 are 32 mL/min and 116 mL/min, respectively, the junction temperature of the two tiers is equalized at approximately 49.5 °C. By mitigating the thermal gradient of the two tiers, thermomechanical stress and thermal induced variations are lowered.

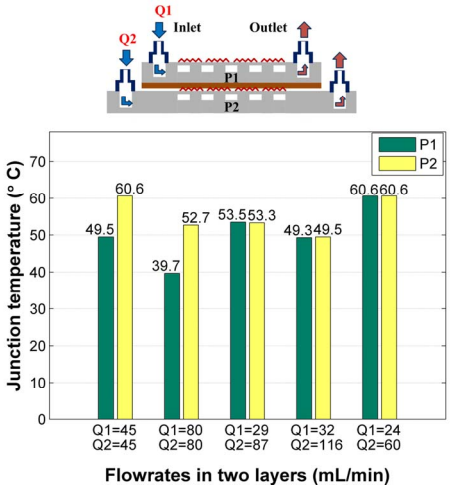

Fig. 9. Junction temperature of P1 and P2 as a function of the flow rates.

Additionally, when an operating temperature is specified, adjusting the flow rate according to the power dissipation saves pumping power by preventing over-cooling. Considering the conventional microfluidic delivery method (i.e., Fig. 1 (a)) in which the flow rate in each tier has to be identical, the total flow rate needed is chosen based on the thermal needs of the tier with the highest power. The conventional method is emulated as the second set of flow rates in Fig. 9. For example, for an operating temperature of 53 °C, Q1 and Q2 need to be 80 mL/min. In our tier-independent cooling (the third set of flow rates in Fig. 9), the needed Q1 and Q2 are 29 mL/min and 87 mL/min, respectively. The pressure drops at 29 mL/min, 80 mL/min and 87 mL/min are measured to be 12 kPa, 60 kPa and 67.9 kPa, respectively. As a result, using tier-independent flow rate, the pumping power is reduced by 37.5% relative to the conventional fluidic delivery method.

Electrical Implications

The main contributor to leakage current is subthreshold current [10]. Based on the leakage current model proposed in [10] and assuming a constant supply voltage, the leakage power is calculated for a single tier with uniform power and lateral thermal gradient due to fluidic coolant flow from the inlet to the outlet (i.e., as discussed in Fig. 5). The leakage power is then normalized to the leakage power at room temperature (25 °C). Junction temperature and the normalized leakage power ($P_{leaknorm}$) of the four cores are listed in Table I for a uniform power density of 100 W/cm^2. $P_{leaknorm}$ increases gradually along the direction of the cooling fluid. In a multicore chip, if all the cores have fixed power budget, the allowable dynamic power decreases along the flow direction. The cores closest to the outlet may operate at a lower frequency than cores closest to the inlet. In order to allow all the cores to work symmetrically, localized coolant delivery to each core using the microscale fluidic chip I/Os is proposed (Fig. 10 (a)). Solder based microfluidic chip I/Os with an outer diameter of 210 µm, an inner diameter of 150 µm, and a height of 12 µm are shown in Fig. 10 (b). The microfluidic chip I/Os have been experimentally shown to withstand a pressure drop of 100 kPa without leakage. These microfluidic chip I/Os are fabricated in parallel to electrical microbumps with a density of 40,000/cm^2 (microbump pitch of 50 µm), which is critical to power delivery and high-bandwidth off-chip signaling.

Table I. Junction temperature and normalized leakage power of the four measured heaters (cores) in Fig. 5.

Uniform power	Core	1	2	3	4
	T (°C)	36.9	41.3	47.5	52.8
	$P_{leaknorm}$	1.4	1.5	1.8	2

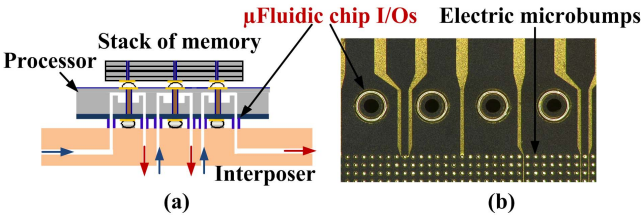

Fig. 10. (a) Prototype of 3D stack with microfluidic chip I/Os for localized coolant delivery and (b) top view of solder based microfluidic chip I/Os and electric microbumps.

Conclusion

This paper proposes and demonstrates tier-independent MFC technology for high performance 3D ICs. The tier-independent cooling approach is shown to equalize the temperatures of two tiers that dissipate different power densities (50 W/cm^2 and 100 W/cm^2), which may be needed in some applications. Tier-independent cooling is also shown to reduce the pumping power by 37.5% by preventing over-cooling when an operating temperature is specified. The lateral thermal gradient induced by coolant temperature increases along the flow direction causing the leakage power to increase for the downstream cores. The use of localized cooling enabled through microfluidic chip I/Os is a promising solution to eliminate this thermal effect.

Acknowledgement

This work has been carried out in part under support from DARPA.

References

[1] S.J. Souri et al., *DAC*, 2000.
[2] K. Banejee et al., *Proc. of the IEEE*, 2001
[3] S. M. Sri-Jayantha et al., *IBM Journal of Research and Development*, 2008
[4] D. B. Tuckerman and R. F. Pease, *EDL*, 1981.
[5] T. Brunschwiler et al., *3D System Integration*, 2009
[6] N. Khan et al., *ECTC*, 2009
[7] L. Zheng and M. S. Bakir, *ECTC*, 2013.
[8] Y. Zhang et al., *ECTC*, 2011
[9] D. Sekar et al., *ECTC*, 2008
[10] K. Shakeri and J. D. Meindl, *ISVLSI*, 2002