Thermal Design and Constraints for Heterogeneous Integrated Chip Stacks and Isolation Technology Using Air Gap and Thermal Bridge

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

***Abstract*—This paper summarizes the thermal challenges in conventional 3-D stacks and proposes a novel stacking structure that eases the thermal problem. The objective of this paper is first to define limits and opportunities for developing different 3-D chip stacks from a thermal perspective, and second to explore our proposed system as a function of microbumps, through silicon vias, die thickness, and other design parameters. In our proposed 3-D stack, the interposer integrated microfluidic heat sink serves as the main heat sink. To thermally decouple stacked dice, we propose air gap isolation between them and a thermal bridge on top of the stack to cool down the isolated die. To evaluate the thermal benefits of the stack, a thermal model is developed based on the finite difference method. Several chip stack scenarios are studied and the simulations are conducted with a processor power of 74.63 W/cm^2 and memory power of 2.82 W/cm^2. The proposed architecture yielded processor and memory temperatures of 64 °C and 40 °C, respectively, compared with 76 °C and 75 °C for the air cooled stack.**

***Index Terms*—3-D integrated circuit (3-D IC), dynamic random-access memory (DRAM), microbumps, microfluidic heat sink (MFHS), multicore processor, through silicon vias (TSVs).**

I. Introduction

3-D INTEGRATION is an emerging technology to address the critical challenges that on- and off-chip interconnects present to nanoelectronics [1]. The use of through silicon vias (TSVs) or monolithic interdie vias enables heterogeneous stacking of multiple dice with very large bandwidth and low energy communication, which improves system performance [2], [3]. However, there are thermal challenges and potential show stoppers due to higher total power density and larger thermal resistance for the dice within the stack [4], [5].

The thermal challenge in a 3-D integrated circuit (IC) is composed of two very distinct components with each requiring separate optimization and technology solutions: first, stacking dice in 3-D increases the total power density while it simultaneously increases the thermal resistance for dice within the stack to the atop-attached heat sink; and second, stacked dice will experience unwanted thermal crosstalk, particularly between high-power die and low-power temperature sensitive components.

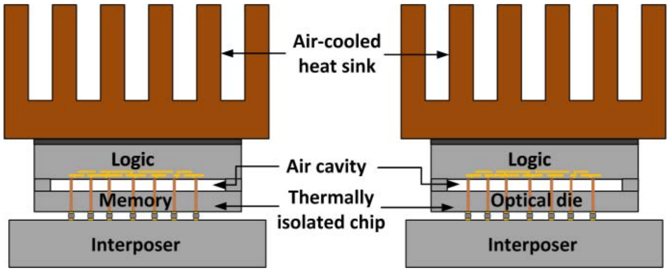


Fig. 1. Illustration of air gap thermal isolation concept.

Microfluidic cooling has been proposed to address the heat removal challenges. The chip- or interposer-embedded staggered micropin fin heat sink has been shown to produce thermal resistance in the range between 0.27 and 0.33 °C · cm^2/W using dimensions suitable for within silicon integration [6]. While significant work has been done to directly address the thermal challenges of 3-D integration, relatively little attention has been paid to the negative effects of thermal coupling between different dice in 3-D IC and, in particular, to minimize interdie thermal coupling. For instance, in a dynamic random-access memory (DRAM)-on-processor 3-D stack, the DRAM usually exhibits a thermal profile similar to the processor die, i.e., a mirror image, and has a relatively high temperature due to strong thermal coupling even though DRAM itself dissipates much lower power than the processor [7]. However, higher DRAM temperature degrades memory performance, resulting in lower overall system performance [8]. Therefore, the memory die must be thermally decoupled from the processor die. To resolve this thermal coupling problem, an air gap isolation concept [9] is presented to prevent thermal coupling from one die to another, as shown in Fig. 1. However, the challenge of removing heat from the isolated die was not previously considered, and the impact of TSVs, microbumps, die thickness, and other die parameters was not fully explored.

The objective of this paper is twofold. First, we explore the limits and opportunities of current 3-D stacked IC and propose novel 3-D IC stack architectures that are thermally centric; specifically, our proposed 3-D stack architectures

Manuscript received March 19, 2014; revised August 1, 2014; accepted October 8, 2014. Date of publication November 19, 2014; date of current version December 5, 2014. This work was supported by the Defense Advanced Research Projects Agency, Arlington, VA, USA, under Grant N66001-12-1-4240. Recommended for publication by Associate Editor R. Prasher 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: steven.zhang@gatech.edu; yzhang324@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.2014.2364742

enable thermal decoupling of dice within the stack as well as microfluidic cooling in the interposer. Second, we thermally analyze our proposed 3-D stack architecture as a function of TSV/microbump diameter, TSV/microbump number, TSV layout, and die thickness. Moreover, we present the design implications of these technology parameters.

This paper is organized as follows. Section II further discusses the motivation of thermal isolation. Section III describes our proposed and conventional architectures of 3-D stack as well as related technologies. In Section IV, thermal modeling based on the finite difference method is described. Section V compares different stack architectures based on simulations using our thermal models. Section VI explores our proposed 3-D stack as a function of various technology parameters. Finally, Section VII concludes this paper.

II. Motivation of Thermal Isolation

A. DRAM Chip

DRAM chips are periodically refreshed to maintain capacitor charge. During the refresh operation, external access of the DRAM is partially disabled, and additional power is drawn for the refresh operation. Therefore, higher DRAM refresh rates decreases throughput and performance while simultaneously increasing power. According to the characterization of 248 DDR3 DRAM chips from five vendors [10], the refresh retention time exponentially reduces as the temperature increases. Using an advanced control engine that refreshes according to different temperature bins [8], the DRAM can maintain a relatively longer average refresh period that results in higher performance and lower power. For example, in the case of a DRAM chip with a hybrid refresh rate of 256, 128, and 64 ms instead of just a single refresh rate of 64 ms, the power of the DRAM chip decreases by 16.1%, while the system performance increases by 8.6% [8]. For future 64 Gbit DRAM chips, the worst throughput loss induced by refresh activities can be 40% and the percentage of refresh power can be as large as 50% [8]. This problem gets aggravated when memory is stacked with high-power processor because the DRAM chip in such stacks exhibits a relatively higher temperature, close to the processor's peak temperature, due to the strong thermal coupling phenomenon in 3-D IC [7]. Higher temperature leads to performance degradation and power overhead. Thus, it is important to decouple the thermal path from processor to DRAM chips to reduce DRAM temperature.

B. Silicon Nanophotonics

In the domain of silicon nanophotonics, a number of components are sensitive to temperature variation. For example, the temperature sensitivity of a silicon-based 10-μm-diameter electrooptic modulator is 0.11 nm/°C [11]. Assuming 64 channels of wavelength division multiplexing sharing the working band, i.e., bands (1530–1625 nm), each wavelength channel is 1.48-nm wide [12]. If we use the above modulators in the two ends (the send and receive), there will be a complete wavelength mismatch when the temperature drifts by only 14 °C (under such a temperature variation, wavelength drift is 1.54 nm $>$ 1.48 nm). 3-D stacked nanophotonics, such as circuit-switched optical interconnection networks on logic, can not only decrease the memory access latency but also can improve the power efficiency [13]; however, the thermal crosstalk between the photonic components and high-power logic circuits presents a potential challenge for 3-D photonics.

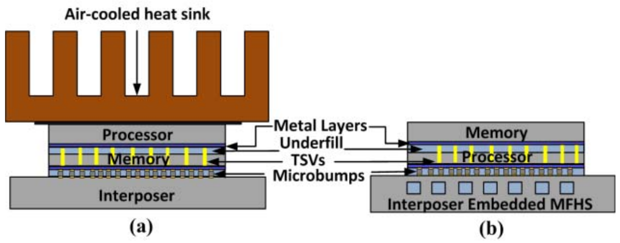


Fig. 2. 3-D stack. (a) With conventional air cooled heat sink. (b) With interposer embedded MFHS.

C. Microelectromechanical System

Similarly, there are temperature coupling challenges for the 3-D stacking of microelectromechanical systems (MEMS) and their readout circuits; for example, consider the silicon on insulator MEMS accelerometer and readout circuit stack in [14]. The read accuracy of the inertial elements, such as accelerometers or gyroscopes, will dramatically degrade with varying temperature [15]. The accuracy of a gyroscope obtained under varying temperature will degrade by almost one order of magnitude compared to the one with stable temperature [16]. Since MEMS modules generally have relatively low power consumption, their temperatures will be strongly influenced by other high-power modules in the 3-D stack. By reducing unwanted thermal coupling, the accuracy of the MEMS component will be enhanced.

From the above discussions, it is clear that thermal isolation in a 3-D stack is important for many applications, in particular, when decoupling the heating of low-power and temperature-sensitive devices from high-power IC. Under such conditions, we propose air gap isolation between the stacked dice to minimize thermal coupling between dice.

III. Benchmark Architecture

In this paper, we focus on heterogeneous 3-D stacking of high- and low-power devices such as memory on logic, photonics on logic, and MEMS on logic. Specifically, we use a memory on processor stack as our application example.

Fig. 2 shows two 3-D stacks with different cooling solutions. The first 3-D stack is based on the current approaches in the literature in which an air cooled heat sink (with heat spreader) is attached to the top of the stack. Since the heat sink is on the top, the thermally optimal architecture for this stack is to place the processor above the memory.

The second 3-D stack is cooled using a microfluidic-cooled interposer [17], [18], as shown in Fig. 2(b). In this case, the processor is on the bottom to minimize the thermal resistance between the processor and the heat sink. Moreover, placing the processor die on the bottom avoids the need to route power, ground, and signal interconnections through the memory die [6]. Even if the microfluidic-cooled interposer

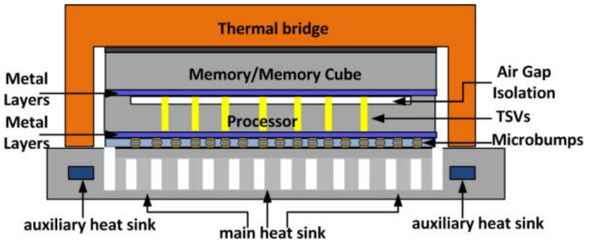


Fig. 3. Proposed architecture with interposer-embedded heat sink, thermal bridge, and air gap isolation.

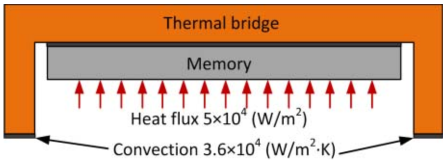


Fig. 4. Thermal bridge on the top of the memory simulated in ANSYS.

[Fig. 2(b)] can lower the system temperature compared to the air cooled heat sink, the thermal coupling issue is still unsolved because there are no mechanisms to prevent heat transfer between the two dice.

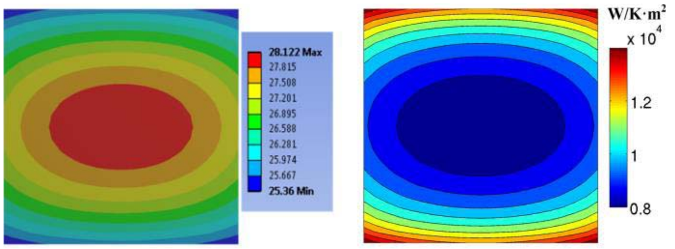
To take advantage of microfluidic cooling as well as to solve the thermal coupling challenge, we propose a new 3-D stack shown in Fig. 3. The proposed architecture has three novel features.

1) A microfluidic heat sink (MFHS) is integrated in the interposer and consists of two separate parts. The main MFHS is directly under the high-power chip, i.e., the processor. It serves as the main thermal path for the stack. The auxiliary MFHS is located at the peripheral of the interposer and is used to cool the thermal bridge (to be discussed later).
2) An air gap thermal isolation is integrated between the high-power and low-power dice to decouple the thermal crosstalk.
3) A thermal bridge is attached on top of the isolated low-power die to provide a cooling path. Next, we will discuss the details of the thermal bridge.

A. Thermal Bridge

The air gap is integrated to prevent the heat generated by the high-power die from traveling up to the low-power die. Likewise, it also prevents the heat generated by the low-power die from traveling downward along the main thermal path. Without an effective thermal path for the isolated die, it will be at a higher temperature. Thus, the thermal bridge is proposed as a solution. This proposed solution is similar to a lid in a conventional 3-D IC [19] with the key difference being that the bottom of the thermal bridge is connected to an auxiliary MFHS in the interposer. The thermal bridge can be made of copper and thus have a small conductive thermal resistance. Here, we calculate the thermal resistance from the bridge to the ambient. When we model the whole system, the thermal bridge can be emulated as a convective boundary condition using simulated thermal resistance.

The structure shown in Fig. 4 is simulated in ANSYS. The dimensions of the memory die are 1 cm × 1 cm × 50 μm.



(a) (b)

Fig. 5. (a) Temperature profile on the top of the memory die. (b) Characterized heat transfer coefficient of thermal bridge.

The top surface of the copper thermal bridge is 1.5 cm × 1.5 cm with a thickness of 500 μm. The bridge is attached to the top of the memory die through a 5-μm-thick thermal interface material (TIM) with thermal conductivity of 3 W/K · m. The two support structures (fins) have a width of 2 mm and a height of 200 μm. A 10 μm TIM is assumed at the bottom to connect the thermal bridge and the interposer. The convective thermal resistance of 0.28 K · cm^2/W was converted to a boundary condition with a heat transfer coefficient of 3.6×10^4 W/m^2 · K. The ambient temperature is 22 °C. In the simulation, the memory die dissipates 5 W.

The objective in this section is to characterize the thermal resistance of the thermal bridge (including the TIM and the convective boundary). Fig. 5 shows the temperature profile of the stack and the corresponding heat transfer coefficient. As can be observed from Fig. 5(a), the highest temperature of the memory die is 28.12 °C and appears in the center. Thus, the temperature gradient from the memory to ambient is 6.12 °C, yielding a total thermal resistance of 1.22 °C/W, which is much better than the natural convective cooling of a package. From Fig. 5(b), the heat transfer coefficient is quite uniform throughout the chip and only the corners exhibit larger values. Considering such phenomenon, we assume a uniform maximum thermal resistance for the bridge, which gives a worst case estimation and simplifies the modeling. In Section VI-A, we show that the difference between the uniform thermal resistance assumption to the position dependent modeling is relatively negligible.

B. TSVs Placement

In 3-D ICs, the TSVs are good heat conductors due to the high thermal conductivity of copper. The location of the TSVs will promote stronger thermal coupling [20]. Thus, the placement of the TSVs is an important thermal consideration. There are three key types of TSV placement methodologies. The first type is to distribute the TSVs throughout the chip uniformly. The second type is to cluster the TSVs in the center or periphery of the chip, as in 3-D DRAM or wide input/output (I/O) technology [1], [21]. The last type uses the objective targeted strategy, which is determined by area, stress, crosstalk, power delivery network noise, and so on [22].

In this paper, we focus on the first two methods, as shown in Fig. 6(a) and (b), respectively. When the TSVs are distributed uniformly in the chip, there will be a uniform thermal coupling between the two dice. However, if we cluster the

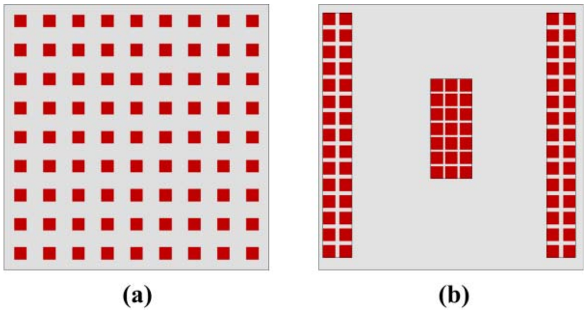


Fig. 6. TSV distribution scenario, where the red square denotes the TSV. (a) Uniformly distributed and (b) clustered TSVs.

TSVs in a farm-like fashion, similar to wide I/O technology, the coupling is expected to be localized. In Section VI, we will discuss the impact of TSV layout in detail.

IV. Thermal Modeling

In a 3-D stack, there are multiple layers with heterogeneous materials. Thus, it is time consuming to simulate the whole stack using finite element method software such as *ANSYS* in such simulations. Modified *hotspot* [23], [24] with weighted average thermal resistance model decreases the complexity of the geometry but reduces the solution accuracy. Other modeling methods based on frequency domain computation, Green's function, and cosine or sine transforms [25] are faster, but their extension to heterogeneous stacks with nonuniform materials is difficult.

Based on the thermal simulator using different sizes of grid in chip and package [26], we developed a nonconformal meshing flow and implemented the finite-difference method in our models. We extended the idea of nonconformal gridding inside the chip domain, and we modified the grid size corresponding to the geometry difference. It was ensured that each grid element was composed of homogeneous material. Thus, the model was able to quantify the impact of each structure even as small as the TSV liner. It was also validated by *ANSYS* simulation.

The finite difference scheme which we use, is described in [27]. In the scheme, there is no requirement that the mesh along different axes should be equal. This allows us to apply these numerical schemes to the nonconformal meshing nodes.

A. Nonconformal Meshing Strategy

The mismatch of the grid size in different areas results in poor convergence. To mitigate such nonconformal gridding, the mesh size should gradually increase during transition from a small to a larger geometry [26]. It is important to choose the proper size ratio of the adjacent meshes to maintain convergence and efficiency. A ratio above 13.3 can lead to nonconvergence [26]. In our case, we set the maximum ratio as 10.

When meshing the chip, we first mesh the TSVs and microbumps. We then add more mesh lines with two requirements: 1) the gradual transition requirement and 2) the maximum mesh size constraints at the chip and interposer. Fig. 7 sketches the gridding results of the interposer and chip.

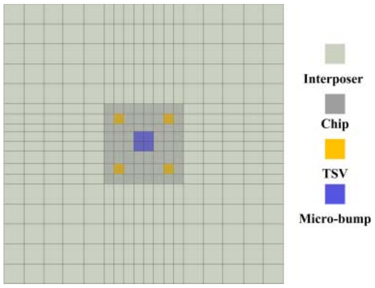


Fig. 7. Illustration of nonconformal meshing. The mesh for TSV liner is not shown.

For simplification and illustration, we assume there are only four TSVs and one microbump. The mesh for the liner is omitted and the TSV size is magnified for visualization.

B. Metal Layer Modeling

The on-chip metal layer consists of metal traces, vias, and dielectrics. The metal layers often contain a large number of dummy metal traces to obtain a relatively homogeneous surface during chemical-mechanical planarization to yield a smoother surface and better wire resistance [28]. The metal percentage is about 20%–80%, so the metal layers can be treated as a lumped single layer. This abstraction has been validated with *COMSOL* models and the maximum error is approximately 6% [29].

Due to the above observations, we use equivalent thermal conductivity to model the metal layers and assume the percentage of copper is 50%. The lateral and vertical thermal conductivity is expressed as

$$1/k_{\text{lateral}} = \sum_{1}^{2} p_i/k_i \tag{1}$$

$$k_{\text{vertical}} = \sum_{1}^{2} p_i * k_i \tag{2}$$

where p_1 and p_2 are the percentage of the metal and the dielectrics, respectively; and k_1 and k_2 are the thermal conductivity of the metal and dielectrics, respectively [20].

C. Thermal Simulation Flow

The thermal models have three macroinputs: 1) the power consumption of each functional block in the chip; 2) the geometry information of the 3-D stack, such as the dimension of each component; and 3) the material properties.

After meshing the whole stack, we can obtain each nonzero element in the thermal conductance matrix Y, which is highly sparse and symmetric. The relationship between the matrix Y, power consumption vector b, and the unknown temperature x is expressed as follows:

$$Yx = b. \tag{3}$$

This matrix equation can be solved either by a direct solver or by the iterative methods described in [30] if memory resources are limited. All algorithms and related methodologies are implemented using MATLAB.

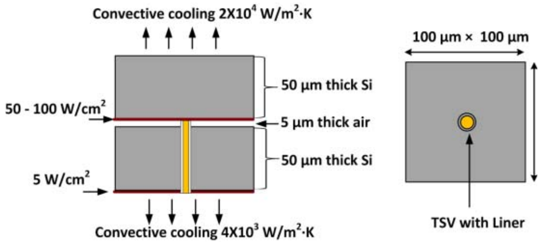


Fig. 8. Configuration of the validation experiments.

TABLE I
VALIDATION OF THE THERMAL TOOLS I

Configuration				ANSYS		Model	
Liner	R_{tsv}	$Thick_{liner}$	P_{top}	T_{top}	T_{bot}	T_{top}	T_{bot}
SiO2	1	1	50	45.5	42.9	45.37	43.44
SiO2	1	1	100	67.3	59.6	66.97	61.20
Air	1	1	50	45.8	41.4	45.80	41.31
Air	1	1	100	68.2	54.9	68.24	54.83
SU8	1	1	50	45.3	42.3	45.52	42.70
SU8	1	1	100	67.6	57.9	67.41	58.99
SU8	1	5	50	45.6	42	45.62	42.17
SU8	1	5	100	67.8	56.9	67.73	57.40

R_{tsv} denotes the radius of TSV copper and $Thick_{liner}$ denotes the thickness of liner, the unit is µm. P_{top} denotes the power density of the top die, the unit is W/cm². T_{top} is the temperature of the top die and T_{bot} is the temperature of the bottom die

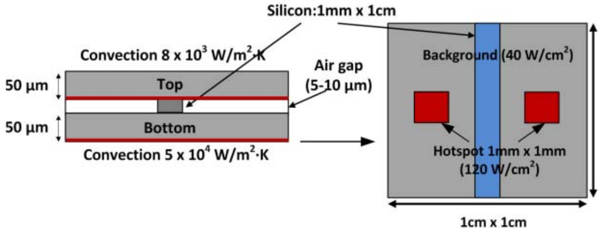


Fig. 9. Schematic of the experiment with two hotspots in the bottom die.

D. ANSYS Validation

To validate the accuracy of the model, we design two simple benchmarks. The first one is similar to that described in [9] and is shown in Fig. 8. The power distribution of both dice is uniform, and the top and bottom dice are assigned power densities of 100 and 5 W/cm^2, respectively. Table I shows the results of all the scenarios. The finite-difference model matches the ANSYS results with a maximum error of 2.7%.

The second benchmark is with two hotspots in the bottom die while keeping a uniform power of 5 W/cm^2 in the top die. The schematic is shown in Fig. 9. In this benchmark, we have two scenarios: 5- and 10-μm air gap. Table II lists the temperature range in both scenarios. Fig. 10 shows the thermal profiles obtained from both ANSYS and the model. In this experiment, the finite-difference model also matches the ANSYS results. The thermal maps are identical and the maximum error of the temperature is 0.42%.

TABLE II
TEMPERATURE RANGE OF EXPERIMENT II

Air gap thickness	ANSYS		Model	
	T_{bot}	T_{top}	T_{bot}	T_{top}
5 µm	28.66-38.21	25.77-28.89	28.72-38.25	25.82-28.87
10 µm	28.58-38.56	24.88-28.84	28.66-38.58	24.91-28.87

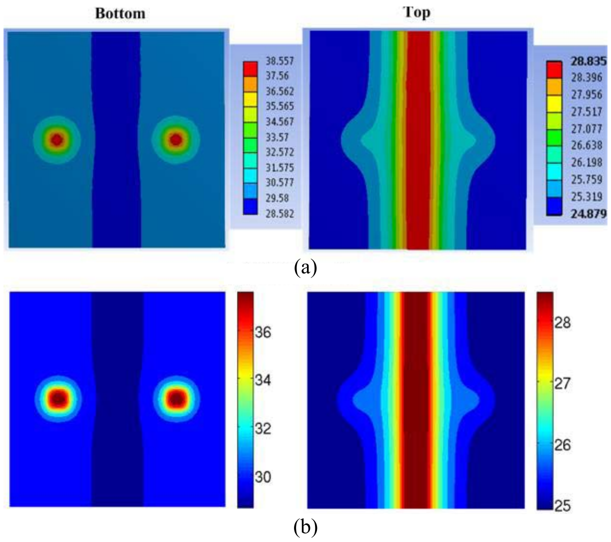


Fig. 10. Thermal profiles of the bottom and top dice in the experiment shown in Fig. 9. (a) ANSYS simulation results. (b) Model simulation results (10 μm).

TABLE III
PARAMETERS

	Conductivity (W/K·m)	Thickness (µm)
TIM	3	5
Memory die	149	100
Underfill layer	0.9	5
Air gap	0.024	5
Processor die	149	50
Micro-bump	60	40
Interposer	149	200
Copper	400	N/A
SiO_2	1.38	N/A

V. COMPARISON OF DIFFERENT 3-D STACKS

In this section, we evaluate the three memory-on-processor stacks described in Section III. This benchmark analysis will highlight the limits and opportunities of conventional 3-D stacks shown in Fig. 2 as well as the thermal benefits of our proposed architecture shown in Fig. 3.

A. Specifications

1) Geometry Parameter and Thermal Boundary Condition: Table III lists the thickness and material of each layer. Those parameters apply for all the stacking architectures. The thermal conductivity of the TIM layer is based on [31]. Due to the fact

TABLE IV
BOUNDARY CONDITION

Stack	Boundary face	Heat Transfer Coefficient (W/K·m^2)
air cooled heat sink	Air heat sink (top)	20000
	Others (near adiabatic)	5
Only with MFHS	MFHS (bottom)	50000
	Others (near adiabatic)	5
Our proposed stack	thermal bridge	8000
	MFHS (bottom)	50000
	Side (near adiabatic)	5

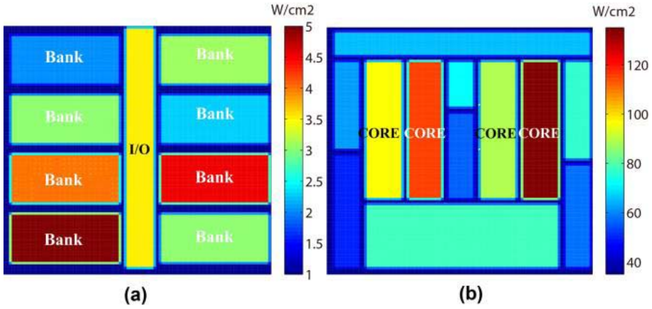


Fig. 11. Power density distribution. (a) Memory die. (b) Processor die.

that the upward heat path is dominated by the resistance of the thermal bridge, the TIM thickness has a minimal impact. Thus, we set the TIM thickness to 5 μm, which is of the same magnitude of published data [31], [32]. The chip size is assumed to be 1 cm × 1 cm. The interposer is set to be 3.5 cm × 3.5 cm. The interposer embedded MFHS is assumed to be the same size as the chip.

In our thermal modeling, we treat all heat sinks as convective boundary conditions. Table IV lists the heat transfer coefficient of each stack. From Section III-A, the thermal resistance from the thermal bridge to ambient is 1.22 °C/W. To give a worst case estimate, we use 1.25 °C/W as the convective resistance for the thermal bridge, which is equivalent to a convective heat transfer coefficient of 8000 W/K · m^2. The ambient temperature is set to be 25 °C.

2) Power Density Maps: Fig. 11 shows the power maps of the memory and processor dice. The memory die layout is based on an 8 Gbit 3-D DDR3 DRAM design from *Samsung* [1]. The size of that DRAM chip is 1.09 cm × 0.9 cm, which is close to the assumed value. The total power is estimated from the Micron DDR3 DRAM datasheet [33], which gives a value of 2.82 W.

The layout of the processor die is based on the *Intel* i7 microprocessor [34]. The thermal design power of i7 K series processor is 84 W. Since our assumed chip size is smaller than the real case, we scale the total power to 74.63 W. According to the estimation from *McPAT,* for a 22-nm 4-core processor, the cores consume approximately 60% of the total power [35]. Hence, we assign the power distribution as shown in Fig. 11(b).

3) Microbumps and TSVs: The microbumps we study are between the interposer and the processor die. They are assumed to be uniformly distributed throughout the chip. The default diameter is 40 μm and the total number is 1600.

With regard to the TSVs, we assume the TSV diameter to be 5 μm with a liner thickness of 0.5 μm. The default total number is assumed to be 10 000, and they are uniformly distributed throughout the chip.

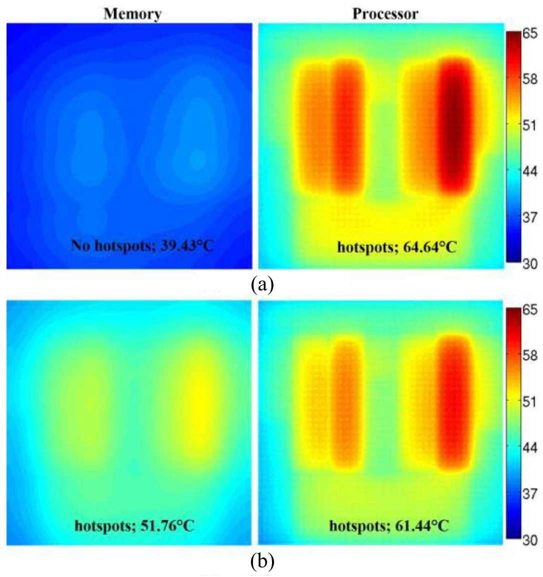
B. Thermal Comparison of Stacks With Different Cooling Methods

The two baseline stacks are shown in Fig. 2. They are configured with an air cooled heat sink and an interposer embedded heat sink, respectively.

Because the TSVs may influence the decoupling results of the air gap, we study the proposed stack with and without TSVs to give a worst and best estimation.

For all stacking scenarios, we also study two conditions: standby processor (24.63-W total power) and active processor (74.63-W total power). When the processor jumps from standby to active state, the temperature will increase, and we seek to see how this variation influences the temperature of the memory die. Table IV shows the maximum temperature of each die under the two different processor states. From the results, there are three key conclusions. First, our proposed architecture has the lowest temperature in both standby and active states. In the active state, the maximum temperatures of the three stacks (first three rows) are 76.44 °C, 66.05 °C, and 64.64 °C, respectively. The maximum temperature of our proposed stack is approximately 12 °C lower than that of the air cooled heat sink. Second, our proposed architecture decouples the heat from the processor to the memory. For the first two scenarios, in both active and standby modes, the memory exhibits a temperature similar to that of the processor. While for the proposed stack with thermal isolation, in standby mode, the memory temperature is 31.96 °C, while the processor temperature is 47.58 °C; in active state, the memory temperature is only 39.63 °C even though the processor temperature is as high as 64.64 °C. Third, our proposed stack maintains the memory die temperature fairly independent of the processor die activity status. In the first two cases, when the processor transits from the standby mode to the active mode, the temperature of the memory die increases by 25 °C and 20 °C, respectively. For our proposed stack, it increases only by 8 °C when the processor changes to active state.

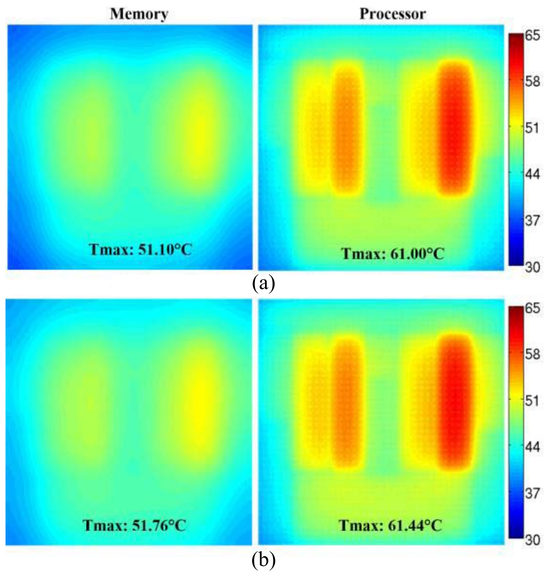
However, when TSVs are inserted in our proposed stack, the thermal coupling increases, as expected, compared with the TSV-free case. In both activity states, the temperature of the memory die is always higher than the case without TSVs, which indicates coupling. The thermal map of each die is shown in Fig. 12 when the processor die is in the active state. In Fig. 12(b), the temperature distribution of both dice is similar to each other, and the temperature difference of the two dice is only 10 °C compared with 25 °C in Fig. 12(a). Thus, the TSVs clearly impact the thermal isolation, and we will discuss this further later in this paper.



(a)

(b)

Fig. 12. Thermal maps of proposed stack when processor is in active mode. (a) Without TSVs. (b) With TSVs.



(a)

(b)

Fig. 13. Thermal maps of proposed stack under two modeling cases of thermal bridge. The simulation is the fourth case in Table V. (a) Nonuniform modeling. (b) Uniform modeling.

VI. Study of the Proposed Architecture

In this section, we thermally study our proposed 3-D stack architecture as a function of the cooling capability of the thermal bridge, TSV/microbump diameter, TSV/microbump density, TSV layout, and die thickness. Through this analysis, the benefits, limits, and challenges of our proposed architecture can be better understood. If not specified, the parameters and power maps are the same as those used in Section VII.

A. Thermal Bridge

1) Thermal Bridge Modeling: As we mention in Section III-A, we model the thermal bridge as a lumped thermal resistance with a uniform value throughout the chip. In this section, we will show the difference between the uniform thermal resistance assumption to the position dependent modeling is relatively negligible.

Fig. 13 shows the thermal maps of the proposed stack with TSVs (the fourth case in Table V) under the two modeling methods. The maximum error is 9.7% and occurs at the corners of the memory die because the corners have a smaller resistance compared with the rest of the chip. Nevertheless, the thermal profiles and maximum temperature in the two cases are close to each other, which allows us to use the simplified model. It is also beneficial to use a lumped model to make quick design guidelines for the thermal bridge.

2) Impact of Thermal Bridge Cooling Capability: The thermal bridge serves as the secondary heat path in the stack, so its cooling capability will influence the maximum temperature of the stack. If the thermal resistance of the bridge is too high, the heat of the memory cannot be removed, resulting in a higher temperature.

We sweep the thermal resistance of thermal bridge from 1 to 3 K/W and plot the maximum temperature of both dice. To model the worst case, the processor is assumed to be in active (high-power) state. We consider the no TSV case and the case with a medium number (2500) of uniformly distributed TSVs. Fig. 14 shows the results. When the thermal resistance of the bridge increases, there is not a big impact for the processor die, but the increased resistance leads to a rapid increase of the temperature of the memory. This is because the air gap decouples the two dice and the thermal bridge mainly cools the memory.

Fortunately in our proposed system, we use a separate (isolated) MFHS to cool the thermal bridge, and this leads

TABLE V
Comparison of Different Stacks

Unit:°C	T_{max} (Memory)		T_{max} (Processor)	
	Standby	Active	Standby	Active
Stack with Air cooled heat sink	50.33	75.06	51.63	76.44
Stack with interposer embedded MFHS	46.59	65.38	47.14	66.05
Proposed stack w/o TSVs	31.96	39.63	47.58	64.64
Proposed stack w TSVs	38.88	51.76	44.75	61.44

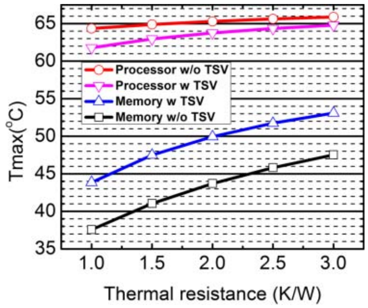


Fig. 14. Impact of the thermal bridge. The x-axis denotes the thermal resistance of the bridge and y-axis is the maximum temperature of the dice.

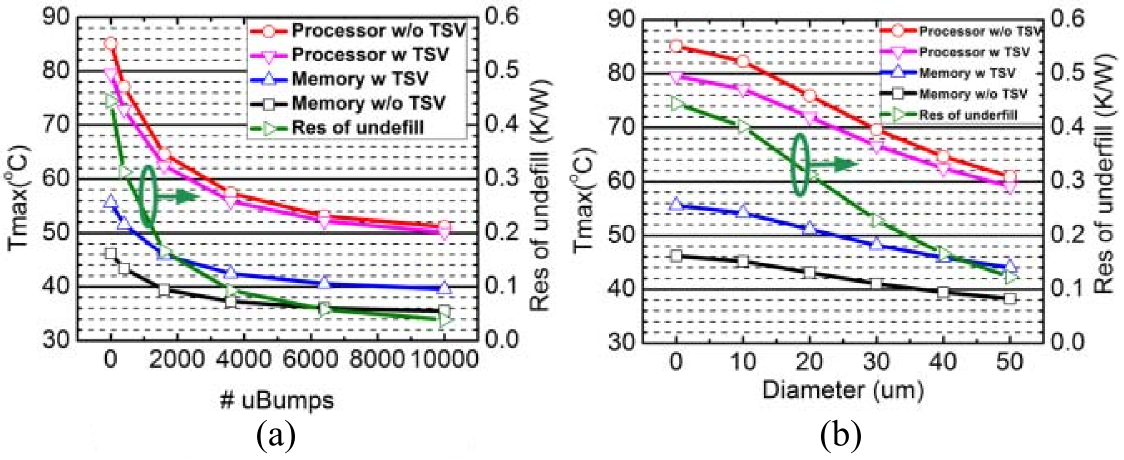


Fig. 15. Impact of the microbumps. (a) Changing the number of microbumps. (b) Changing the diameter of microbumps.

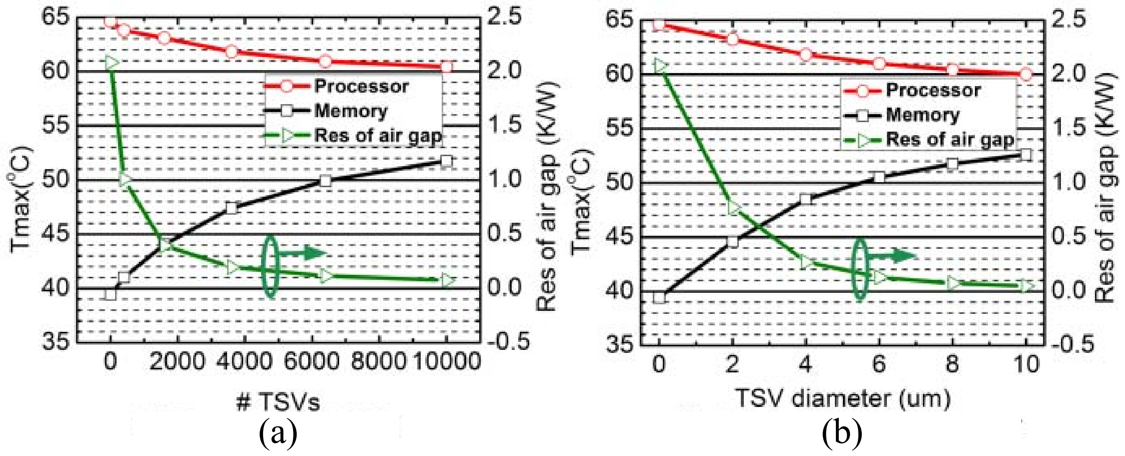


Fig. 16. Impact of the TSVs. (a) Changing the number of TSVs. (b) Changing the diameter of TSVs.

to a resistance of about 1.25 K/W; thus, the memory die can be cooled down effectively.

B. Microbump

Microbumps affect the primary heat path. Their diameter and total number influence the equivalent thermal resistance between the processor and the interposer.

First, we fix the diameter of the microbump to 40 μm and change the total number of microbumps from 1600 to 10 000. Second, we fix the total number of microbumps to 1600 and change their diameter from 10 to 50 μm. We also evaluate the cases where there are no TSVs and where there are 2500 TSVs between the processor and the memory dice.

The results are shown in Fig. 15. As expected, a larger number or diameter of microbumps leads to a decrease of the equivalent thermal resistance of the layer, which improves the primary heat path. With 3600 40-μm-diameter microbumps, the temperature of the processor die is below 60 °C, which is tolerable. In reality, we have a very fine pitch (50 μm) electrical microbumps between the chip and the interposer, which leads to a total of 40 000 microbumps; thus, the thermal requirement of the microbumps can be easily met.

C. TSVs

The air gap thermal isolation in our proposed structure enables dice in a 3-D stack to be thermally decoupled. However, 3-D stacks may require a large number of TSVs for interdie signaling and power and delivery, which will bridge the air gap and reduce its equivalent thermal resistance. As shown previously in Fig. 12, the inclusion of TSVs causes the temperature of the memory die to closely track that of the processor die.

1) TSV Number and Diameter: The diameter and number of TSVs impact the equivalent thermal resistance of the thermal isolation air gap. To quantify this impact, we perform two experiments. First, we fix the TSV diameter to 5 μm with a 0.5 μm liner and sweep the TSV number from 1600 to 10 000. Then, we fix the total TSV number at 10 000 and sweep the TSV diameter from 2 μm to 10 μm with a fixed 0.5-μm liner.

Fig. 16(a) and (b) shows the impact of the TSV number and diameter, respectively. As the TSV total volume increases, the air gap isolation layer becomes more thermally conductive, and the interdie heat coupling becomes stronger, reducing the temperature difference between the two dice.

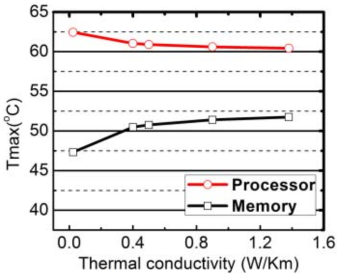
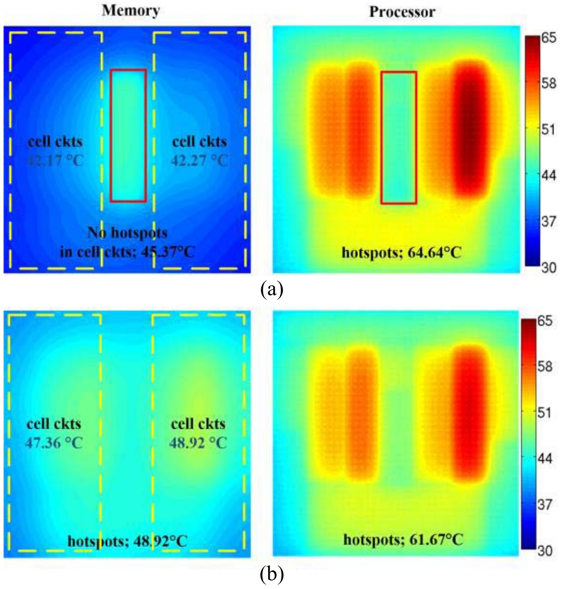


Fig. 17. Impacts of the TSV liner. The x-axis denotes the thermal conductivity of the liner.



(a)

(b)

Fig. 18. Thermal maps of clustered and uniform TSVs. (a) TSVs are clustered in the solid-line box. (b) Same amount of TSV are uniformly distributed.

If 2-μm-diameter TSVs are used rather than 10 μm, the memory temperature is only 44 °C, compared with 54.5 °C for the 10 μm TSV case. Further scaling of the TSV dimensions will yield additional improvements in the thermal isolation of the air gap.

2) TSV Liner: The TSVs are covered by the liner so the properties of the TSV liner are also thermally important. If the liner consists of thermally resistive materials, the isolation will be improved. To investigate this effect, we run several simulations with different liner thermal conductivities. The results are shown in Fig. 17. Using a low conductivity liner reduces the temperature of the memory die, which helps the thermal isolation.

3) TSV Distribution: In 3-D DRAM stacks or wide I/O applications, the TSVs are usually clustered instead of uniformly distributed. When the TSVs are clustered in a certain area, thermal coupling is expected to occur only in that area. In this way, the heat from the processor die will be localized.

For the memory die, the clustered TSVs or TSV farm usually acts as the I/O pins and are outside the memory cell circuits [labelled by a dashed-line box in Fig. 18(a)]. Hence, the memory cell circuits will become relatively free from the impact of the processor because there are no TSVs in this area.

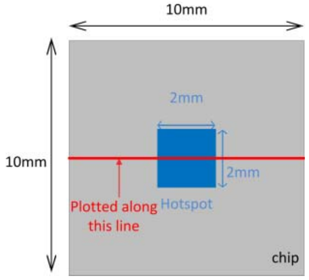


Fig. 19. Power map of the processor die. The hotspot (blue square) has power density of 135 W/cm^2 and the background (gray area) power density is 35 W/cm^2.

Inspired by the above analysis, we cluster the TSVs only in the center. The farm is assumed to be 1 mm × 5 mm and is assumed to have 49 × 100 TSVs, which are labeled by the solid-line rectangle in Fig. 18(a). To make a fair comparison, we also consider a uniformly distributed TSVs case with 4900 TSVs. The results are shown in Fig. 18(b).

In the clustered TSVs case, the maximum temperature of the whole DRAM die drops by 3.55 °C compared with the uniform TSV case. However, the maximum temperature of the cell array circuits is only 42.27 °C, which is a drop of 6.65 °C and is much closer to 39.63 °C of the ideal case without TSVs. By clustering the TSVs far from the memory cells, the most thermally sensitive portion of the die is effectively isolated from the high-power die.

Thus, we conclude that clustering TSVs can localize thermal coupling in 3-D stacks.

D. Die Thickness

The silicon substrate itself serves a useful role as a heat spreader, further reducing the impact of localized hotspots. Thus, as the die thickness scales down, it becomes very difficult to spread the heat of the hotspot due to increased lateral thermal resistance. In our proposed system, due to the air gap, the temperature of the stack will be more sensitive to die thickness.

In our test case, we assume there is a 2 mm × 2 mm 135 W/cm^2 hotspot in the center of the processor die and the background power density is 35 W/cm^2. The memory die has a uniform power density of 1 W/cm^2. The power map of the processor die is shown in Fig. 19. We separately sweep the die thickness of the memory and processor from 1 μm to 100 μm while fixing the other die at the default thickness (50 μm for processor and 100 μm for memory, respectively). We also compare the results with the conventional bonding case (using underfill).

The impact of the processor die thickness is shown in Fig. 20(a). Several observations can be made. First, thinning the processor die will increase the temperature of both dice, especially when there is air gap isolation. With air gap thermal isolation, the maximum temperature of a 100-μm-thick processor die is only 56 °C, while for an ultrathin 1-μm-thick die, the maximum temperature is 78 °C.

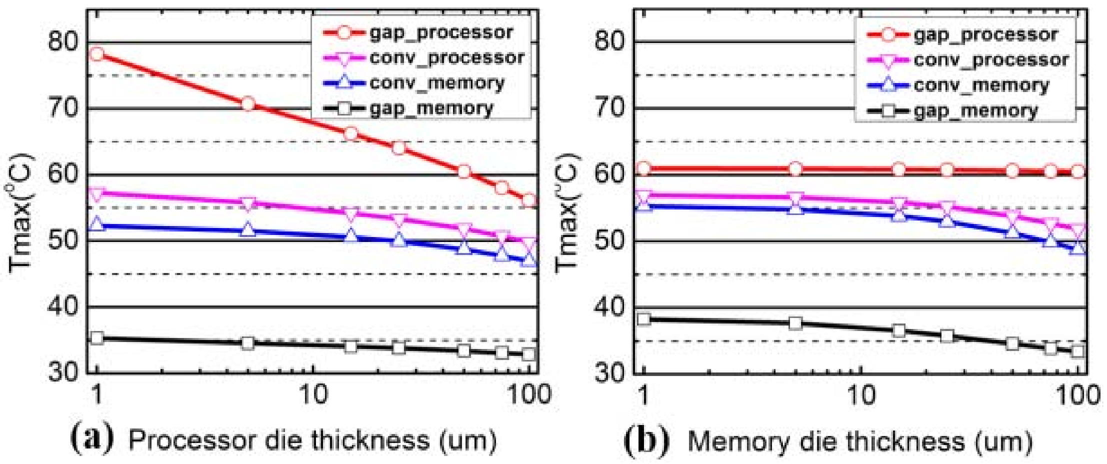


Fig. 20. Impact of die thickness. (a) Changing the thickness of processor. (b) Changing the thickness of memory.

Second, if there is an air gap, when the processor die is thinned below 50 μm, the processor temperature increases very fast. Third, for the conventional bonding case, the memory die serves as the heat spreader for both dice. In this case, thinning the processor die has a limited impact on system temperature, as shown by the two intermediate lines in Fig. 20(a).

For the impact of the memory die thickness shown in Fig. 20(b), when the air gap exists, because of the changing thermal isolation, the memory die thickness has only a small impact on the processor. For the conventional bonding case, the processor and the memory dice are strongly coupled, which yield results similar to changing the processor die thickness.

VII. Conclusion

In this paper, we propose a novel stacking structure with microfluidic cooling embedded in the interposer, thermal isolation between the memory and processor dice, and a thermal bridge on the top of the memory die. The new architecture exhibits thermal benefits over conventional stacks and is of high value in the heterogeneous integration of high-power and low-power dice. In addition, we thermally explore our proposed system as a function of microbumps, TSVs, die thickness, and other system parameters. Specifically, this paper benchmarks a memory on processor stack, but the methodologies, analysis, and conclusions can be applied to any high-power and low-power stack.

Second, tuning of all system parameters is necessary to build a thermally tolerant system: 1) the microbumps influence the primary heat path and must be considered when assessing the thermal performance of the system; 2) the TSVs have important impact on the thermal isolation layer and with smaller/less/clustered TSVs, the thermal coupling between dice can be minimized; and 3) die thickness also plays an important role. Thinning the processor die below 50 μm will lead to a rapid temperature increase of the stack.

Several key conclusions can be drawn from this paper. First, our proposed stack can realize lower temperatures for both dice in a memory on processor stack. More importantly, the use of an air gap for thermal isolation causes the memory die to be maintained at lower temperature and to be fairly independent of the fluctuating temperature of the processor because of the thermal isolation.

References

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

[2] U. Kang *et al.*, "8Gb 3D DDR3 DRAM using through-silicon-via technology," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Feb. 2009, pp. 130–132.

[3] T. Zhang *et al.*, "A 3D SoC design for H.264 application with on-chip DRAM stacking," in *Proc. IEEE 3D Syst. Integr. Conf. (3DIC)*, Nov. 2010, pp. 1–6.

[4] P. Leduca *et al.*, "Challenges for 3D IC integration: Bonding quality and thermal management," in *Proc. IEEE Int. Interconnect Technol. Conf.*, Jun. 2007, pp. 210–212.

[5] J. Cong, W. Jie, and Y. Zhang, "A thermal-driven floorplanning algorithm for 3D ICs," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design (ICCAD)*, Nov. 2004, pp. 306–313.

[6] Y. Zhang, A. Dembla, Y. Joshi, and M. Bakir, "3D stacked microfluidic cooling for high-performance 3D ICs," in *Proc. IEEE Electron. Compon. Technol. Conf.*, May/Jun. 2012, pp. 1644–1650.

[7] H. Oprins, V. O. Cherman, B. Vandevelde, G. Van der Plas, P. Marchal, and E. Beyne, "Numerical and experimental characterization of the thermal behavior of a packaged DRAM-on-logic stack," in *Proc. IEEE 62nd Electron. Compon. Technol. Conf. (ECTC)*, May/Jun. 2012, pp. 1081–1088.

[8] J. Liu, B. Jaiyen, R. Veras, and O. Mutlu, "RAIDR: Retention-aware intelligent DRAM refresh," in *Proc. IEEE/ACM 39th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2012, pp. 1–12.

[9] Y. Zhang, H. Oh, and M. S. Bakir, "Within-tier cooling and thermal isolation technologies for heterogeneous 3D ICs," in *Proc. IEEE 3D Syst. Integr. Conf. (3DIC)*, Oct. 2013, pp. 1–6.

[10] J. Liu, B. Jaiyen, Y. Kim, C. Wilkerson, and O. Mutlu, "An experimental study of data retention behavior in modern DRAM devices: Implications for retention time profiling mechanisms," in *Proc. IEEE/ACM 40th Int. Symp. Comput. Archit.*, Jun. 2013, pp. 60–71.

[11] S. Manipatruni *et al.*, "Wide temperature range operation of micrometer-scale silicon electro-optic modulators," *Opt. Lett.*, vol. 33, no. 19, pp. 2185–2187, 2008.

[12] Z. Li *et al.*, "Reliability modeling and management of nanophotonic on-chip networks," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 1, pp. 98–111, Jan. 2012.

[13] D. Brunina, D. Liu, and K. Bergman, "An energy-efficient optically connected memory module for hybrid packet- and circuit-switched optical networks," *IEEE J. Sel. Topics Quantum Electron.*, vol. 19, no. 2, Mar./Apr. 2013, Art. ID 3700407.

[14] R. Nadipalli, J. Fan, K. H. Li, K. H. Wee, H. Yu, and C. S. Tan, "3D integration of MEMS and CMOS via Cu-Cu bonding with simultaneous formation of electrical, mechanical and hermetic bonds," in *Proc. IEEE Int. 3D Syst. Integr. Conf. (3DIC)*, Jan./Feb. 2012, pp. 1–5.

[15] K. L. Wang, Y. Li, and C. Rizos, "The effect of the temperature-correlated error of inertial MEMS sensors on the integration of GPS/INS," in *Proc. IGNSS Symp.*, Surfers Paradise, Qld, Australia, 2009, paper 35.

[16] D. L. DeVoe, "Thermal issues in MEMS and microscale systems," *IEEE Trans. Compon. Packag. Technol.*, vol. 25, no. 4, pp. 576–583, Dec. 2002.

[17] K. Matsumoto, S. Ibaraki, M. Sato, K. Sakuma, Y. Orii, and F. Yamada, "Investigations of cooling solutions for three-dimensional (3D) chip stacks," in *Proc. IEEE 26th Annu. Semiconductor Thermal Meas. Manage. Symp.*, Feb. 2010, pp. 25–32.

[18] D. Kearney, "A numerical model of an inter-strata liquid cooling solution for a 3D IC architecture," in *Proc. IEEE 16th Int. Workshop Thermal Invest. ICs Syst. (THERMINIC)*, Oct. 2010, pp. 1–6.

[19] K. Sikka, J. Wakil, H. Toy, and H. Liu, "An efficient lid design for cooling stacked flip-chip 3D packages," in *Proc. 13th IEEE Intersoc. Conf. Thermal Thermomech. Phenomena Electron. Syst.*, May/Jun. 2012, pp. 606–611.

[20] K. Athikulwongse, M. Pathak, and S.-K. Lim, "Exploiting die-to-die thermal coupling in 3D IC placement," in *Proc. ACM/EDAC/IEEE 49th Design Autom. Conf. (DAC)*, Jun. 2012, pp. 741–746.

[21] J.-S. Kim *et al.*, "A 1.2 V 12.8 GB/s 2 Gb mobile wide-I/O DRAM with 4 × 128 I/Os using TSV based stacking," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Jan. 2011, pp. 107–116.

[22] J. Knechtel, I. L. Markov, J. Lienig, and M. Thiele, "Multiobjective optimization of deadspace, a critical resource for 3D-IC integration" in *Proc. Int. Conf. Comput.-Aided Design (ICCAD)*, San Jose, CA, USA, Nov. 2012, pp. 705–712.

[23] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, D. Tarjan, and K. Sankaranarayanan, "Temperature-aware microarchitecture," in *Proc. IEEE/ACM 30th Int. Symp. Comput. Archit. (ISCA)*, May 2003, pp. 2–13.

[24] J. Meng, K. Kawakami, and A. K. Coskun, "Optimizing energy efficiency of 3-D multicore systems with stacked DRAM under power and thermal constraints," in *Proc. ACM/EDAC/IEEE 49th Design Autom. Conf. (DAC)*, Jun. 2012, pp. 648–655.

[25] Y. Zhan and S. S. Sapatnekar, "A high efficiency full-chip thermal simulation algorithm," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2005, pp. 635–638.

[26] J. Xie and M. Swaminathan, "Fast electrical-thermal co-simulation using multigrid method for 3D integration," in *Proc. IEEE 62nd Electron. Compon. Technol. Conf. (ECTC)*, May/Jun. 2012, pp. 651–657.

[27] J. Xie and M. Swaminathan, "Electrical-thermal co-simulation of 3D integrated systems with micro-fluidic cooling and Joule heating effects," *IEEE Trans. Compon., Packag. Manuf. Technol.*, vol. 1, no. 2, pp. 234–246, Feb. 2011.

[28] S. Lakshminarayanan, P. J. Wright, and J. Pallinti, "Electrical characterization of the copper CMP process and derivation of metal layout rules," *IEEE Trans. Semicond. Manuf.*, vol. 16, no. 4, pp. 668–676, Nov. 2003.

[29] 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 *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2012, pp. 14.21–14.2.4.

[30] J. Xie and M. Swaminathan, "DC IR drop solver for large scale 3D power delivery networks," in *Proc. IEEE 19th Conf. Elect. Perform. Electron. Packag. Syst. (EPEPS)*, Oct. 2010, pp. 217–220.

[31] Dow Corning. *Dow Corning TC-5622 Thermally Conductive Compound*. [Online]. Available: http://www.dowcorning.com/DataFiles/090276fe801909c2.pdf, accessed Nov. 2014.

[32] J.-J. Park and M. Taya, "Design of thermal interface material with high thermal conductivity and measurement apparatus," *J. Electron. Packag.*, vol. 128, no. 1, pp. 46–52, 2006.

[33] *DDR3 SDRAM System-Power Calculator*, Micron Technology, Boise, ID, USA, Jul. 2011.

[34] Intel. *3rd Generation Intel® Core Processor Family Quad Core Launch Product Information*. [Online]. Available: http://download.intel.com/newsroom/kits/core/3rdgen/pdfs/3rd_Generation_Intel_Core_Product_Information.pdf, accessed Nov. 2014.

[35] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures," in *Proc. 42nd Annu. IEEE/ACM Int. Symp. Microarchit.*, Dec. 2009, pp. 469–480.

Yang Zhang (S'13) received the B.S. degree in microelectronics and mathematics 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.

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

Muhannad S. Bakir (SM'12) received the B.E.E. (*summa cum laude*) degree from Auburn University, Auburn, AL, USA, in 1999, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology (Georgia Tech), Atlanta, GA, USA, in 2000 and 2003, respectively.

He is currently an Associate Professor and the ON Semiconductor Junior Professor with the School of Electrical and Computer Engineering, Georgia Tech.

Dr. Bakir is a member of the International Technology Roadmap for Semiconductors Technical Working Group for Assembly and Packaging. He was a recipient of the Intel Early Career Faculty Honor Award in 2013, the DARPA Young Faculty Award in 2012, and the IEEE CPMT Society Outstanding Young Engineer Award in 2011. He was also a recipient of the Semiconductor Research Corporation Inventor Recognition Awards in 2002, 2005, and 2009. He and his research group have received 14 conference and student paper awards, including five 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. He was an Invited Participant in the 2012 National Academy of Engineering Frontiers of Engineering Symposium. He is an Editor of the IEEE TRANSACTIONS ON ELECTRON DEVICES, an Associate Editor of the IEEE TRANSACTIONS ON COMPONENTS, PACKAGING AND MANUFACTURING TECHNOLOGY, and was a Guest Editor of the 2011 Special Issue of the IEEE JOURNAL OF SELECTED TOPICS IN QUANTUM ELECTRONICS.