Thermal Evaluation of 2.5-D Integration Using Bridge-Chip Technology: Challenges and Opportunities

Yang Zhang, *Student Member, IEEE*, Thomas E. Sarvey, and Muhannad S. Bakir, *Senior Member, IEEE*

***Abstract*— In this paper, 2.5-D integrated circuits (ICs) using bridge-chip technology are thermally evaluated to investigate thermal challenges and opportunities for such multi-die packages. To this end, the objectives of this paper are twofold. First, thermal benchmarking of a number of 2.5-D integration approaches is performed and compared to 3-D ICs for completeness. Thermal modeling shows that the evaluated 2.5-D integration approaches exhibit similar thermal characteristics, but show significant improvements compared to 3-D IC solutions with the same power consumption. Second, this paper explores bridge-chip-based 2.5-D integrated systems as a function of bridge-chip thickness, thermal interface material properties, microbump properties, die thickness, die thickness mismatch, and die-to-die spacing along with transient analysis to investigate time-domain thermal coupling. Results suggest that a die thickness mismatch of 100 μm can increase the maximum temperature by 25.8% Therefore, the die thickness mismatch should be kept as small as possible, and the hottest die should be the thickest. Moreover, reducing die-to-die space from 1 to 0 mm increases thermal coupling, and the low-power dice, such as DRAM, can have a temperature increase of 6.1%.**

***Index Terms*— 2.5-D, 3-D integrated circuits (ICs), bridge chip, thermal challenges.**

I. Introduction

EMERGING applications, such as the Internet of Things, cloud computing, and machine learning-based artificial intelligence, have presented performance, communication bandwidth (BWD), and functionality challenges to conventional integrated circuits (ICs) and electronic systems [1]. To address these challenges, novel heterogeneous computing fabrics based on processor (CPU), field-programmable gate array (FPGA)/GPU/application-specified integrated circuit (ASIC) accelerators, and high-density memory [2] have been widely proposed and studied [3]–[5] to increase system throughput, computation capability, and efficiency. However, one of the biggest bottlenecks for such systems is the inter-die BWD, which can cause functional blocks to be idle during data transfer [6], leading to lower system performance.

Manuscript received August 12, 2016; revised March 30, 2017; accepted May 15, 2017. Date of publication June 28, 2017; date of current version July 17, 2017. This work was supported by NSF under Grant CCF-1302297. Recommended for publication by Associate Editor K. Ramakrishna upon evaluation of reviewers' comments. *(Corresponding author: Yang Zhang.)*

The authors are with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: steven.zhang@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.2017.2710042

In order to keep up with rapidly evolving off-chip communication requirements, multiple integration platforms using advanced interconnect technologies have been explored and demonstrated. Silicon interposer-based 2.5-D integration of FPGAs achieves an aggregate BWD in excess of 400 Gb/s [7]. The 3-D processor-on-memory integration using through silicon vias (TSVs) exhibits a maximum memory BWD of 510.4 Gb/s at 277 MHz [8]. Monolithic 3-D integration is another promising option, which achieves even higher BWD than TSV-based 3-D integration resulting from the utilization of shorter and denser nanoscale vertical vias [9]. Moreover, there has been recent interest in multi-die packages using bridge-chip technology, including embedded multi-interconnect bridge technology [1], [10] and heterogeneous interconnection stitching technology [11] to enable 2.5-D microsystems, as shown in Fig. 1(a) and (c), respectively. In its simplest form, bridge-chip technology utilizes a silicon die with high-density interconnects for inter-die communication. The performance metrics of these 2.5-D integration technologies are comparable to interposer-based 2.5-D solutions, but many other benefits are offered, including the elimination of TSVs.

However, as multi-die packaging continues the trend of using more high-peformance (i.e., CPU, GPU, and FPGA) chips in a package, thermal challenges will increase. It is expected that air cooling will become ineffective at cooling such systems without keeping much of the silicon dark. Thermal coupling from high-power chips to low-power chips will also lower the overall system performance [12]. Therefore, in spite of the benefits brought by these emerging integration platforms, there are critical thermal issues that are potential show stoppers.

Thermal analysis and optimization have been extensively conducted for interposer-based 2.5-D integration [13], [14], as well as TSV-based [15], [16] and monolithic 3-D IC integrations [17], [18]. However, there are no thermal modeling efforts focused on 2.5-D bridge-chip-based interconnection platforms. Moreover, previous thermal efforts have generally focused on one of the above-mentioned technologies; there is a need for thermal benchmarking of all these approaches. Therefore, the objectives of this paper are twofold. First, we explore the thermal attributes of bridge-chip-based 2.5-D ICs and benchmark with other 2.5-D and 3-D solutions. Second, we conduct a deep-dive look at bridge-chip-based 2.5-D integration and evaluate thermal performance as a function of

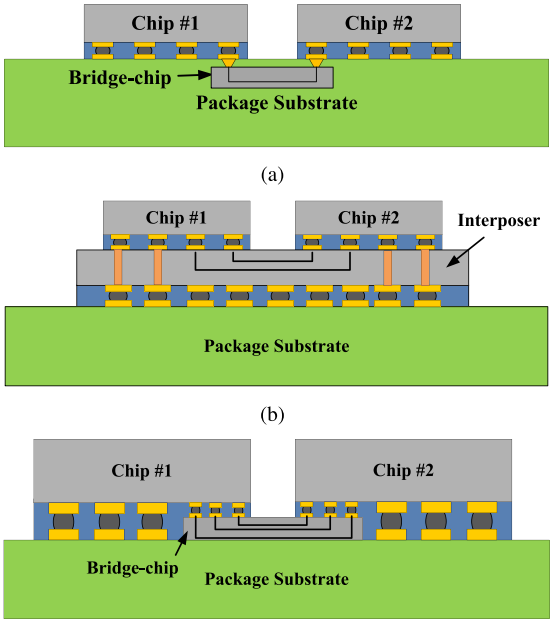


Fig. 1. 2.5-D chip stack using (a) bridge-chip technology and (b) interposer technology. (c) Non-embedded bridge-chip using multiheight microbumps technology.

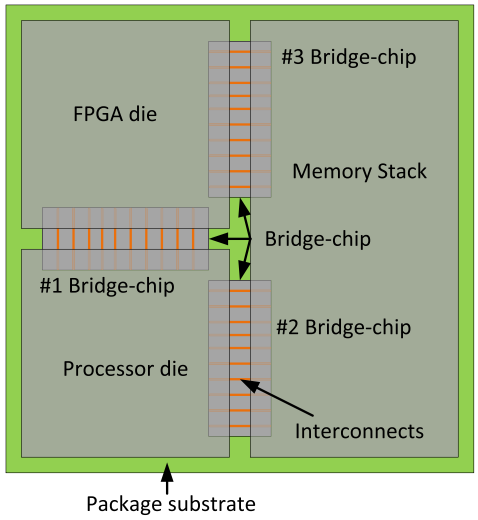


Fig. 2. Illustration of the envision FPGA-CPU-Memory 2.5-D chip stack using bridge-chip technology (top view).

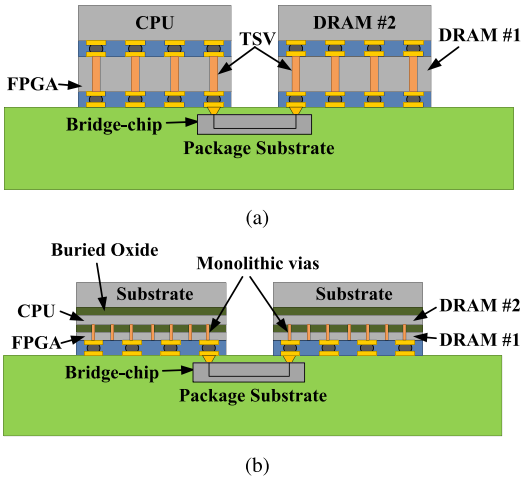


Fig. 3. 3-D chip stack. (a) TSV-based. (b) Monolithic nanoscale via-based.

various technology parameters, such as bridge-chip thickness, thermal interface material (TIM) properties, microbump properties, die thickness, and die spacing. These studies will help the community understand the thermal limits and challenges facing bridge-chip-based integration technologies.

This paper is organized as follows: Section II reports the benchmarking of 2.5-D and 3-D stacks using different integration technologies. In Section III, thermal modeling and simulation specifications are described. Sections IV and V compare 2.5-D bridge-chip-based integration to a number of 2.5-D and 3-D solutions. Section VI explores bridge-chip-based 2.5-D integration as a function of various technology parameters. Finally, Section VII summarizes this paper.

II. 2.5-D and 3-D Benchmark Architectures

A. 2.5-D Integration

Fig. 1 shows three 2.5-D integration technologies with differing chip-to-chip interconnection architectures. The first approach uses bridge-chip technology in which a silicon chip, called the "bridge" chip, is embedded into the package. Chip-to-chip interconnects are routed on the bridge chip, and fine-pitch microbumps are used to connect the bridge chip and the active dice, as shown in Fig. 1(a). The second approach is more traditional and uses interposer technology, as shown in Fig. 1(b). The last approach de-embeds the silicon bridge chip and places it between the active chips and the package, as shown in Fig. 1(c).

With the above-mentioned off-chip technologies, 2.5-D heterogeneous integration of multifunctional chips can be realized. In this paper, we focus on high-performance 2.5-D integration of processor, accelerator (GPU, FPGA, or ASIC), and a memory stack. Specifically, we envision a CPU-FPGA-Memory 2.5-D microsystem as a test vehicle for the bridge-chip technology, and thus, all benchmarks are based on this chip set, as shown in Fig. 2. The FPGA, processor, and memory dice are placed side by side in a package with bridge chips underneath the active dice. We assume that the memory stack consists of five tiers, of which the bottom tier is the controller circuit and the other four tiers are memory cells.

B. 3-D Integration

Fig. 3 shows two 3-D integration architectures, both of which have been extensively explored in the literature. The key enabler for the two stacks is the utilization of vias as chip-to-chip interconnections. The vias in monolithic 3-D are much shorter ($\sim$ a few 100 nm) and smaller ($\sim$100 nm) compared to TSVs ($\sim$5 μm diameter and about 40 $\sim$ 100 μm tall).

In both cases, we still consider a CPU-FPGA-DRAM integrated microsystem. We assume there are two separate 3-D stacks: the first is a CPU-on-FPGA computation stack

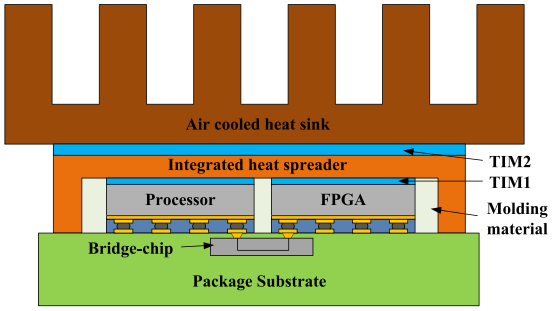


Fig. 4. 2.5-D integration using bridge-chip technology with detailed layer information.

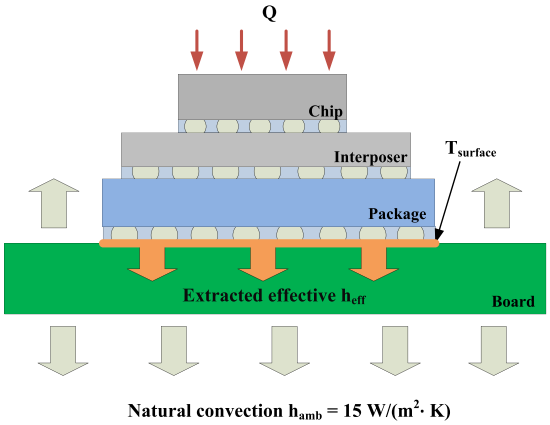


Fig. 5. Setup for characterizing effective convection coefficient.

(CPU is placed on top for thermal consideration), and the second is a DRAM chip stack. To simplify this case study, we assume the FPGA, CPU, and memory chips are of the same size. The memory stack has one controller tier and eight memory cell tiers (the same storage capacity as 2.5-D cases). The two 3-D chip stacks are placed side by side and employ a bridge chip in the package for communication.

III. Thermal Modeling and Specification

A. Thermal Modeling

The thermal framework used in this paper for benchmarking is reported in [19]. To reduce the computation costs, nonconformal meshing methods are implemented in the x-, y-, and z-axes. The mesh sizes are tuned in accordance to geometry differences (chip, heat spreader, interposer, and package). The backward Euler scheme is used to implement the transient analysis. Choleskey factorization is used for both steady- and transient-state analyses. The preconditioned conjugate gradient method is applied when the memory is limited. The above-mentioned steady-state and transient-state models are validated against ANSYS, and the maximum relative error in junction temperature rise is less than 7% [20], [21].

Fig. 4 shows a 2.5-D stack with an air-cooled heat sink. The 3-D IC configurations used for evaluation are quite similar except that the chips are stacked. For thermal modeling, we abstract the heat sink and the printed circuit board (PCB) as primary and secondary cooling boundaries, respectively. Both boundaries are modeled using a uniform convection coefficient applied to the top surface of the heat spreader and to the bottom surface of the package substrate, respectively.

B. Thermal Specifications

1) Layer Thickness and Material Property: The layers' information and material properties are summarized in Table I. On-chip and package metal layers are modeled using in-plane and through-plane thermal conductivity formulated in [22]. Moreover, effective thermal conductivity modeling of the layers with "vertical interconnects" (microbumps, TSVs, etc.) [22] is implemented to further reduce the mesh number. In Sections IV–VI, all the reported simulations are based on these values as shown in Table I.

2) Geometry Parameter and Boundary Conditions: The processor and FPGA dice are assumed to be 1×1 cm^2 large. For the 2.5-D cases, the memory die size is assumed to be 1×2 cm^2, and for the 3-D cases, the memory die size is assumed to be the same size as the processor and FPGA die, i.e., 1×1 cm^2. As mentioned in Section II, the memory stack has five and nine tiers for the 2.5-D and 3-D cases, respectively. The default spacing between dice is 0.3 mm, and the bridge chip is set to be 2.3×7 mm^2. The package size is 2.23×2.23 cm^2. The heat spreader size is assumed to be 4×3.5 cm^2.

The air-cooled heat sink is assumed to have a case-to-ambient thermal resistance of 0.218 K/W [23], and the ambient temperature is assumed to be 38 °C. For the secondary heat path, we used the method described in [24] to characterize the effective cooling capability of the PCB. We assume the stack is assembled on a 10×10-cm^2 PCB, and a natural cooling of 15 W/°C · m^2 is applied to both surfaces of the PCB, as shown in Fig. 5. In order to extract an effective heat transfer coefficient, so that the whole board does not need to be modeled with the package and dice, a power dissipation of 1 W is applied to the top surface. The effective convection coefficient that the PCB provides can be calculated using the equation shown as follows:

$$R = \frac{1}{h_{\text{eff}} \cdot A} = \frac{T_{\text{surface}} - T_{\text{amb}}}{Q}. \quad (1)$$

This value (h_{eff}) is calculated using weighted average temperature of the bottom surface of the package and is found to be 311 W/m^2 · K.

3) Power Maps: The layouts of the emulated processor and DRAM dice are shown in [19]. They are based on Intel Core i7 processor and Samsung 3-D DRAM, respectively. The processor is assumed to dissipate 74.49 W. For the DRAM chip, the bottom controller is assumed to dissipate a uniform power of 5 W, and each DRAM cell tier dissipates 1.46 W. The power profiles of the processor and the DRAM cell tiers are shown in Fig. 6(b) and (c), respectively. The emulated FPGA layout is based on Altera Stratix V and Stratix 10 FPGAs [25]. The FPGA chip power is dependent on application, and in our case, we envision the FPGA chip as the server (processor)

TABLE I
THERMAL SPECIFICATION

Layer	Conductivity (W/mK)		Thickness (μm)	Heat capacity $(J/^\circ C \cdot Kg)$	Mass Density (Kg/m^3)
	In-plane	Through-plane			
TIM2	3		30	1,000	2,900
Heat spreader	400		1,000	385	8,690
TIM1	3		25	1,000	2,900
CPU/FPGA Die	149		125	705	2,329
Memory Die	149		15	705	2,329
Molding	0.28		N.A.	915	1,790
Metal	61.17	1.62	5	433	7,783
Package	30.4	0.38	1,000	600	1850
Interposer	149		200	705	2,329
Bridge	149		200/25[1]	705	2,329
Microbump: CPU/FPGA	60		40	227	12,000
Microbump: memory	60		40	227	12,000
Bonding layer: CPU/FPGA	0.9		40	1,000	2,100
Bonding layer: memory	0.9		7.5	1,000	2,100
Copper	400		N.A.	385	8,690
SiO_2	1.38		N.A.	705	2,648
Tungsten	179		N.A.	135	19,250

[1] 200 μm for bridge-chip case and 25 μm height for non-embedded bridge case.

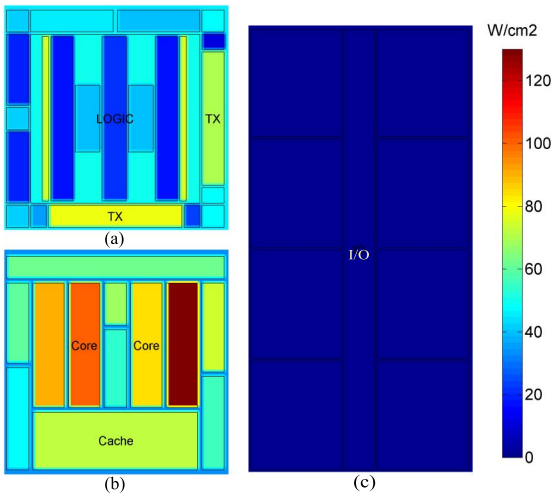


Fig. 6. Power density maps of each die. (a) FPGA die, 44.8 W. (b) Processor die, 74.49 W. (c) DRAM die (cell circuit), 5.65 W for cell circuit.

accelerator. Based on [25], the total power is approximately 44.8 W for the server accelerator, and by using the open-source power calculator [26], we can further estimate the power per functional block and emulate the power profile, as shown in Fig. 6(a).

4) Microbumps and TSVs: The microbumps we study are categorized into two groups: the first is between the chip and the package or interposer, and the second is between the chip and the bridge chip. For the first group, the diameter and the pitch of the microbumps are 40 and 200 μm, respectively. For the second group, dense microbumps are used, and the diameter and the pitch are 20 and 40 μm, respectively. Both groups of microbumps are assumed to be uniformly distributed in the corresponding area.

IV. COMPARISON OF DIFFERENT 2.5-D INTEGRATIONS

All the three 2.5-D IC cases are configured with an air-cooled heat sink, as shown in Fig. 4. The chip placement and power maps are shown in Figs. 2 and 6, respectively. If not stated, all of the thermal analyses are steady-state results using the maximum power maps shown in Fig. 6 to give a worst-case estimate.

Fig. 7 shows the thermal profiles of each die in all cases. The figures are to scale according to die size and spacing listed in Section III. All of the cases exhibit a similar thermal profile because most of the heat is conducted through the attached air cooled heat sink on top (98.18%, 97.17%, and 97.19% for the bridge chip, interposer, and non-embedded bridge-chip cases, respectively). Nevertheless, there is a small difference in the junction temperature resulting from different secondary heat paths in each case. For the interposer-based 2.5-D configurations, heat spreading is enhanced due to the high thermal conductivity of silicon interposer. Therefore, the maximum temperature is the lowest of the three cases ($T_{\max}$ is 102.80 °C). Likewise, the bridge chip is placed closer to the die in the non-embedded bridge-chip case, and as a consequence, the temperature is lower than the first case (0.69 °C cooler). Fig. 8 shows the difference in the spreading capability of the bridge-chip and interposer-based 2.5-D integration cases. For the interposer 2.5-D case, the thermal profile of the layer beneath the chip (interposer layer) is smoother and exhibits a smaller $T_{\max} - T_{\min}$ than that of the bridge-chip case (package layer), which is approximately 4.85 °C lower.

Another observation is the clear lateral thermal coupling between different dice in all cases due to the heat conduction in the heat spreader. The edges of the FPGA and DRAM dice near the processor die are greatly influenced, which creates a relatively larger hotspot area in the FPGA and DRAM dice. To minimize this thermal coupling, it is necessary to apply either tier-specific microfluidic cooling [27] or thermal isolation technology using an insulator [12] to eliminate thermal coupling.

A. Impact of the Thickness of Interposer and Bridge Chip

Based on the above-mentioned analysis, the thickness of the interposer and the bridge chip impact heat spreading. With a

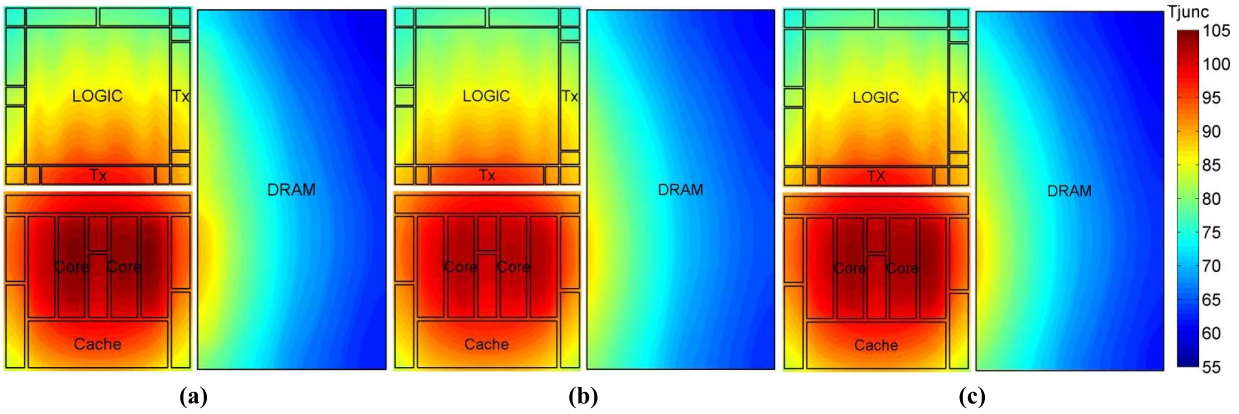


Fig. 7. Top view of thermal profiles of each die in all cases. The bottom die and the hottest die of the DRAM stack are plotted. (a) Embedded bridge chip, $T_{\max}$: 104.92 °C. (b) Interposer, $T_{\max}$: 102.80 °C. (c) Non-embedded bridge-chip, $T_{\max}$: 104.23 °C.

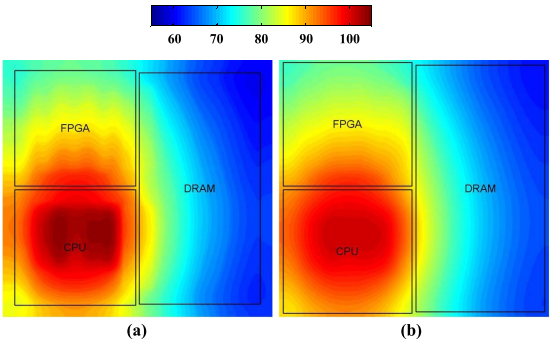


Fig. 8. Illustration of heat spreading effects of (a) package layer in embedded bridge-chip-based 2.5-D, 60.22 °C ∼ 104.60 °C and (b) interposer layer in interposer-based 2.5-D, 61.61 °C ∼ 101.14 °C.

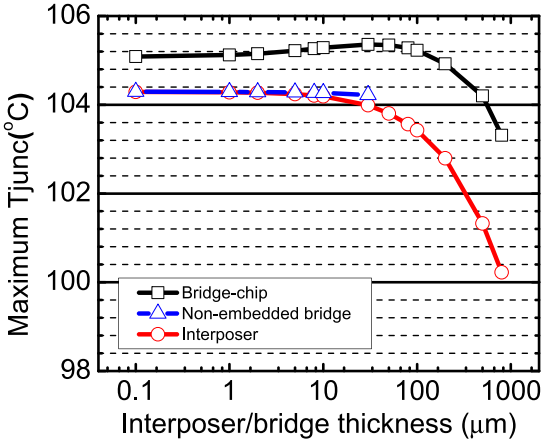


Fig. 9. Impact of interposer and bridge thickness.

thicker silicon layer beneath the dice, the heat spreading is improved and the junction temperature of the hottest die is reduced. Therefore, we sweep the thickness of the interposer and bridge chips from 100 nm to 800 μm (for non-embedded bridge-chip case, the upper bound is 30 μm) and plot the maximum junction temperature of each case, as shown in Fig. 9. For the bridge-chip and interposer cases, $T_{\max}$ decreases as the thickness becomes larger because of the improved spreading of the thicker interposer and bridge chips. Increasing the thickness from 100 nm to 800 μm, $T_{\max}$ is reduced by 1.78 °C and 4.07 °C for the bridge-chip and interposer cases, respectively. The temperature reduction trend is significant when the thickness is larger than 100 μm. For the non-embedded case, the impact is not significant due to the fact that the thickness of the bridge chip is limited by the height of the microbumps.

TABLE II
THERMAL COMPARISON OF BRIDGE-CHIP 2.5-D AND 3-D INTEGRATION

Unit:°C	CPU		FPGA		DRAM [1]	
	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}
Bridge-chip	104.92	83.08	98.28	75.02	89.17	60.01
Monolithic	122.29	93.61	124.22	94.25	96.25	63.57
TSV	121.37	94.64	125.62	98.94	98.18	66.57

[1] For DRAM, we show the maximum temperature of the bottom die in the stack (the hottest die).

V. THERMAL COMPARISON BETWEEN 2.5-D AND 3-D INTEGRATION

When multiple chips are stacked vertically, the power density of the resulting 3-D stack will be larger than that of 2.5-D configurations. As a result, the thermal challenges become more difficult to address. In this section, we thermally explore two types of 3-D integration approaches and compare to bridge-chip-based 2.5-D integration.

For the TSV-based 3-D IC case, we assume a die thickness of 125 μm, a TSV diameter of 5 μm, a liner thickness of 0.5 μm, and a pitch of 100 μm. For the monolithic 3-D IC case, the thickness of both the die and buried oxide is 100 nm; the handle layer is assumed to be 100 μm. The monolithic via diameter is 100 nm and assumed to be tungsten.

Table II lists the maximum junction temperature of bridge-chip-based 2.5-D and the two 3-D IC cases. The results show that 2.5-D integration has better thermal attributes than the two 3-D IC cases. The maximum junction temperature of the CPU in 2.5-D integration is 17.37 °C and 16.45 °C

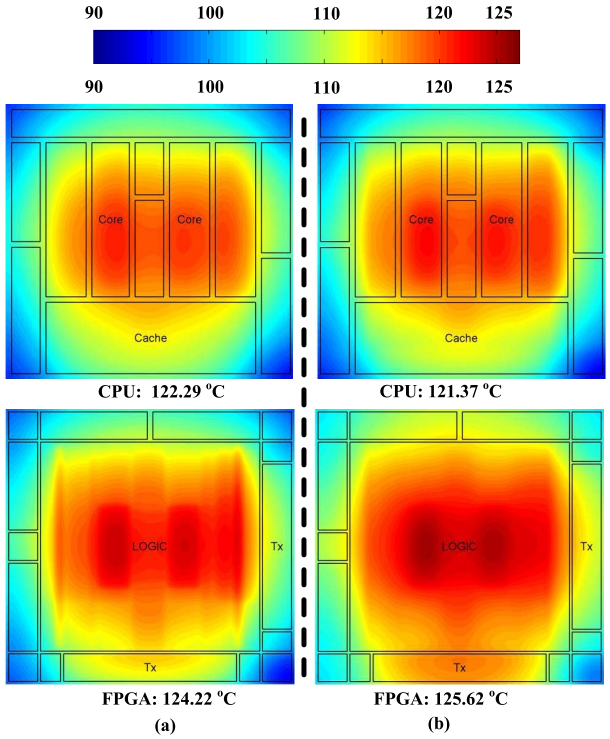


Fig. 10. Thermal profile of each die in the 3-D stack cases. (a) Monolithic 3-D. (b) TSV-based 3-D.

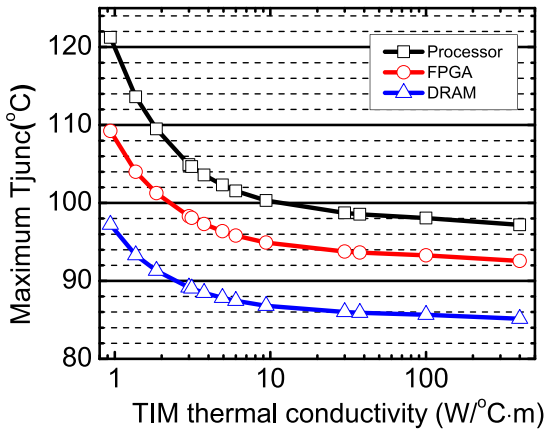


Fig. 11. Impact of thermal conductivity of TIM.

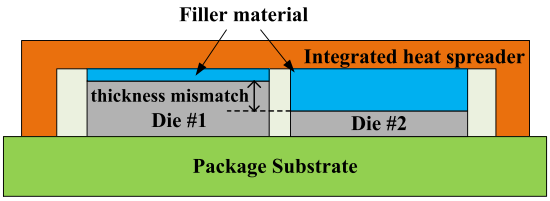


Fig. 12. Illustration of die thickness mismatch.

lower than monolithic and TSV 3-D ICs, respectively. The maximum junction temperature of the FPGA in 2.5-D integration is 25.94 °C and 27.34 °C lower than the monolithic and TSV-based 3-D ICs, respectively. For the DRAM chips, the maximum junction temperature using 2.5-D integration is 7.08 °C and 9.01 °C lower than TSV and monolithic 3-D ICs, respectively. From a thermal perspective, high-power stacks, such as CPU-on-FPGA, may not be practical using 3-D technology; while for a low-power stack, such as DRAM chips, despite the temperature rise, the maximum temperature does not exceed the thermal limit. Additionally, the DRAM chip can be cooler if less tiers are stacked.

Fig. 10 shows the thermal profiles of the CPU and FPGA dice in the two 3-D IC cases. Compared to the bridge-chip 2.5-D case shown in Fig. 7(a), the 3-D cases have stronger thermal coupling, and the thermal profile of one die exhibits a mirror image of the other. The monolithic case has a smaller active layer thickness; thus, the spreading is worse than in the 3-D IC TSV case, as shown in Fig. 3. However, the thermal resistance from the FPGA to the heat sink is slightly smaller than the 3-D TSV case, which results in a lower maximum temperature.

VI. Thermal Study of Bridge-Chip 2.5-D Integration

In this section, we focus on bridge-chip-based 2.5-D integration and thermally evaluate as a function of TIM thermal properties, die thickness mismatch, die thickness, and die spacing. In addition, transient analysis is performed to further understand die-to-die thermal coupling. Through these analyses, the limits and challenges of bridge-chip-based 2.5-D integration are better understood. If not specified, the parameters and power maps are the same as those used in Section III

A. TIM Properties and Die Thickness Mismatch

There are two TIM layers: The first is between the heat spreader and each die (TIM1), and the second is between the heat spreader and the heat sink (TIM2), as shown in Fig. 4. With a good TIM material, the junction temperature will decrease. To evaluate the impact of TIM properties, we sweep the thermal conductivity of TIM1 and TIM2 from 0.9 to 400 W/°C · m (TIM1 and TIM2 are assumed to be the same material). The results are shown in Fig. 11. There is a crossing point in the thermal conductivity at approximately 3 W/°C · m, beyond which better TIM material does not yield significant benefits. Likewise, changing the TIM thickness leads to similar results.

In heterogeneous 2.5-D integration, the chips may be fabricated in different technology nodes and vendors (Intel 22 nm, 14 nm and TSMC 28 nm, 16 nm). Therefore, the die thickness may be different, and it is necessary to use a material to fill the gap, as shown in Fig. 12, using TIM and/or copper (customized heat spreader [28]).

To investigate the impact of die thickness mismatch, we make the following assumptions. First, we assume that the die thickness mismatch is only attributed to the dice; second, we only change the die thickness of one chip and fix the thickness of the other two to the default value; finally, we assume the die with thickness mismatch to be thicker than the other chips.

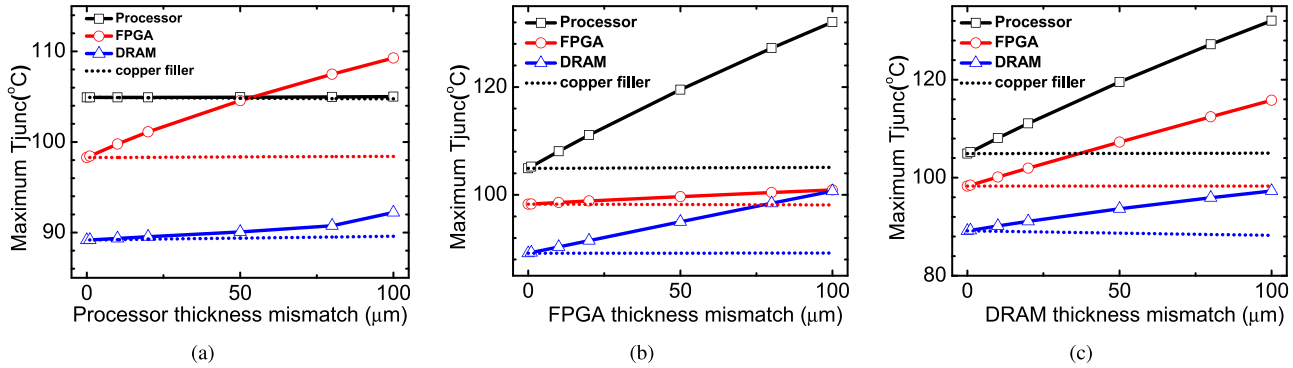


Fig. 13. Impact of die thickness mismatch of (a) processor, (b) FPGA, and (c) DRAM. The solid line in the figures represents the cases using default TIM filler (3 W/°C · m), and the dashed line represents the cases using copper filler (400 W/°C · m).

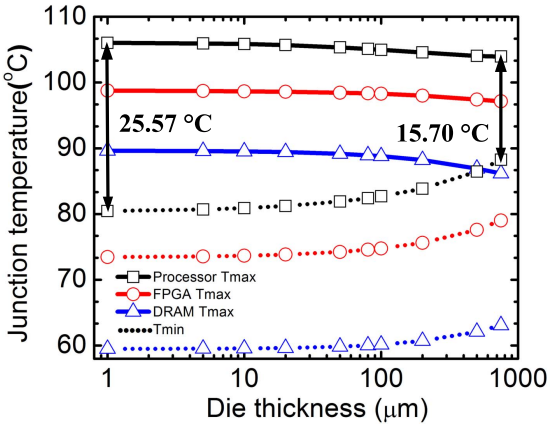


Fig. 14. Impact of die thickness scaling. The dotted line plots $T_{\min}$ of each die.

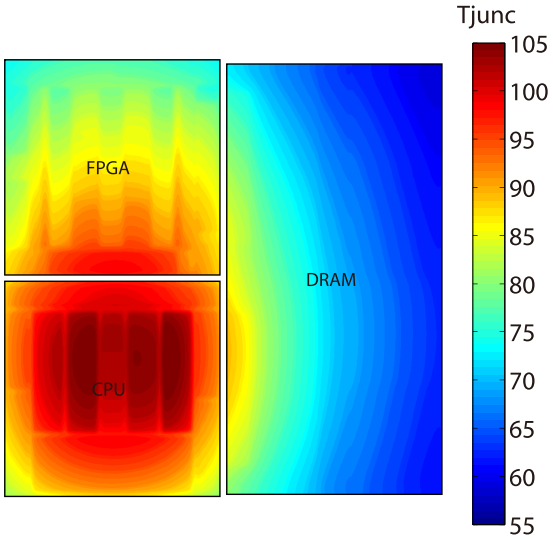


Fig. 15. Thermal profile of each die (die thickness is 1 μm). The heat spreading is confined, and the block outlines are clearly observed from the thermal maps.

The results are shown in Fig. 13. There are three observations. First, good fillers are preferred to avoid elevated temperature. If we use copper instead of a TIM (which is not practical), the temperature of each die experiences a nominal change. Second, using the TIM, the temperature increases as the thickness mismatch increases. Third, when the low-power die is thicker [Fig. 13(b) and (c)], it results in a higher maximum temperature for the whole microsystem. Thus, it is necessary to guarantee that the die with the largest power density has the largest thickness and uses the least filler. In this case, Fig. 13(a) shows when the processor die has a height mismatch of less than 50 μm, the maximum temperature of the whole microsystem does not increase significantly.

B. Die Thickness

The die layer plays an important role in heat spreading, which reduces the localized hotspot temperature. Therefore, as the die thickness scales down, the lateral thermal resistance increases and heat spreading becomes reduced. We sweep the die thickness from 1 to 750 μm, and the results are shown in Fig. 14. Although the maximum temperature of each die does not change significantly (2.08 °C, 1.63 °C, and 3.48 °C for processor, FPGA, and DRAM die, respectively), the intradie variation experiences a relatively larger change. For example, $T_{\max} - T_{\min}$ for the processor changes from 25.57 °C to 15.70 °C when die thickness is changed from 1 to 750 μm.

Fig. 15 shows the thermal profile of each die when the die thickness is 1 μm. The block layout is demarcated in the thermal profile as a result of poor heat spreading.

C. Impact of Microbump and Underfill

The thermal properties of the microbump and underfill impact the secondary heat path. When the effective thermal resistance is reduced, $T_{\max}$ of the whole assembly will minimally decrease. On the other hand, due to the fact that most of the heat is conducted through the main heat path, even if the effective thermal resistance of the microbump layer becomes poor, it is expected that $T_{\max}$ would not change significantly.

However, if the microbump and underfill become thermally resistive, they form an insulation between the active device layer and the interposer or bridge layer. As a consequence, the interconnection wires will have a lower temperature. Based on the modeling of the thermal impact on the interconnection metric of BWD over energy per bit (EPB), the BWD/EPB

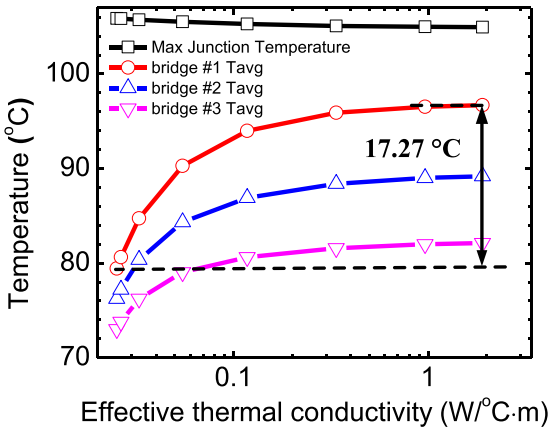


Fig. 16. Impact of effective thermal conductivity of microbump layer for bridge-chip case.

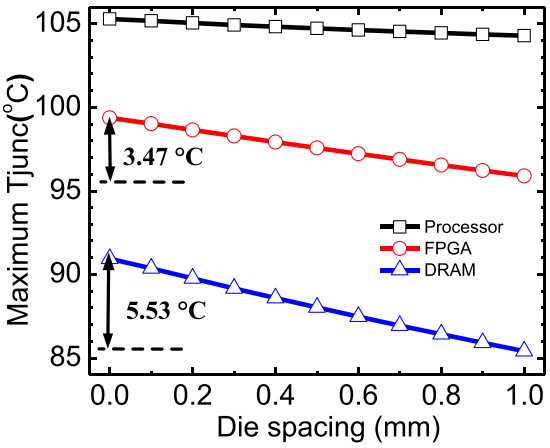


Fig. 17. Impact of die spacing. As the die spacing increases, the junction temperature decreases.

metric can be improved by approximately 7.76% if the temperature is reduced by 30 °C [29]. To investigate the impact of microbump and underfill, we fix the number and diameter of microbumps at the default value and only change the underfill and microbump thermal conductivity to change the effective resistance of the microbump layer. The effective conductivity of the microbump layer is defined as

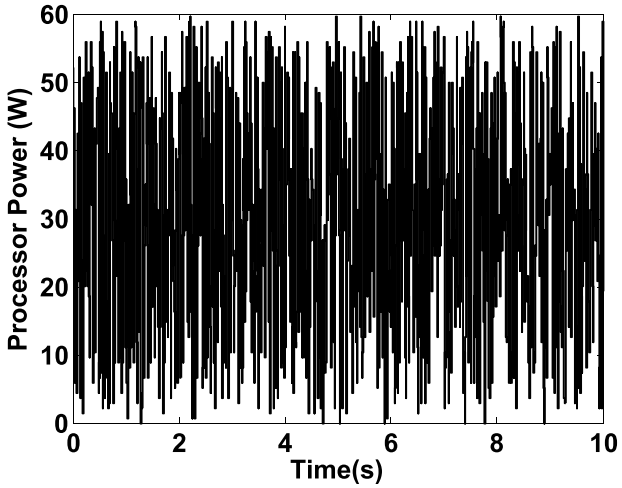
$$k_{\text{eff}} = k_{\text{bump}} \cdot \frac{N \cdot A_{\text{bump}}}{A_{\text{chip}}} + k_{\text{underfill}} \cdot \left(1 - \frac{N \cdot A_{\text{bump}}}{A_{\text{chip}}}\right)$$
$$A_{\text{bump}} = \frac{\pi * D^2}{4} \tag{2}$$

where A_{chip} is the chip area, and N and D are the number and diameter of the microbumps.

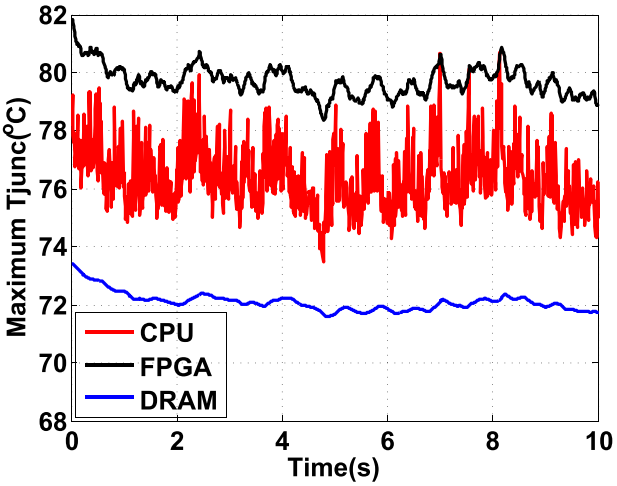
Fig. 16 shows the junction and interconnection temperatures as a function of effective thermal conductivity of the microbump layer for the bridge-chip case (similar for interposer and non-embedded bridge-chip cases). When the effective conductivity of the microbump layer is reduced, there is minimal change in the maximum junction temperature. On the other hand, the average temperature of the bridge chips (where the chip-to-chip interconnections are located) experiences a temperature change as high as 17.27 °C (bridge chip #1). This implies that a low-conductivity material for the microbump layer helps maintain a lower temperature for chip-to-chip interconnections.

D. Die Spacing

Fig. 17 shows that as the die spacing increases (heat spreader is kept the same size), the junction temperature of each die decreases. However, the rate of temperature reduction of each die is not the same. For the die with smaller power, the rate is larger and implies that the low-power die is more vulnerable to thermal coupling. The FPGA and DRAM junction temperatures drop by 3.47 °C and 5.53 °C, respectively. On the other hand, the processor die temperature has a nominal temperature change when the die spacing increases from 0 to 1 mm. Since DRAM performance degrades in the extended temperature range (above 85 °C) [30], it is meaningful to take advantage of this effect and carefully design the spacing between the DRAM die and the other high-power chips (of course, there are tradeoffs between thermal



(a)



(b)

Fig. 18. (a) Emulated processor power and (b) transient analysis results of bridge-chip 2.5-D integration.

considerations and off-chip interconnection metrics, such as BWD/EPB and integration density).

E. Transient Thermal Coupling

The thermal profiles shown in Fig. 7 represent the final steady-state results, but thermal coupling between dice is evolving as the chip activity changes. To investigate the time-domain impact of thermal coupling, we perform a transient analysis to show these time-varying activities. We emulate a processor workload with the activity factor shown in Fig. 18(a). The emulated activity factor has a range of 0.01–0.80. To simplify the case, we assume the FPGA and DRAM dice maintain a constant power.

The maximum junction temperature of each die is shown in Fig. 18(b); time-domain thermal coupling can be observed. When the temperature of the processor changes, the other two dice also experience a temperature change, but with a relatively larger response time and smaller variation. The thermal variation of the processor, FPGA, and DRAM is 7.23 °C, 3.52 °C, and 1.81 °C, respectively. Due to the lateral distance between the FPGA and DRAM dice to the hotspots on the processor, the two dice respond more slowly to power changes in the processor.

VII. Conclusion

This paper presents a comprehensive thermal study for 2.5-D integration focusing on bridge-chip-based technology to identify the thermal limits and challenges in such integration approaches. A CPU-FPGA-DRAM assembly is used as an application example. Bridge-chip 2.5-D integration is compared to interposer and non-embedded bridge-chip 2.5-D integration. Compared to bridge-chip 2.5-D integration, interposer 2.5-D integration offers modest improvements in terms of maximum die junction temperature due to better heat spreading in the interposer layer. Bridge-chip 2.5-D integration is also compared to TSV and monolithic 3-D integration and shows improved thermal response due to smaller power density. We also study the bridge-chip-based 2.5-D integration as a function of bridge-chip thickness, microbump properties, TIM thickness, die thickness mismatch, die spacing, and die thickness. The simulation results show that the die thickness mismatch should be kept as low as possible and that the hottest die should be the thickest. Moreover, from a thermal perspective, low-power dice, such as DRAM, benefit in maximum temperature by 6.1% when the die spacing is increased from 0 to 1 mm. The die thickness plays an important role in heat spreading with thicker die reducing intra-die thermal gradient. Finally, time-domain thermal coupling is investigated. As the temperature of the CPU changes, FPGA and DRAM dice temperatures follow this change.

References

[1] Altera. *Enabling Next-Generation Platforms Using Altera's 3D System-in-Package Technology*, accessed on Jun. 13, 2017. [Online]. Available: https://www.altera.com/content/dam/

[2] J. Jeddeloh and B. Keeth, "Hybrid memory cube new DRAM architecture increases density and performance," in *Proc. Symp. VLSI Technol. (VLSIT)*, Jun. 2012, pp. 87–88.

[3] S. W. Keckler, W. J. Dally, B. Khailany, M. Garland, and D. Glasco, "GPUs and the future of parallel computing," *IEEE Micro*, vol. 31, no. 5, pp. 7–17, Sep./Oct. 2011.

[4] A. Putnam *et al.*, "A reconfigurable fabric for accelerating large-scale datacenter services," in *Proc. ACM/IEEE 41st ISCA*, Jun. 2014, pp. 13–24.

[5] C. Farabet, B. Martini, B. Corda, P. Akselrod, E. Culurciello, and Y. LeCun, "NeuFlow: A runtime reconfigurable dataflow processor for vision," in *Proc. CVPR Workshops*, Jun. 2011, pp. 109–116.

[6] J. Cong, M. A. Ghodrat, M. Gill, B. Grigorian, K. Gururaj, and G. Reinman, "Accelerator-rich architectures: Opportunities and progresses," in *Proc. ACM Design Autom. Conf.*, New York, NY, USA, 2014, pp. 180:1–180:6.

[7] C. Erdmann *et al.*, "A heterogeneous 3D-IC consisting of two 28 nm FPGA die and 32 reconfigurable high-performance data converters," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 258–269, Jan. 2015.

[8] D. H. Kim *et al.*, "Design and analysis of 3D-MAPS (3D massively parallel processor with stacked memory)," *IEEE Trans. Comput.*, vol. 64, no. 1, pp. 112–125, Jan. 2015.

[9] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "High-density integration of functional modules using monolithic 3D-IC technology," in *Proc. IEEE Asia South Pacific Design Autom. Conf.*, Jan. 2013, pp. 681–686.

[10] R. Mahajan *et al.*, "Embedded multi-die interconnect bridge (EMIB)—A high density, high bandwidth packaging interconnect," in *Proc. IEEE 66th Electron. Compon. Technol. Conf. (ECTC)*, May/Jun. 2016, pp. 557–565.

[11] X. Zhang, P. K. Jo, M. Zia, G. S. May, and M. S. Bakir, "Heterogeneous interconnect stitching technology with compressible microinterconnects for dense multi-die integration," *IEEE Electron Device Lett.*, vol. 38, no. 2, pp. 255–257, Feb. 2017.

[12] Y. Zhang, Y. Zhang, T. Sarvey, C. Zhang, M. Zia, and M. Bakir, "Thermal isolation using air gap and mechanically flexible interconnects for heterogeneous 3-D ICs," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 6, no. 1, pp. 31–39, Jan. 2016.

[13] H. Oprins and E. Beyne, "Generic thermal modeling study of the impact of 3D-interposer material and thickness options on the thermal performance and die-to-die thermal coupling," in *Proc. IEEE Intersoc. Conf. Thermal Thermomech. Phenomena Electron. Syst.*, May 2014, pp. 72–78.

[14] X. Zhang *et al.*, "Heterogeneous 2.5D integration on through silicon interposer," *Appl. Phys. Rev.*, vol. 2, no. 2, 2015, Art. no. 021308.

[15] B. Black *et al.*, "Die stacking (3D) microarchitecture," in *Proc. Annu. Int. Symp. Microarchitecture*, Dec. 2006, pp. 469–479.

[16] J. Cong, G. Luo, and Y. Shi, "Thermal-aware cell and through-silicon-via co-placement for 3D ICs," in *Proc. ACM Design Autom. Conf.*, New York, NY, USA, Jun. 2011, pp. 670–675.

[17] S. K. Samal, S. Panth, K. Samadi, M. Saedi, Y. Du, and S. K. Lim, "Fast and accurate thermal modeling and optimization for monolithic 3D ICs," in *Proc. ACM Design Autom. Conf.*, Jun. 2014, pp. 206:1–206:6.

[18] H. Wei, T. F. Wu, D. Sekar, B. Cronquist, R. F. Pease, and S. Mitra, "Cooling three-dimensional integrated circuits using power delivery networks," in *IEDM Tech. Dig.*, Dec. 2012, pp. 14.2.1–14.2.4.

[19] Y. Zhang, Y. Zhang, and M. S. Bakir, "Thermal design and constraints for heterogeneous integrated chip stacks and isolation technology using air gap and thermal bridge," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 4, no. 12, pp. 1914–1924, Dec. 2014.

[20] Y. Zhang, T. E. Sarvey, and M. S. Bakir, "Thermal challenges for heterogeneous 3D ICs and opportunities for air gap thermal isolation," in *Proc. Int. 3D Syst. Integr. Conf. (3DIC)*, Dec. 2014, pp. 1–5.

[21] Y. Zhang and M. S. Bakir, "Integrated thermal and power delivery network co-simulation framework for single-die and multi-die assemblies," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 7, no. 3, pp. 434–443, Mar. 2017.

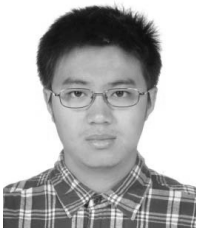
[22] H. Qian, H. Liang, C.-H. Chang, W. Zhang, and H. Yu, "Thermal simulator of 3D-IC with modeling of anisotropic TSV conductance and microchannel entrance effects," in *Proc. 18th Asia South Pacific Design Autom. Conf. (ASP-DAC)*, Jan. 2013, pp. 485–490.

[23] Intel. *Intel Core i7 Processor Families for the LGA2011-0 Socket, Thermal Mechanical Specification and Design Guide*, accessed on Jun. 13, 2017. [Online]. Available: http://www.intel.com/content

[24] Z. Wan, H. Xiao, Y. Joshi, and S. Yalamanchili, "Co-design of multicore architectures and microfluidic cooling for 3D stacked ICs," *Microelectron. J.*, vol. 45, no. 12, pp. 1814–1821, Dec. 2014.

[25] Altera's. *Leveraging HyperFlex Architecture in Stratix 10 Devices to Achieve Maximum Power Reduction*, accessed on Jun. 13, 2017. [Online]. Available: https://www.altera.com/products/fpga/stratix-series/stratix-10/overview%.html

[26] Altera. *PowerPlay Early Power Estimators (EPE) and Power Analyzer (Stratix IV and Stratix V)*, accessed on Jun. 13, 2017. [Online]. Available: https://www.altera.com/support/support-resources/operation-and-testing/%power/pow-powerplay.html
[27] T. E. Sarvey *et al.*, "Embedded cooling technologies for densely integrated electronic systems," in *Proc. IEEE Custom Integr. Circuits Conf.*, Sep. 2015, pp. 1–8.
[28] K. Sikka, J. Wakil, H. Toy, and H. Liu, "An efficient lid design for cooling stacked flip-chip 3D packages," in *Proc. IEEE Intersoc. Conf. Thermal Thermomech. Phenomena Electron. Syst.*, May/Jun. 2012, pp. 606–611.
[29] L. Zheng, Y. Zhang, and M. S. Bakir, "A silicon interposer platform utilizing microfluidic cooling for high-performance computing systems," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 5, no. 10, pp. 1379–1386, Oct. 2015.
[30] J. Liu, B. Jaiyen, R. Veras, and O. Mutlu, "RAIDR: Retention-aware intelligent DRAM refresh," in *Proc. IEEE Int. Symp. Comput. Archit.*, Jun. 2012, pp. 1–12.



Yang Zhang (S'13) received the B.S. degree in microelectronics and math (double major) from Peking University, Beijing, China, in 2012. He is currently pursuing the Ph.D. degree in electrical engineering with the Georgia Institute of Technology, Atlanta, GA, USA.

Thomas E. Sarvey received the B.S. degree in physics and computer engineering from the University of Maryland at College Park, College Park, MD, USA, in 2012. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the Georgia Institute of Technology, Atlanta, GA, USA.

His current research interests include densely integrated, 2.5-D, and 3-D, electronic systems and their enabling thermal technologies.

Muhannad S. Bakir (S'98–M'03–SM'12) is currently a Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. His current research interests include 3-D electronic system integration, advanced cooling and power delivery for 3-D systems, biosensors and their integration with CMOS circuitry, and nanofabrication technology.

Dr. Bakir was a recipient of the 2013 Intel Early Career Faculty Honor Award, the 2012 DARPA Young Faculty Award, and the 2011 IEEE CPMT Society Outstanding Young Engineer Award. He and his research group have received over 20 conference and student paper awards including six from the IEEE Electronic Components and Technology Conference, four from the IEEE International Interconnect Technology Conference, and one from the IEEE Custom Integrated Circuits Conference. His group received the 2014 Best Paper of the IEEE TRANSACTIONS ON COMPONENTS, PACKAGING, AND MANUFACTURING TECHNOLOGY in the area of advanced packaging. In 2015, he was elected by the IEEE CPMT Society to serve as a Distinguished Lecturer for a four-year term. He is an Editor of the IEEE TRANSACTIONS ON ELECTRON DEVICES and an Associate Editor of the IEEE TRANSACTIONS ON COMPONENTS, PACKAGING, AND MANUFACTURING TECHNOLOGY.