3-D Stacked Tier-Specific Microfluidic Cooling for Heterogeneous 3-D ICs

Yue Zhang, *Student Member, IEEE*, Li Zheng, and Muhannad S. Bakir, *Senior Member, IEEE*

***Abstract*— Cooling is a significant challenge for high-performance and high-power 3-D ICs. In this paper, tier-specific microfluidic cooling technology is proposed and experimentally evaluated in a 3-D stack. Different stacking scenarios are experimentally evaluated including: 1) a memory-on-processor stack; 2) a processor-on-processor stack with equal power dissipation; 3) a processor-on-processor stack with different power dissipations; and 4) a two-tier 3-D stack with each tier containing four cores along the flow direction. Compared with conventional microfluidic cooling, the tier-specific cooling is shown to reduce the pumping power by 37.5% by preventing overcooling when an operating temperature is specified. The results are benchmarked with an air-cooled heat sink. The impact of the lateral thermal gradient along the flow direction on the electrical performances, including leakage power and clock frequency, is analyzed.**

***Index Terms*— 3-D IC, leakage power, microchannel heat sink, micropin-fin (MPF) heat sink, multicore processors.**

I. Introduction

With continued aggressive CMOS scaling, interconnect performance and power dissipation have become limiting factors for high-performance integrated circuits [1]–[3]. 3-D ICs offer new opportunities for improving chip performance and reducing power dissipation by enabling shorter interconnects (both on- and off-chip) as well as the possibility of heterogeneous integration. However, a number of challenges must be overcome before 3-D ICs can be adopted for high-performance and high-power applications [4]–[6]. Cooling is a key issue for 3-D ICs since both the power dissipation per unit area and the thermal resistance for the dice within the stack increase with the number of tiers. For reference, a few ITRS projections of interest are shown in Fig. 1 [7]. To address the challenges in heat removal, innovative cooling solutions have been proposed, including single-phase forced microfluidic cooling [8]–[11], two-phase microfluidic cooling [12], [13], and active thermoelectric coolers to address hotspots [14], [15]. This paper focuses on integrated single-phase microfluidic cooling in 3-D ICs. Advantages and disadvantages of conventional air cooling, prior microfluidic cooling technology in the literature, and the proposed tier-specific microfluidic cooling are summarized as follows.

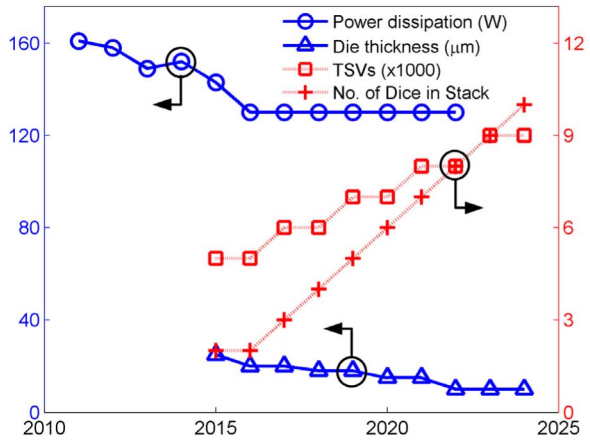


Fig. 1. ITRS projections for the number of dice in a stack, number of through-silicon vias (TSVs), die thickness, and power of a single high-performance chip.

A. Air-Cooled Heat Sink

Adopting an air-cooled heat sink (ACHS) to reject heat from a 3-D stack is simple [Fig. 2(a)], but it has limited cooling capability [16]. In a memory-processor stack, the processor chip should be placed next to the heat sink to have the lowest thermal resistance. However, placing the processor away from the package substrate requires possibly a large number of power and ground interconnects that need to go through the memory tier. A single processor requires a few thousand power and signaling interconnects [7]. This large number of interconnects have to be routed through the memory chip, which affects memory design, density, and performance. The memory performance is also impacted by the thermal crosstalk between the two tiers. An additional point of value is the fact that an ACHS (and its heat spreader) requires large lateral footprint, which limits how close two chips (or stacks of chips) can be placed laterally if each has its own heat sink. This clearly would impact off-chip interconnect length and thus energy dissipation and bandwidth density.

B. Embedded Microfluidic Heat Sink in the Literature

Due to the limitations of ACHS, there has been much research effort to investigate the use of integrated microfluidic heat sinks (MFHSs) to reject heat from a 3-D stack. Fig. 2(b)

Manuscript received April 25, 2013; revised July 24, 2013; accepted July 29, 2013. Date of publication September 30, 2013; date of current version October 28, 2013. This work has been carried out in part under support from Defense Advanced Research Projects Agency (DARPA) with Grant N66001-12-1-4240 and from the Interconnect Focus Center and the Semiconductor Research Center under contract 2011-KJ-2189. Recommended for publication by Associate Editor S. Ankireddi upon evaluation of reviewers' comments.

The authors are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: yzhang324@gatech.edu; lizheng@gatech.edu; muhannad.bakir@mirc.gatech.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCPMT.2013.2281605

2156-3950 © 2013 IEEE

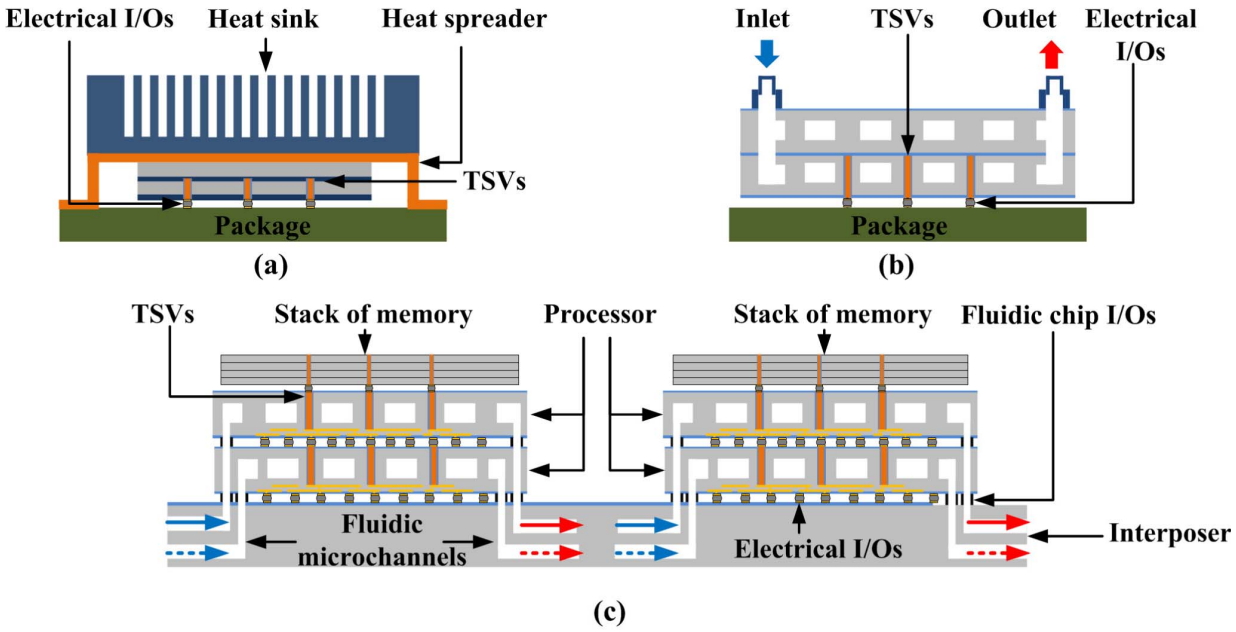


Fig. 2. Illustration of (a) conventional air cooling technology, (b) prior integrated microfluidic cooling technology, and (c) tier-specific microfluidic cooling technology in this paper.

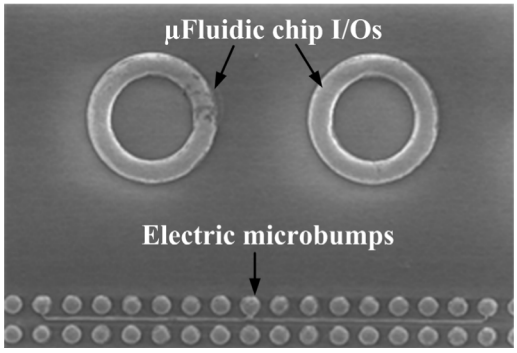


Fig. 3. SEM of solder microfluidic chip I/Os and electric microbumps.

shows a typical chip configuration with embedded MFHS in the literature, in which the fluid is supplied through a single inlet from the top of the stack [8], [10]. References [8] and [10] demonstrated the cooling of a four-tier and a two-tier stack with total power dissipation of 390 and 200 W, respectively. With this approach, it is not possible to control or tailor the flow rate in each tier. However, in a realistic 3-D 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 in Section VI. 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. 3 [17]. With this innovative chip I/O technology, this paper proposes and experimentally demonstrates a tier-specific MFHS in a two-tier stack.

C. Tier-Specific MFHS

Fig. 2(c) shows our vision for the tier-specific MFHS in a heterogeneous high-performance 3-D IC system. The proposed 3-D IC system features a silicon interposer with embedded fluidic delivery microchannels and an array of 3-D stacked processor and memory tiers. Each of the processor tiers contains an embedded MFHS. TSVs are routed through the integrated MFHS. Each tier has its dedicated microfluidic chip I/Os (formed using either solder [18] or polymer [19]) for fluid delivery from the interposer. The coolant flow rate in each tier can be tailored independently, according to the heat dissipation of each tier, i.e., tier-specific cooling. This approach helps to minimize the vertical thermal gradient across the stack when power dissipation varies in the stack. Pumping power may be reduced by adjusting the flow rate to the needed value for a given power dissipation per tier. The proposed local coolant delivery mechanism, which is also based on the solder chip I/O technology (discussed in Section VI), may minimize the lateral thermal gradient within a single tier as well. Moreover, since the MFHS is chip scale, this approach allows high lateral scalability of the electronic components, i.e., placing an array of 3-D ICs laterally next to each other. This greatly enhances off-chip connectivity.

II. Thermal Testbed and Experimental Setup

A. Fabrication of the Thermal Testbed

The MFHS uses a staggered micropin-fin (MPF) array, which is designed to be compatible with TSVs [20]. Fig. 4 shows the fabrication steps of a 3-D thermal testbed with integrated MPF heat sink. The fabrication starts with a double-side polished silicon wafer with a 2-μm-thick PECVD SiO_2 film at the bottom. Using the standard Bosch process, which alternates between SF_6 (plasma etch step) and inert C_4F_8

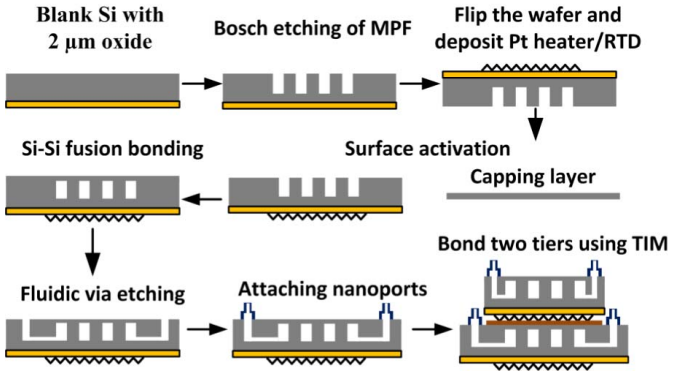


Fig. 4. Fabrication process flow of the two-tier thermal testbed.

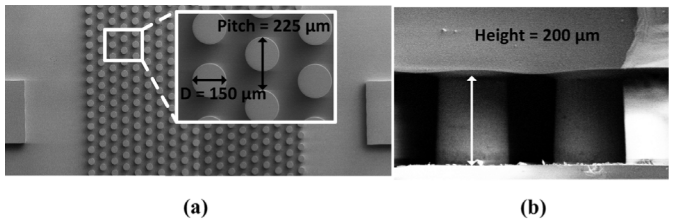


Fig. 5. SEM images of (a) overall and zoomed-in views of MFHS and (b) cross-sectional view of bonded MFHS.

(passivation step), 200-μm ($\pm$2 μm) deep MPF array is etched. The diameter of a single MPF is 150 μm, and the pitch of the MPF is 225 μm [20]. SEM images of an overall view and a magnified view of the MPFs are shown in Fig. 5(a). After the MPFs are etched, the wafer is flipped over and thin-film platinum (Pt) heater, to emulate the heating of a microprocessor, is sputter coated on the backside of the wafer. Due to the linear resistance–temperature relationship and its chemical inertness, the Pt heater also serves as a resistance thermal detector (RTD) during the thermal measurements. Next, silicon-to-silicon fusion bonding is performed to encapsulate the microfluidic channels. Initially, oxygen plasma is used to activate the silicon surface and enable the bonding of two silicon wafers at room temperature. Annealing at 400 °C increases the bonding strength by forming Si–O–Si bonds [21]. Fig. 5(b) shows a cross-sectional SEM image of the bonded MFHS. Fluidic vias are then etched to enable liquid circulation into and out of the MFHS. Nanoports (from Idex Health & Science) are attached to the fluidic vias through adhesive pad for providing consistent fluid connections into the testbed. The final step in preparing the testbed is to bond two tiers using a thermal interface material (TIM) with a thermal resistance of 0.25–0.28 K/W (depending on the pressure applied during the experiments).

To attain an initial insight into the benefits of the embedded microfluidic cooling, a 3-D ACHS testbed is constructed similarly without the embedded MFHS. This will be used to benchmark the thermal results of the MFHS cooled chip.

B. Thermal Testbed Features and Experimental Setup

Three types of thermal testbeds (fabricated as described in Section II-A) are used in this paper. All the microfluidic

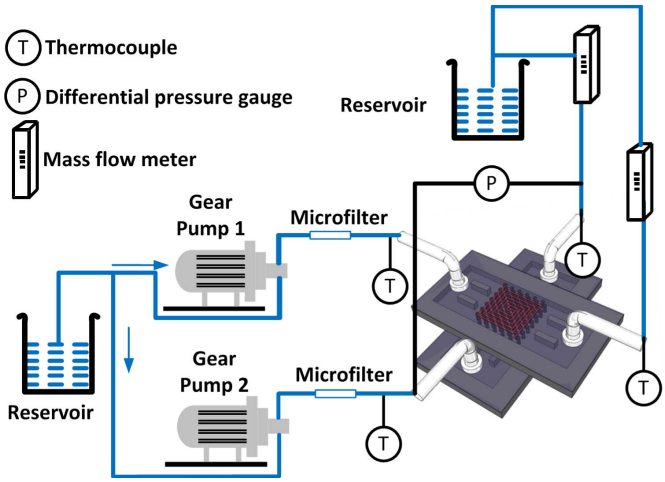


Fig. 6. Schematic diagram of the microfluidic test setup for Testbed 1.

testbeds contain two tiers with embedded MFHS. Each tier has its own set of inlet and outlet ports allowing tier-specific control of the flow rate. Each tier has Pt thin-film heater/RTD to emulate the heating of a microprocessor. In Testbed 1, the heating area is formed by a single Pt heater/RTD with dimensions of 1 cm $\times$ 1 cm. Testbed 1 is built to emulate different 3-D stacking scenarios cooled using MFHS. As a benchmark, a 3-D ACHS testbed (Testbed 2) (fabricated as discussed in Section II-A) was also made. In the ACHS testbed, a high-performance ACHS containing three copper heat pipes and 45 aluminum fins designed for the Intel i5/i7 CPU is attached to the backside of the ACHS 3-D chip stack through a TIM. The ACHS chip is tested while the fan rotates at its maximum speed (2500 r/m $\pm$ 15%). The corresponding air flow is 54.8 CFM. The same 3-D stacking scenarios are emulated using both Testbeds 1 (MFHS) and 2 (ACHS). The benefits of MFHS are reported in Sections III-A, B, and C.

It is well known that the coolant temperature increases as it passes through the MFHS, and thus the chip temperature will increase. To capture the lateral chip temperature gradient and emulate the stacking of multicore processors, Testbed 3 is constructed. Each tier features four independently controlled segmented heaters/RTDs along the flow direction. The dimensions of each heater are 0.22 cm $\times$ 1 cm with a spacing of 0.03 cm. The total heating area of each tier is 1 cm $\times$ 1 cm. The related results are reported in Sections IV-A and B.

In the microfluidic experimental setup for Testbed 1 (Fig. 6), two pumps are connected to the two inlets in the stack (i.e., each tier has its own inlet and pump). Deionized (DI) water is pumped from a nearby reservoir. Polyester-based filters are connected to the outlet of the pump for eliminating any particles ($>$20 μm) that may potentially block the MFHS. An acrylic block flow meter that measures up to 100 mL/min is connected to each inlet serially for measuring the flow rate. For the sake of simplified port access, the two tiers are stacked orthogonally such that the inlets and outlets are easily accessible (Fig. 6). An Agilent N6705B power analyzer with four outputs is used to source current to the thin-film Pt heaters/RTDs for emulating chip power dissipation [16].

TABLE I
SUMMARY OF SOME REPRESENTATIVE DATA POINTS FOR MICROFLUIDIC COOLING AND AIR COOLING

Emulated scenarios	Cooling method	Power density (W/cm^2)		Flow rate (mL/min)		Junction temperature rise (°C)	
		BTM	TOP	BTM	TOP	BTM	TOP
Memory-on-processor stack (III.A)	Microfluidic	5	49.7	0	100	9.9	14
		5	99.2	0	100	15.3	28.7
	Air	5	57.1	-	-	38.9	36.7
		49.3	5	-	-	57.0	28.4
Processor-on-processor stack (III.B)	Microfluidic	49	49.6	100	100	14.2	14.6
		98.3	99.2	100	100	28.7	30.2
	Air	45.6	48.4	-	-	73.6	54.4
Two tiers with different power dissipations (III.C)	Q1=Q2	100	50	80	80	39.7	52.7
	Q1≠Q2	100	50	29	87	53.7	53.5

*BTM=bottom; Q1 and Q2 are flow rates in the top and bottom tier respectively. Room temperature is 21.4 °C.

The heater resistance in each tier is measured and tracked using an Agilent 34970A data logger at 1 Hz. In Testbeds 1 and 2, the measured resistance, and thus junction temperature, represents the average junction temperature in each tier since we use a single heater/RTD per tier. In Testbed 3, where four segmented heaters are used, the measured temperature represents the average junction temperature of each of the four zones. The experimental setup for Testbed 3 is similar to that shown in Fig. 6 with the difference being the two tiers are stacked in parallel.

III. MICROFLUIDIC COOLING BENCHMARKED WITH ACHS

The thermal experiments begin with the characterization of the Pt heater/RTD. For various fabricated Pt heaters/RTDs used in this paper, the temperature coefficient of resistance varies in the range 0.0025–0.003 K^{-1}. These results show good consistency. The experimental results reported in this section are to investigate and compare an air-cooled 3-D stack with a microfluidic-cooled 3-D stack under different stacking (heating) scenarios. The temperature of the inlet DI water is 20 ± 1 °C. This temperature is used to compute the junction temperature rises. For the convenience of reader, the key experimental results that are to be discussed in the following sections are summarized in Table I.

A. Memory-on-Processor Stack

In Fig. 7(a), the fluid is pumped only into the processor tier of the memory-processor stack at a flow rate of 100 ± 5 mL/min. In this experiment, the power density of the memory chip is held at 5 W/cm^2. Since the memory tier is stacked on the processor tier with integrated microfluidic-cooled heat sink, the MFHS serves as a path for cooling of the memory tier as well. The junction temperature rise of the memory and processor tiers are 15.3 °C and 28.7 °C, respectively, when the power density of the processor tier is 99.2 W/cm^2. As a comparison, a memory-processor stack is

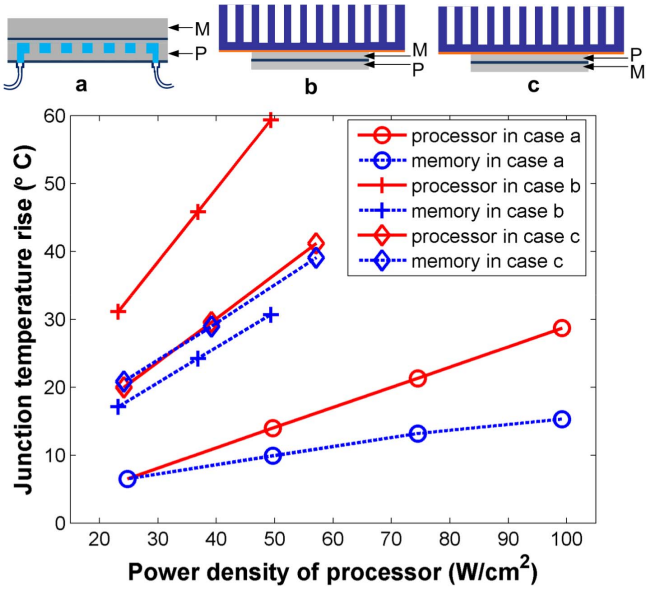


Fig. 7. Junction temperature rise in a memory-processor stack under MFHS and ACHS.

tested under ACHS [Fig. 7(b) and (c)]. For the case where memory is placed close to the ACHS [Fig. 7(b)], the junction temperature rise of the memory and processor tiers are 30.6 °C and 59.3 °C, respectively, when the power density of the processor is 49.3 W/ cm^2 [22]. For the case where processor is placed close to the ACHS [Fig. 7(c)], the junction temperature rise of the memory and processor tiers are 39 °C and 41.1 °C when power density of the processor tier is 57.1 W/ cm^2. For the same power density, the absolute junction temperatures of the chips under MFHS are lower than those under ACHS by at least 12 °C and by 48 °C in the worst case. In the ACHS experiments, due to the overheating of the chips, the power densities of the two tiers are limited to <60 W/cm^2.

B. Processor-on-Processor Stack

In Fig. 8, the two-tier chip stack (Testbed 1) dissipates up to 100 W/cm^2 per tier to emulate the stacking of processors. An

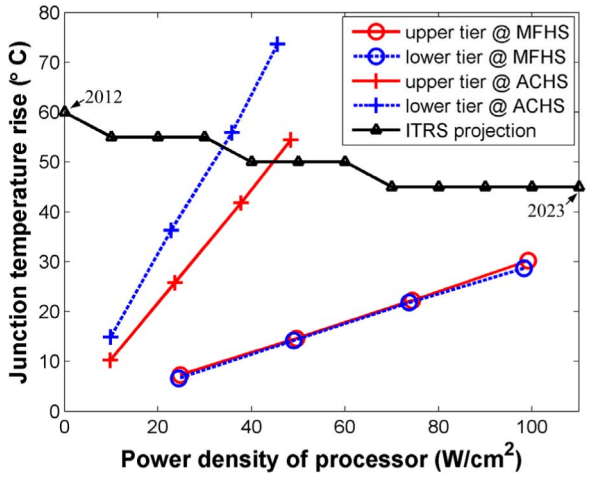


Fig. 8. Junction temperature rise in a processor-on-processor stack under MFHS and ACHS.

MFHS is integrated into each tier. The flow rate in each tier is 100 mL/min. Two sets of measurements were performed for the MFHS cooled stack, and the average junction temperature rise above the inlet coolant temperature in each tier is shown in Fig. 8. The difference in the two measurements did not exceed 1.1 °C. As seen from the plots, when the power density in each tier is 100 W/cm^2, the junction temperature rise in either tier is 30 °C, resulting in an absolute junction temperature of 50 °C. In contrast, the testbed under ACHS has a temperature rise of $>$54 °C at 50 W/cm^2. The maximum junction temperature rise trend according to ITRS is also shown in Fig. 8 as a reference. The processor-on-processor stack cooled using MFHS can dissipate $>$100 W/cm^2 in each tier without reaching the ITRS projected maximum junction temperature. Please note that the thermal results obtained by ACHS testbed may have been better if a better TIM is used.

C. Tier-Specific Flow Rates in ICs With Different Power Dissipations

In this section, we experimentally evaluate a 3-D stack of two high-power tiers with different power densities: 50 W/cm^2 ($P1$) and 100 W/cm^2 ($P2$). The tier-specific flow rate (and thus cooling) that we proposed in Section I-C 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 temperature of $P1$ and $P2$ is 49.5 °C and 60.6 °C, respectively. Next, the flow rate of each tier is varied independently so that the junction temperature of the two tiers is equalized at the higher and lower ends. For example, in the case where flow rates $Q1$ and $Q2$ are 32 and 116 mL/min, respectively, the junction temperature of the two tiers is equalized at ~49.5 °C. By mitigating the thermal gradient of the two tiers, thermomechanical stress and thermal-induced variations are lowered. In addition, when an operating temperature is specified, adjusting the flow rate according to the power dissipation saves pumping power by preventing overcooling. Considering the conventional microfluidic delivery method [i.e., Fig. 2(b)] in which the flow rate in each tier has to be

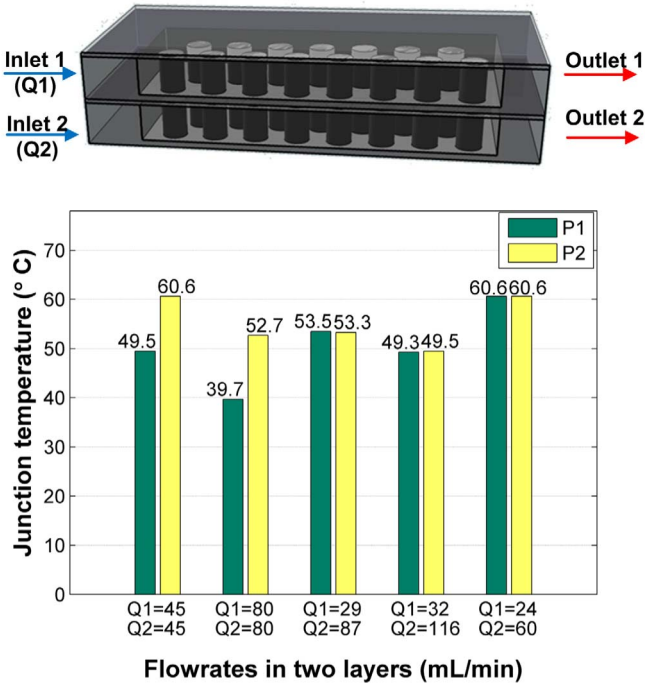


Fig. 9. Junction temperature of $P1$ and $P2$ as a function of the flow rates.

identical, the total flow rate 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 to maintain both tiers at a temperature $<$53 °C. In our tier-specific cooling (the third set of flow rates in Fig. 9), the needed $Q1$ and $Q2$ are 29 and 87 mL/min, respectively. The pressure drops at 29, 80, and 87 mL/min are measured to be 12, 60, and 67.9 kPa, respectively. As a result, using tier-specific flow rate, the pumping power is reduced by 37.5% relative to the conventional fluidic delivery method.

IV. Microfluidic Cooling in Multicore Processor Stacking

A. Single Tier With Uniform Power Dissipation

To capture the lateral temperature increase as a coolant flows from the inlet to the outlet, a single-tier measurement is performed using Testbed 3, which contains four segmented heaters along the flow direction. Fig. 10 shows the temperature of each heater on the chip as the total chip power density ramps from 25 to 100 W/cm^2. The DI water flow rate is 80 mL/min in all of the measurements in this section 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 heater 1 increases by only 17 °C. This result is expected since the coolant temperature increases as it flows from the inlet to the outlet, and thus the chip junction temperature also increases. The chip design was simulated using ANSYS Fluent at a power density of 100 W/cm^2. Since the MFHS structure is symmetric, only half of the MPF array is modeled. Figs. 11 and 12 show the temperature maps of the base and coolant, respectively.

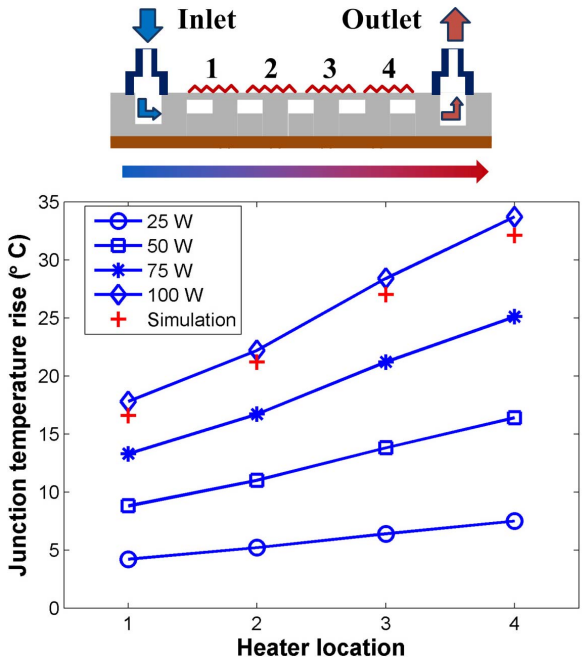


Fig. 10. Junction temperature rise at different heater locations on the chip for different power dissipations. ANSYS simulation for 100-W case is also plotted for reference.

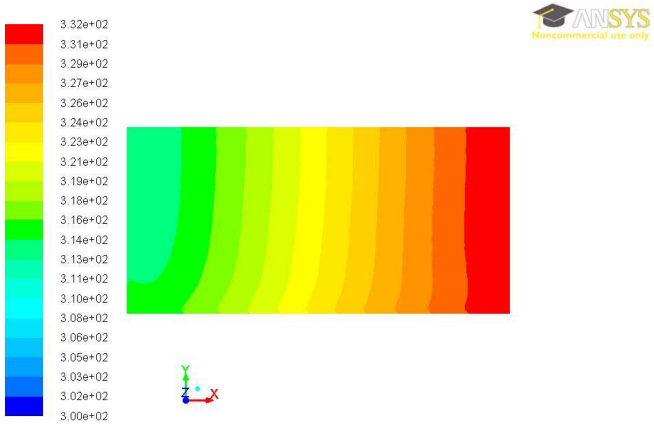


Fig. 11. Base temperature map in ANSYS simulation at 100 W/cm^2.

Average junction temperatures are extracted from the simulation results and are also plotted in Fig. 10. The difference between the experimental results and the simulations is $<$1.6 °C. 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. However, the pressure drop and the pumping power will increase. In Section VI, we propose an alternative solution to mitigate the lateral thermal gradient based on local coolant delivery.

B. Vertical Thermal Coupling

Vertical thermal coupling between two tiers with embedded MFHS is investigated in this section. In Cases A and B (Fig. 13), DI water is only pumped into the top tier such that the two tiers share the same MFHS. In Case A, heaters 1 and 4 of the top tier are each powered up to 25 W. In Case B,

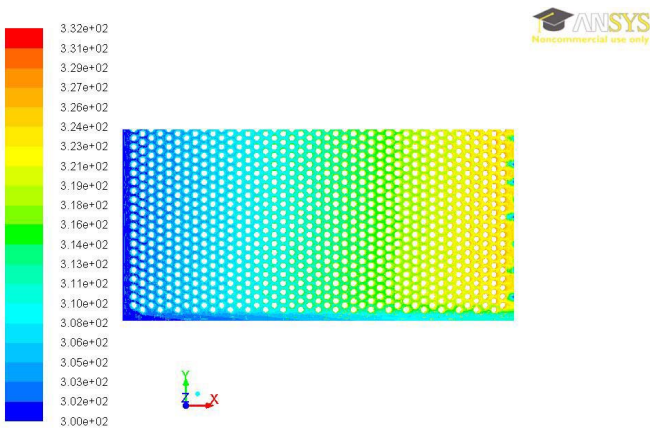


Fig. 12. Fluid temperature map in ANSYS simulation at 100 W/cm^2.

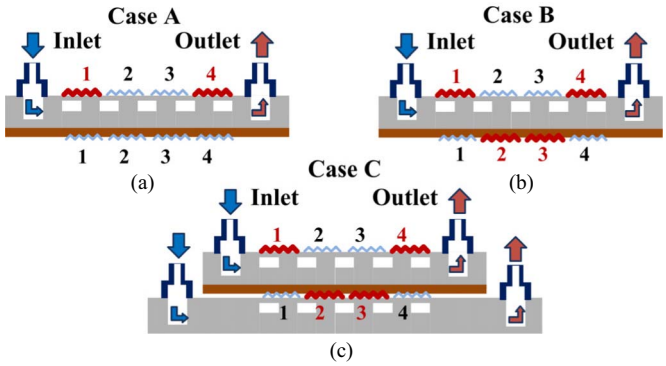


Fig. 13. Vertical thermal coupling test cases. (a) Heaters 1 and 4 in upper tier are powered. (b) Heaters 1 and 4 in upper tier and heaters 2 and 3 in lower tier are powered. (c) Heaters 1 and 4 in upper tier and heaters 2 and 3 in lower tier are powered. DI water is pumped into both tiers in Case C.

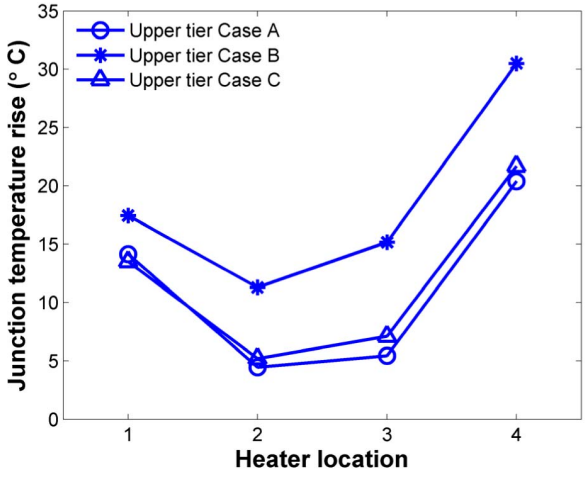


Fig. 14. Junction temperature rise of the upper tier at different heater locations on the chip for the cases shown in Fig. 13.

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. 14, the junction temperature of heaters 1–4 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

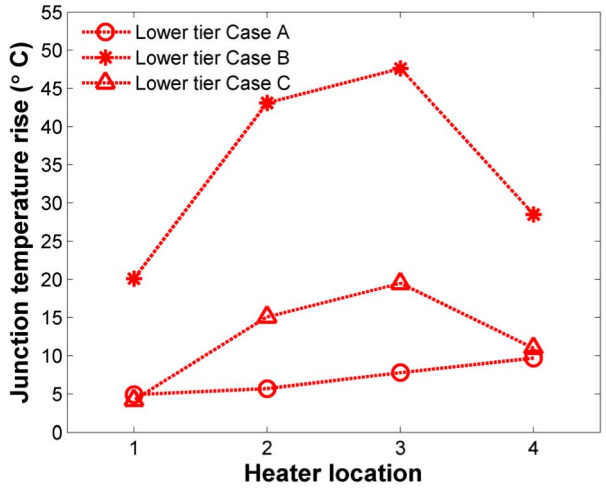


Fig. 15. Junction temperature rise of the lower tier at different heater locations on the chip for the cases shown in Fig. 13.

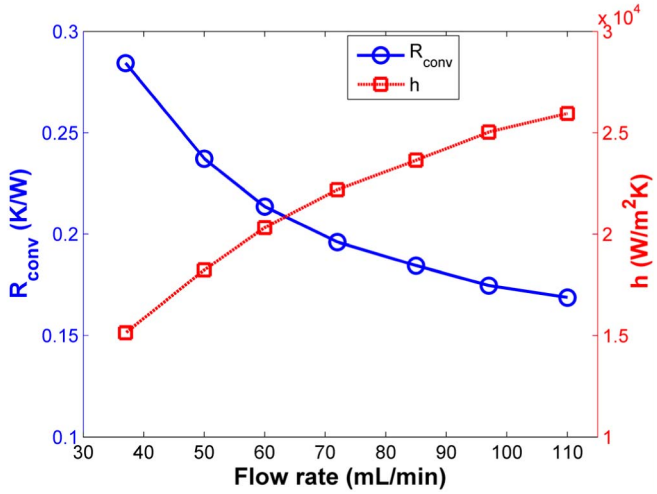


Fig. 16. Convective thermal resistance as a function of the flow rate.

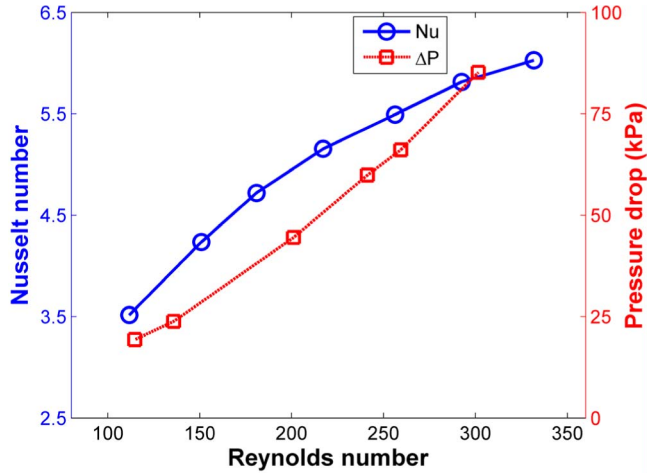


Fig. 17. Nusselt number and pressure drop as a function of Reynolds number.

that in Case B. The difference is that DI water is pumped into both tiers. Clearly, the temperature of the upper tier (Fig. 14) in Cases A and C overlaps, suggesting the impact of the lower tier is minimal. In Case C, embedding an MFHS in the bottom tier provides a heat flow path with a lower thermal resistance. This would greatly diminish the heat transfer to the upper tier. In Fig. 15, 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 1–4 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 [23]. To reduce the vertical thermal coupling between tiers in microfluidic cooling, each tier can have its own MFHS (Case C) instead of sharing one heat sink (Case B).

V. Data Analysis

In this section, single-tier measurements are presented. DI water with different flow rates is pumped into the tier. The power density (P) is kept at 40 W/cm^2. DI water inlet temperature (T_{in}), outlet temperature (T_{out}), and chip junction temperatures (T_j) are monitored and used to calculate the convective thermal resistance (R_{conv}) of the MFHS using [24]

$$R_{\text{conv}} = (T_j - T_f)/P - R_{\text{cond}} \tag{1}$$

where R_{cond} is the conductance from the circuit through the silicon base to the heat sink interface given by (2); T_f is the average fluid temperature calculated by (3)

$$R_{\text{cond}} = \frac{t_{\text{base}}}{k_{\text{si}} A_{\text{base}}} + \frac{t_{\text{ox}}}{k_{\text{ox}} A_{\text{base}}} \tag{2}$$

$$T_f = \frac{1}{2} \times (T_{\text{in}} + T_{\text{out}}). \tag{3}$$

Since R_{cond} is dependent on the thickness of the silicon base (t_{base}) and silicon dioxide (t_{ox}), the area of the base (A_{base}), and the thermal conductivities of silicon (k_{si}) and silicon dioxide (k_{ox}), it is a constant throughout the experiments and the value is calculated to be 0.05 K/W. Fig. 16 plots R_{conv} as a function of flow rate. The heat transfer coefficient (h) is derived from using R_{conv} and the effective total heat transfer area (A_t) by

$$h = \frac{1}{R_{\text{conv}} \times A_t} \tag{4}$$

$$A_t = A_b + \eta \times A_{\text{fin}} \tag{5}$$

where A_b is the base area exposed to the fluid, η is the fin efficiency and is a function of the MPF height (H_{fin}) and diameter (D) given by (6), and A_{fin} is the aggregate surface area of the MPFs exposed to the fluid [20]. Furthermore, Nusselt number (Nu) as a function of Reynolds number (Re) is plotted in Fig. 17. Nusselt number and Reynolds number are calculated using (7) and (8), respectively, based on the hydraulic diameter (D_h) [9]

$$\eta = \frac{\tanh(2H_{\text{fin}}\sqrt{h/k_{\text{si}}D})}{2H_{\text{fin}}\sqrt{h/k_{\text{si}}D}} \tag{6}$$

$$Nu = (h \cdot D_h)/k \tag{7}$$

$$Re = V_{\max} \times D_h/\upsilon \tag{8}$$

$$D_h = \frac{2 \cdot H_{\text{fin}} \cdot w_c}{H_{\text{fin}} + w_c} \tag{9}$$

$$V_{\max} = \frac{Q}{H_{\text{fin}}(W - n \cdot D)} \tag{10}$$

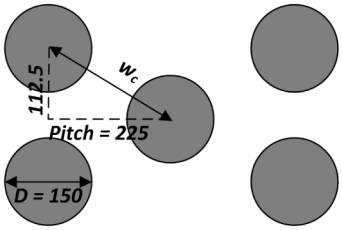


Fig. 18. MPF layout and dimensions in micrometers.

where k and υ are the thermal conductivity and the kinematic viscosity of the fluid, respectively. The hydraulic diameter (D_h) is calculated using (9), where w_c is the diagonal pitch (shown in Fig. 18) [24]. The maximum velocity ($V_{\max}$) crossing the minimum cross section is calculated using (10), where Q is the volumetric velocity, W is the width of the MPF heat sink, and n is the number of the MPFs in the vertical direction (shown in Fig. 18). Although increasing flow rate (Reynolds number) decreases R_{conv} and increases h, the pressure drop (ΔP) will increase, which is not desirable in electronic systems since it increases the pumping power and may introduce reliability issues. The measured pressure drop data for different Reynolds number is also plotted in Fig. 17 to show the tradeoff between the heat transfer characteristic and the pressure drop.

VI. Electrical Implications

Processor performance and power dissipation are a function of temperature. Leakage power plays a significant role in the total power dissipation. This section discusses the leakage power for different cores along the coolant flow direction. The main contributor to leakage current is subthreshold current (I_{sub}) given by [23]

$$I_{\text{sub}} = \frac{W_{\text{eff}}}{L_{\text{eff}}}\mu(T)C_{\text{OX}}(V_T)^2 e^{\left(\frac{V_{\text{GS}}-V_{\text{TH}}}{nV_T}\right)} \tag{11}$$

where W_{eff} and L_{eff} are the effective channel width and length of the transistor. C_{OX} is the gate capacitance of a single transistor, $\mu(T)$ is the carrier mobility, and n is the subthreshold slope factor, which is assumed to be 1.5. V_T is the thermal voltage; V_{GS} and V_{TH} are the gate-source current and the threshold voltage, respectively. With (11) [23] and assuming the supply voltage does not change dynamically, the leakage power is calculated for a single tier with uniform power and lateral thermal gradient due to fluidic coolant temperature gradient from the inlet to the outlet (i.e., as discussed in Fig. 10). 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 II for a uniform power density of 100 W/cm^2. P_{leaknorm} increases gradually along the direction of the cooling fluid. As shown in Table II, leakage power of core 4 is 43% more than that of core 1 for a chip spanning 1 cm × 1 cm. This ratio will increase as the chip size grows. Assuming the chip spans 2.5 cm × 2.5 cm and contains 10 cores in the flow direction, the leakage power of core 10 will be 2.67 times of that of core 1. 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.

TABLE II
Junction Temperature and Normalized Leakage Power of the Four Measured Heaters (Cores) in Fig. 10

	Core 1	Core2	Core 3	Core 4
Junction temperature (°C)	36.9	41.3	47.5	52.8
Normalized leakage power	1.4	1.5	1.8	2

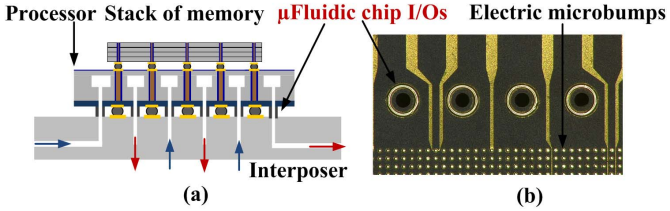


Fig. 19. (a) Prototype of 3-D stack with microfluidic chip I/Os for localized coolant delivery. (b) Top view of solder-based microfluidic chip I/Os and electric microbumps.

To allow all the cores to work symmetrically, localized coolant delivery to each core using the microscale fluidic chip I/Os is proposed [Fig. 19(a)]. This enables the delivery of fresh coolant to each core or cluster of cores at the inlet temperature irrespective of core location. 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. 19(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.

VII. Conclusion

In this paper, we report and compare the experimental results for different 3-D stacking scenarios cooled with microfluidic and air cooling technologies. Based on the thermal testbed configurations investigated in this paper, the key results are summarized as follows:

1) The MFHS maintains the stack temperature <50 °C for a total power density of 200 W/cm^2 in a two-tier processor-on-processor stack.
2) MFHS-based tier-specific cooling is shown to equalize the temperature in two tiers with different power dissipations (50 and 100 W/cm^2) by supplying coolant at different flow rates in each tier. This reduces pumping power by 37.5% by preventing overcooling compared with conventional microfluidic coolant delivery.
3) The temperature of the core near the outlet is 16 °C higher than that of the core near the inlet in a multicore processor and the vertical coupling drastically reduces when each tier has its own embedded MFHS.

4) The lateral thermal gradient induced by coolant temperature increases along the flow direction causes the leakage power to increase by 43% for the downstream core. The use of localized coolant delivery enabled through microfluidic chip I/Os is a promising solution to eliminate this thermal gradient and allow the cores to work symmetrically.

REFERENCES

[1] M. T. Bohr, "Interconnect scaling—The real limiter to high performance ULSI," in *Proc. IEDM*, 1995, pp. 241–244.

[2] J. D. Meindl, "Interconnect opportunities for gigascale integration," *IEEE Micro*, vol. 23, no. 3, pp. 28–35, May/Jun. 2003.

[3] S. Borkar, "Thousand core chips: A technology perspective," in *Proc. 44th Annu. Design Autom. Conf.*, 2007, pp. 746–749

[4] 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 Design Test Comput.*, vol. 22, no. 6, pp. 498–510, Dec. 2005.

[5] P. G. Emma and E. Kursun, "Is 3D chip technology the next growth engine for performance improvement?" *IBM J. Res. Develop.*, vol. 52, no. 6, pp. 541–552, 2008.

[6] S. M. Sri-Jayantha, G. McVicker, K. Bernstein, and J. U. Knickerbocker, "Thermomechanical modeling of 3D electronic packages," *IBM J. Res. Develop.*, vol. 52, no. 6, pp. 623–634, 2008.

[7] Semiconductor Industry Association. (2012). *International Technology Roadmap of Semiconductors* [Online]. Available: http://public.itrs.net

[8] 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 Syst. Integr.*, 2009, pp. 1–10.

[9] D. B. Tuckerman and R. F. W. Pease, "High-performance heat sinking for VLSI," *IEEE Electron Device Lett.*, vol. 2, no. 5, pp. 126–129, May 1981.

[10] N. Khan, Y. Hong, P. T. Siow, H. S. Wee, S. Nandar, H. W. Yin, V. Kripesh, J. H. Lau, and C. T. Kok, "3D packaging with through ilicon via (TSV) for electrical and fluidic interconnections," in *Proc. Electron. Compon. Technol. Conf.*, 2009, pp. 1153–1158.

[11] Y. Peles, A. Kosar, C. Mishra, C.-J. Kuo, and B. Schneider, "Forced convective heat transfer across a pin fin micro heat sink," *Int. J. Heat Mass Transf.*, vol. 48, no. 17, pp. 3615–3627, 2005.

[12] 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," *Int. J. Heat Mass Transf.*, vol. 51, pp. 5400–5414, Oct. 2008.

[13] W. Qu and A. Siu-Ho, "Experimental study of saturated flow boiling heat transfer in an array of staggered micro-pin-fins," *Int. J. Heat Mass Transf.*, vol. 52, pp. 1853–1863, Mar. 2009.

[14] V. Sahu, Y. K. Joshi, and A. G. Fedorov, "Hybrid solid state/fluidic cooling for hotspot removal," in *Proc. Thermal Thermomech. Phenomena Electron. Syst.*, 2008, pp. 626–631.

[15] Y. Bao, W. Peng, and A. Bar-Cohen, "Thermoelectric mini-contact cooler for hot-spot removal in high power devices," in *Proc. Electron. Compon. Technol. Conf.*, 2006, pp. 1–6.

[16] L. Sheng-Chih and K. Banerjee, "Cool chips: Opportunities and implications for power and thermal management," *IEEE Trans. Electron Devices*, vol. 55, no. 1, pp. 245–255, Jan. 2008.

[17] L. Zheng and M. S. Bakir, "Electrical and fluidic microbumps and interconnects for 3D-IC and silicon interposer," in *Proc. SOC Conf.*, 2012, pp. 159–164.

[18] L. Zheng and M. S. Bakir, "Design, fabrication and assembly of a novel electrical and microfluidic I/Os technology for 3-D chip stack and silicon interposer," in *Proc. Electron. Compon. Technol. Conf.*, May 2013, pp. 2243–2248.

[19] 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. Electron. Compon. Technol. Conf.*, 2008, pp. 1–7.

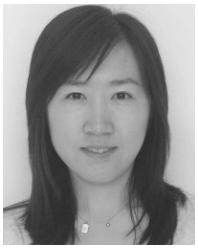
[20] Y. Zhang, A. Dembla, and M. S. Bakir, "A silicon micropin-fin heat sink with integrated TSVs for 3D ICs: Trade-off analysis and experimental testing," *IEEE Trans. Compon. Packag. Manuf. Technol.*, to be published.

[21] M. Shimbo, K. Furukawa, K. Fukuda, and K. Tanzawa, "Silicon-to-silicon direct bonding method," *J. Appl. Phys.*, vol. 60, no. 8, pp. 2987–2989, 1986.

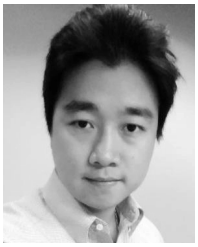
[22] Y. Zhang, A. Dembla, Y. Joshi, and M. S. Bakir, "3D stacked microfluidic cooling for high-performance 3D ICs," in *Proc. Electron. Compon. Technol. Conf.*, 2011, pp. 1644–1650.

[23] K. Shakeri and J. D. Meindl, "Temperature variable supply voltage for power reduction," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI*, Apr. 2002, pp. 64–67.

[24] T. Brunschwiler, B. Michel, H. Rothuizen, U. Kloter, B. Wunderle, H. Oppermann, and H. Reichl, "Interlayer cooling potential in vertically integrated packages," *Microsyst. Technol.*, vol. 15, no. 1, pp. 57–74, 2009.



Yue Zhang (S'13) received the B.S. degree in electrical engineering and automation from the Harbin Institute of Technology, Harbin, China, and the University of Science and Technology of Lille, Lille, France, in 2007, and the M.S. degree in micro and nano technology from the University of Science and Technology of Lille in 2009. She is currently pursuing the Ph.D. degree in electrical engineering with the Georgia Institute of Technology, Atlanta, GA, USA.



Li Zheng received the B.S. degree from Zhejiang University, Hangzhou, China, in 2006, and the dual M.S. degree in electrical and computer engineering from Shanghai Jiao Tong University, Shanghai, China, and the Georgia Institute of Technology, Atlanta, GA, USA, in 2009, where he is currently pursuing the Ph.D. degree in electrical and computer engineering.

Muhannad S. Bakir (SM'12) received the B.E.E. degree from Auburn University, Auburn, AL, USA, in 1999, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology (Georgia Tech), Atlanta, GA, USA, in 2003.

He is currently an Associate Professor and the ON Semiconductor Junior Professor with the School of Electrical and Computer Engineering, Georgia Tech. He is the Editor of a book entitled *Integrated Interconnect Technologies for 3-D Nanoelectronic Systems* (Artech House, 2009).

Dr. Bakir was an Invited Participant in the 2012 National Academy of Engineering Frontiers of Engineering Symposium and has been awarded the 2012 DARPA Young Faculty Award and the 2013 Intel Early Career Faculty Honor Program Award. He is the recipient of the 2011 IEEE Components, Packaging and Manufacturing Technology Society Outstanding Young Engineer Award. He is a member of the International Technology Roadmap for Semiconductors Technical Working Group for Assembly and Packaging.