Thermal Impact Study of Block Folding and Face-to-Face Bonding in 3D IC

Yarui Peng[1], Moongon Jung[2], Taigon Song[1], Yang Wan[3], , and Sung Kyu Lim[1]
[1]School of ECE, Georgia Institute of Technology, Atlanta, GA, USA
[2]Intel Corp., Santa Clara, CA, USA [3]Google Inc., Mountainview, CA, USA
yarui.peng@gatech.edu

***Abstract*—In this paper we study the thermal impact of two high impact design/technology choices for 3D ICs, namely, block folding and face-to-face bonding. A recent study shows that block folding and face-to-face improve wirelength, power, and performance, but the impact on thermal issue is not studied. Based on commercial-quality 3D IC layouts of large-scale OpenSPARC T2 designs and a highly accurate GDSII-level thermal analysis flow, our results first show that block folding, despite its power density increase, does not worsen thermal issues because of the TSVs that act as heat conductors. In addition, face-to-face bonding, despite its thermal benefit from the absence of BCB bonding layer and underfill, still does not improve temperature much because of the small F2F via sizes.**

I. Introduction

3D ICs can provide more functionality in a smaller footprint area by stacking multiple dies in vertical direction. However, it raises more concerns on the thermal impact. High temperature not only introduces thermal-induced stress into the chip, but also degrades device performance and increases leakage power. Therefore, understanding of the thermal properties in 3D ICs is required for design reliability and variation control.

Through silicon vias (TSVs) have been used as a vertical interconnection in 3D ICs. Unlike the traditional Face-to-Back (F2B) bonding, Face-to-Face (F2F) bonding utilizes F2F vias instead of TSVs for vertical interconnections and introduces smaller RC delay and power consumptions. Block folding, a 3D design technique, is also found to enhance power saving in 3D ICs by reducing wirelength and buffer count in [1]. However, due to the power density increase, the thermal impact of block-folding needs to be carefully studied. Various logic-memory stacking options were discussed in [2], but the thermal impact study is performed without considering the power and 3D connections.

However, there are few studies on the thermal impact of block-folding and 3D bonding style. A test chip 3D processor is fabricated using F2F technology [3], but no thermal analysis was provided. A recent study [1] shows that the 3D designs with F2F bonding provide benefits compared with F2B due to smaller via size and flexible placement, but the thermal impact is unknown.

The main contributions of this work include the following: (1) The impact of block folding on thermal is shown with large-scale OpenSPARC T2 designs. (2) Thermal impact of bonding styles, i.e., F2B and F2F, is studied considering power benefits and detailed layouts. To the best of our knowledge, this is the first work that studies the thermal impact of bonding style and block folding using large-scale 3D designs.

This work is supported by Intel Corporation through Semiconductor Research Corporation (ICSS Task 2293) and the Center for Integrated Smart Sensors funded by the Ministry of Science, ICT & Future Planning of the Korean Government under the Global Frontier Project (CISS-2012366054194).

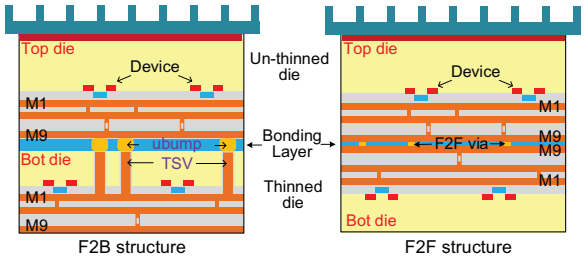


Fig. 1. Thermal structure of F2B and F2F bonding.

II. Thermal Analysis Flow

The structures of F2B and F2F are shown in Figure 1. In this work, TSVs are $3\mu m$ in radius while the F2F vias are $0.5\mu m$. Since the sizes of F2F vias are much smaller than TSVs, they introduce less overhead compared with F2B structure in terms of delay, area and power consumption. In F2B structure, a BCB layer is often used as an adhesive between dies since they provide a cost-effective solution to form a strong and reliable bonding. However, the thermal conductivity of BCB is very low and results in a limitation of vertical heat flow. On the other hand, F2F technology uses a direct copper bonding with no adhesive. The background material of the bonding layer is SiO_2 which has about 5 times larger thermal conductivity than BCB. Both of these improve the thermal conductivity of F2F bonding layer.

3D IC thermal analysis tools such as 3D-ICE[4], which takes a few parameter to compute the layer thermal conductivity and a floorplan of the design, can only be used to obtain a thermal estimation results in early-stage IC design. To accurately study the thermal impact, we first build a mesh structure where each layer contains thousands of thermal cells that are elements when solving the differential equations. Then a layout analyzer is built to read all the layout information including gates, wires, and TSVs from GDSII file. It calculates the thermal conductivity of each thermal cells based on the material portion within the thermal cell. A detailed power distribution map is generated for thermal analysis and heat sources are added to the device layers of each die. Finally, the mesh structure is imported into ANSYS Fluent that solves the thermal differential equations and obtains the thermal map of each layer.

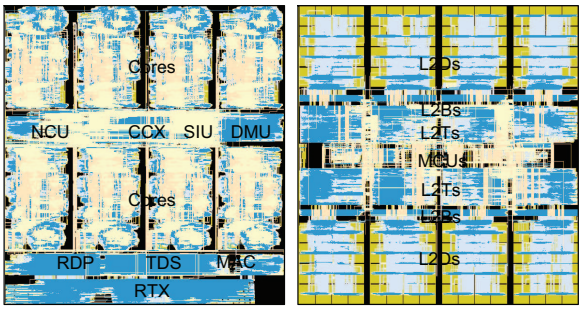
III. OpenSPARC T2 Designs

A commercial-grade 28nm 8-core OpenSPARC T2 processor is used in this study. We performed full-chip static timing analysis using Primetime and obtained the power consumption results for all the designs. For our thermal analysis, two benchmarks are used: an integer program “gcc”, and a float point program “spice”. Detailed design metrics are listed in Table I.

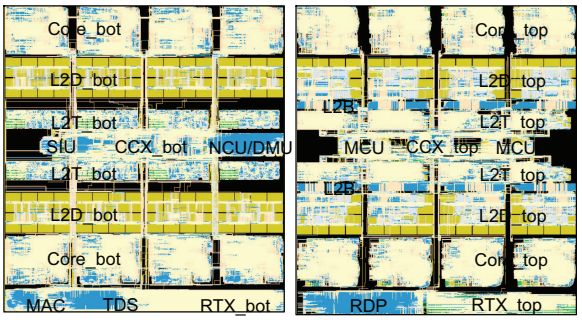
According to McPAT, the cores have much higher power than other modules, e.g., L2 caches (L2T, L2B, L2D) and memory control

TABLE I
DESIGN METRICS OF T2 DESIGNS BASED ON ISO-PERFORMANCE POWER COMPARISON. POWER IS REPORTED IN W.

Design	Single core				Full-chip			
Bonding	2D	F2B	F2B	F2F	2D	F2B	F2B	F2F
Folded?	no	no	yes	yes	no	no	yes	yes
LPD(ns)	1.52	1.50	1.48	1.44	2.05	2.02	1.99	1.97
Area(μm^2)	3.10	1.58	1.54	1.54	71.1	38.4	39.7	39.7
Buffer#	214k	128k	121k	114k	7.4M	7.0M	6.7M	6.6M
WL(m)	21.8	19.0	17.5	17.1	340	320	307	303
TSV #	0	2979	9551	10.3k	0	3263	69.1k	112k
gcc power	1.40	1.19	1.13	1.12	20.5	17.4	16.2	15.9
spice power	1.54	1.32	1.27	1.25	21.3	18.1	16.8	16.6



(a) 2-tier T2 with no folded blocks (TSV: 3263)



(b) 2-tier T2 with folded blocks (TSV: 69,091)

Fig. 2. GDSII layouts of full-chip T2 in 3D IC. (a) 2-tier design with F2B bonding, no block folding (6.0×6.4mm^2), (b) 2-tier design with F2B bonding with block folding (6.0×6.6mm^2).

units (MCUs). Detailed design metrics are listed in Table I. 27°C is assumed as the room temperature, and the temperature difference percentage is measured by temperature increase above the room temperature. Compared with 2D design. 3D design consumes smaller net power and uses fewer buffers due to wirelength reduction. This results in a 15.4% power reduction and more than 50% footprint reduction in 3D design. However, due to smaller power density and die thickness, 2D design shows much lower maximum temperature (45.4°C) than the 3D design (61.7°C). Therefore, the thermal issue in 3D T2 processor needs to be carefully considered.

A. Block Folding Approach

By partitioning the design into 3D, the long wire usage for inter-block connection can be reduced. Block folding serves as the second level of partitioning and reduces the footprint of each block. Thus the wirelength within folded blocks is reduced further and leads to smaller wire capacitance. For the full-chip design, eight cores, eight L2Ts, eight L2Ds, one RTX and one CCX module are folded. The inter-block wirelength also decreases since the blocks are smaller

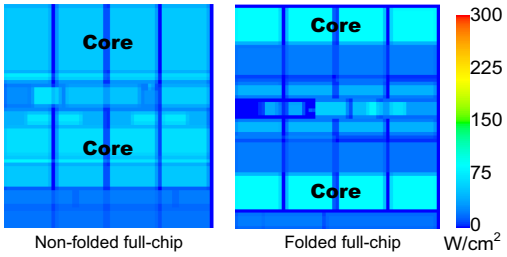


Fig. 3. Power map (gcc) comparison between non-folded and folded designs.

TABLE II
THERMAL IMPACT OF BLOCK-FOLDING.

Design	Benchmark	Folded?	Temperature range (°C) Bottom die	Top die
Full-chip	gcc	no	53.6~**61.7**	52.9~57.6
		yes	51.1~**59.2**	50.6~57.3
	spice	no	54.6~**63.2**	53.9~58.8
		yes	52.1~**60.7**	51.6~58.6

and have more flexibility in 3D floorplan. Therefore the block folding leads to wire capacitance and buffer number reduction and saves total power consumption. Compared to its 2D counterparts, a maximum of 21.0% power reduction is observed in F2B folded designs full-chip level.

B. F2F Bonding Design

F2F bonding is favored by many foundries due to the yield and cost. Compared with TSV, the F2F vias are much smaller. This results in parasitic capacitance and silicon area reduction. Also, since the overhead of F2F vias is much smaller, during the partition stage, we can focus on timing and power quality improvement. Thus in F2F design, more 3D vias are used to improve the overall design quality. On the other hand, unlike F2B bonding, where TSVs are placement blockages and cannot be placed over devices and hard macros, the F2F vias are routing blockages only on the top metal layers and thus they can be placed anywhere. The P&R tool has more flexility and better optimization opportunity. This also contributes to the design quality improvement in F2F designs. We implement F2F designs using our F2F via placement flow and the results are listed in Table I. Compared to F2B design, power is reduced by 1.8% due to shorter wirelength and less buffer count.

IV. THERMAL IMPACT OF BLOCK-FOLDING

A. Power Density Increase with Block-folding

Even though block folding reduces power consumption, it increases the maximum power density especially if the design folds high power density modules such as cores. For the non-folded design, since there are more flexibilities in floorplaning and placement, this problem can be solved by a thermal-aware floorplaning so that the hot spots of each die are not overlapping. However, tiers of a folded block need to be placed at the same location so that TSVs can be placed for vertical interconnection, the maximum power density is still much higher than that of non-folding designs even with power reduction considered. The power maps are shown in Figure 3 and power density increases by 72% in the core area.

B. Thermal analysis with Block-folding

Thermal analysis results are summarized in Table II and bottom die thermal maps are shown in Figure 4. Interestingly, the block-folding does not worsen thermal results, even though the maximum

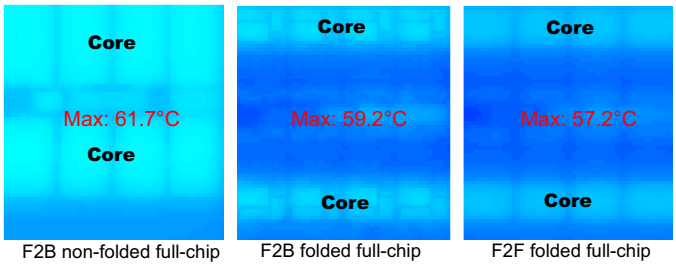


Fig. 4. Full-chip level bot die temperature map (gcc) comparison.

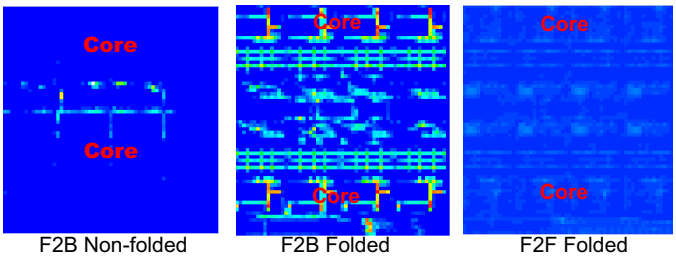


Fig. 5. Full-chip level bonding layer thermal conductivity map comparison. Red region has a thermal conductivity of 80 W/m/K.

power density increased significantly. In all cases, the maximum temperature of a folded design is in a similar range of its non-folded counterpart. This is because by block-folding, physical design impacts help compensate the hot spots overlapping impact. First impact comes from power partitioning. With block-folding, half of the high power density blocks is moved into the top die. Therefore, block-folding has similar effects as swapping dies. Secondly, the number of TSVs is increased using block folding since more TSVs are inserted within each folded blocks. For T2 full-chip, TSV count increases to 21.2 times, and thus the additional TSVs show a much higher impact on full-chip temperature. These additional TSVs have a much higher thermal conductivity than silicon. Thus it improves the vertical heat flow and makes heat dissipation easier.

Another impact comes from the TSV location. Figure 5 shows the thermal conductivity map. In non-folded designs, TSVs are placed at the boundary of each block, which introduces a longer path from heat source to TSVs. This results in a larger intra-die temperature variation, where the functional blocks are hot spots while TSV farms are cooler spots. However, in block-folding case, TSVs are placed inside each blocks, thus it results in a shorter lateral heat dissipation path and cools the block more evenly. Moreover, since any signal TSVs are paired with microbumps, they improve the thermal benefit further. With block-folding, the thermal conductivity of the bonding layer increases. Finally, the overall power consumption decreases in block folding design and this leads to an average temperature reduction for both dies.

V. Thermal Impact of F2F Bonding

A. Bonding Layer with F2F Bonding

As discussed in II, in F2F structure, a direct copper bonding is used instead of a BCB layer. This leads to a background thermal conductivity improvement in the bonding layer. Also, the bonding layer thickness in F2F structure is thinner. Therefore, F2F bonding has a smaller limitation on vertical heat flow. Since both metal layers and F2F bonding layer have the same background materials, F2F bonding layer is no longer the bottle neck in vertical heat flow.

TABLE III
Thermal impact of F2F bonding.

Design	Benchmark	Bonding	Temperature range (°C) Bottom die	Top die
Full-chip	gcc	F2B	51.1~**59.2**	50.6~57.3
		F2F	50.8~**57.2**	50.6~56.5
	spice	F2B	52.1~**60.7**	51.6~58.6
		F2F	51.8~**58.6**	51.6~57.8

The thermal conductivity maps are shown in Figure 5. The background thermal conductivity in F2F structure is 4.75 times larger than that in F2B. However, F2F vias are much smaller than TSVs. This is an advantage for physical design, but not for thermal results. F2F vias introduce less copper into the bonding layer than the microbumps. Thus, in regions where TSVs are located, the F2B designs have better thermal conductivity than its F2F counterparts. The overall thermal impacts of F2F bonding depends on design and technology implementation.

B. Thermal analysis with F2F Bonding

Thermal analysis results of F2F designs are summarized in Table III, and the thermal maps are shown in Figure 4. First we observe that whether the 3D designs are folded or not, designs using F2F show a lower maximum temperature than its F2B counterparts. This is because a better vertical heat flow and a lower power consumption using F2F bonding. Compared to F2B design, the bonding layer temperature drop and die-to-die temperature variation are smaller. For all the cases, F2F shows smaller die-to-die temperature difference. Any temperature reduction in the bottom die results in a temperature increase on the top die.

Moreover, F2F bonding and block-folding help each other in a complementary fashion. Using F2F bonding does not degrade the benefits coming from block-folding and vice versa. This is because their impacts on wirelength reduction come from different ways and power reduction is the most by using both F2F and block-folding. Therefore, we conclude that F2F will not degrade thermal quality and it reduces maximum temperature with better vertical heat flow and lower power consumption. However, this improvement is small since F2F via is much smaller than TSV.

VI. Conclusion

In this paper, we demonstrate the thermal impact of block-folding and F2F bonding using a commercial-grade OpenSPARC T2 design. We implement our designs with TSV and F2F placer and calculate power consumption based on real design quality. Results show that block folding, despite its power density increase, does not worsen thermal issues because of the TSVs that act as heat conductors. F2F bonding, despite its thermal benefit from the absence of BCB bonding layer and underfill, still does not improve temperature much due to the small F2F via sizes.

References

[1] M. Jung *et al.*, "On enhancing power benefits in 3D ICs: Block folding and bonding styles perspective," in *Design Automation Conference (DAC), 2014 51st ACM/EDAC/IEEE*, June 2014, pp. 1–6.

[2] B. Black *et al.*, "Die Stacking (3D) Microarchitecture," in *International Symposium on Microarchitecture*, Dec 2006, pp. 469–479.

[3] D. Kim *et al.*, "Design and Analysis of 3D-MAPS (3D Massively Parallel Processor with Stacked Memory)," *IEEE Transactions on Computers*, vol. PP, no. 99, pp. 1–1, 2013.

[4] A. Sridhar *et al.*, "3D-ICE: A Compact Thermal Model for Early-Stage Design of Liquid-Cooled ICs," *Computers, IEEE Transactions on*, vol. 63, no. 10, pp. 2576–2589, Oct 2014.