Fine-Grained 3-D IC Partitioning Study With a Multicore Processor

Moongon Jung, *Member, IEEE*, Taigon Song, *Student Member, IEEE*, Yarui Peng, *Student Member, IEEE*, and Sung Kyu Lim, *Senior Member, IEEE*

***Abstract*— Low power is widely considered as a key benefit of 3-D integrated circuits (ICs), yet there have been few thorough design studies on how to maximize power benefits in 3-D ICs. In this paper, we present design methodologies to reduce power consumption in 3-D ICs using a large-scale commercial-grade multicore microprocessor (OpenSPARC T2). To further improve power benefits in 3-D ICs on the top of the traditional 3-D floorplanning, we evaluate the impact of 3-D IC partitioning: block folding and bonding styles. In addition, the impact of block folding and bonding style on 3-D thermal is investigated. Last, we examine the power distribution network impact on 3-D power benefit. With aforementioned methods combined, our 3-D designs provide up to 21.7% power reduction over the 2-D counterpart under the same performance.**

***Index Terms*— 3-D integrated circuits (ICs), block folding, bonding style, power benefit, power distribution network (PDN), thermal analysis.**

I. Introduction

POWER reduction has been one of the most critical design considerations for integrated circuit (IC) designers. Minimizing both dynamic and leakage power is imperative to meet power budgets for both low-power and high-power applications. The power efficiency also directly affects IC's packaging and cooling costs. In addition, the power of an IC has a significant impact on its reliability and manufacturing yield.

Because of the increasing challenges in achieving efficiency in power, performance, and cost beyond 32–22 nm, the industry began to look for alternative solutions. This has led to the active research, development, and deployment of thinned and stacked 3-D ICs with Through Silicon Vias (TSVs). Black *et al.* [1] studied the potential to achieve 15% power reduction as well as 15% performance gain of a high-performance microprocessor by a 3-D floorplan. Kang *et al.* [2] demonstrated 25% dynamic and 50% leakage power reduction in 3-D DRAM.

Manuscript received April 14, 2015; revised June 19, 2015; accepted July 19, 2015. Date of publication September 8, 2015; date of current version October 2, 2015. This work was supported by Intel Corporation through Semiconductor Research Corporation under Grant ICSS Task 2293. Recommended for publication by Associate Editor B. Dang 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: moongon@gatech.edu; taigon.song@gatech.edu; yarui.peng@gatech.edu; limsk@ece.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.2015.2470124

Most of the previous works showed 3-D power benefit by 3-D floorplanning. In this paper, we present a fine-grained 3-D IC partitioning study to enhance 3-D power benefit with a multicore processor example. Specifically, we investigate 3-D block folding methods to further reduce power in 3-D ICs on the top of the traditional 3-D floorplanning. We also study impacts of bonding styles, i.e., face to back (F2B) and face to-face (F2F), on 3-D power consumption. In addition, we examine how the block folding and bonding style affect 3-D IC thermal. The impact of power distribution network (PDN) on 3-D power benefit is also examined.

Our study is based on the OpenSPARC T2 [an 8-core 64-b SPARC system-on-a-chip] design database [3] and a Synopsys 28-nm PDK with nine metal layers [4]. We build GDSII-level 2-D and two-tier 3-D layouts, analyze, and optimize designs using the standard sign off CAD tools.

Based on this design environment, we first discuss how to rearrange blocks into 3-D to reduce power in Section II. In Section III, we explore block folding methods, i.e., partitioning a block into two sub-blocks and bonding them, to achieve power savings in the 3-D design. Then, we study how bonding styles affect the folded design quality. In Section IV, we demonstrate the system-level 3-D power benefits by assembling folded blocks in different bonding scenarios. Next, the impact of block folding and bonding style on the 3-D IC thermal is discussed in Section V. Finally, in Section VI, the impact of PDN on 3-D power benefit is examined.

II. Simulation Settings

A. Benchmark

The OpenSPARC T2, an open source commercial microprocessor from Sun Microsystems with 500 million transistors used, consists of 53 blocks, including eight SPARC cores (SPC), eight L2-cache data (L2D) banks, eight L2-cache tags (L2Ts), eight L2-cache miss buffers (L2Bs), and a cache crossbar (CCX). Each block is synthesized with 28-nm cell and memory macrolibraries. Seven blocks that do not directly affect the CPU performance are dropped from our implementation, including five SerDes blocks, an electronic fuse, and a miscellaneous I/O unit. In addition, the phase-locked loop (analog block) in a clock control unit is replaced by ideal clock sources. Thus, a total of 46 blocks are floorplanned. For the 2-D design, we try to follow the original T2 floorplan [5] as much as possible, as shown in Fig. 7(a). In addition, special cares are taken to use both

2156-3950 © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

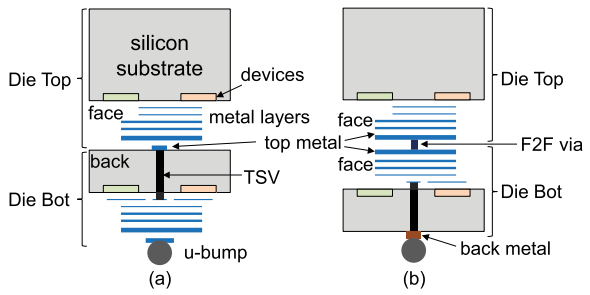


Fig. 1. Die bonding styles. (a) F2B. (b) F2F.

connectivity and data flow between blocks to minimize interblock wirelength.

B. 3-D IC Design Flow

Our RTL-to-GDSII tool chain for the 3-D IC design is based on commercial tools and enhanced with our in-house tools to handle TSVs and 3-D stacking. With initial design constraints, the entire 3-D netlist is synthesized. The layout of each die is done separately based on the 3-D floorplanning result. With given target timing constraints, cells and memory macros are placed in each block. Note that we only utilize regular-Vth (RVT) cells as a baseline. The netlists and the extracted parasitic files are used for the 3-D static timing analysis using Synopsys PrimeTime to obtain new timing constraints for each block's I/O pins as well as die boundaries (=TSVs).[1]

With these new timing constraints, we perform block-level and chip-level timing optimizations (buffer insertion and gate sizing) as well as power optimizations (gate sizing) using Cadence Encounter. We improve the design quality through iterative optimization steps, such as pre-clock tree synthesis (CTS), post-CTS, and postroute optimizations. In our design implementations, most blocks showed the shorter wirelength and the lower power consumption with more metal layers available for routing. However, for interblock routing, top metal layers are needed, especially for long nets. Thus, we utilize seven metal layers for all blocks except the SPC design that requires the most routing resources. Hence, the top two metal layers can be utilized for over-the-block routing in the chip-level design.

C. Die Stacking Technology

In this paper, we design two-tier 3-D ICs. As shown in Fig. 1, two possible bonding styles for 3-D ICs are used: F2B and F2F. In F2B bonding, TSVs are used for interdie connections. Thus, the number of 3-D connections can be limited by the TSV pitch as well as the TSV area overhead. The F2F bonding employing F2F vias is another attractive technology, as this does not require additional silicon area for 3-D connections.

Our 3-D interconnect settings are summarized in Table I. TSV resistance and capacitance values are calculated based on the model in [6]. We assume that TSV diameter is much larger than F2F via size as manufacturing reliable submicrometer TSVs is challenging. Additionally, the physical size of F2F via can be made comparable with the top metal dimension, around twice the minimum top metal (M9) width in our setup.

[1]TSVs are treated as normal I/O pins, and TSV RC values using π-model are included in parasitic files for 3-D static timing anlaysis (STA).

TABLE I
3-D INTERCONNECT SETTINGS

	diameter (μm)	height (μm)	pitch (μm)	R (Ω)	C (fF)
TSV	3	18	6	0.043	8.4
F2F via	0.5	0.38	1	0.1	0.2

TABLE II
COMPARISON BETWEEN 2-D AND 3-D BLOCK-LEVEL DESIGNS WITH A TARGET CLOCK FREQUENCY OF 500 MHz. NUMBERS IN PARENTHESES ARE DIFFERENCES AGAINST THE 2-D DESIGN

	2D	3D (core/cache)	3D (core/core)
footprint (mm^2)	71.1	38.4 (-46.0%)	38.4 (-46.0%)
# cells ($\times 10^6$)	7.39	7.21 (-2.4%)	7.26 (-1.8%)
# buffers ($\times 10^6$)	2.89	2.42 (-16.3%)	2.45 (-15.2%)
Wirelength (m)	343.0	326.0 (-5.0%)	324.5 (-5.4%)
Total power (W)	**9.107**	**8.171 (-10.3%)**	**8.273 (-9.1%)**
Cell power (W)	1.779	1.502 (-15.6%)	1.537 (-13.6%)
Net power (W)	4.499	4.122 (-8.4%)	4.131 (-8.2%)
Leakage power (W)	2.828	2.547 (-9.9%)	2.605 (-7.9%)

D. Baseline Design: 3-D Floorplan Without Block Folding

The T2 chip contains eight copies of SPC and L2-cache blocks (L2D, L2T, and L2B) that occupy most of the chip area. These blocks need to be arranged in a specific order and a regular fashion for communication between them. Considering this constraint, area balance between dies, and connectivity between blocks, the T2 netlist is partitioned into two dies. We design two 3-D floorplan cases to examine their impact on power.

1) *Core/Cache Stacking:* All cores are in one die, and all L2D blocks are in another die, as shown in Fig. 7(b).
2) *Core/Core Stacking:* Four cores and L2-cache blocks are located in each die.

We use the F2B bonding style only for the 3-D block-level designs as a baseline. The 3-D floorplanner in [7] is modified to handle user-defined floorplans, and then used to determine TSV locations with an objective of minimizing interblock wirelength. TSV arrays are treated as additional blocks in this flow, and hence all TSVs can be placed outside blocks only.

We compare our 2-D and 3-D block-level designs with a target CPU clock frequency of 500 MHz that is the highest performance that our 2-D design achieves.[2]

Design metrics in 2-D and 3-D designs are shown in Table II. First, we observe 16.3% buffer count and 5% wirelength reduction in the core/cache 3-D design and 15.2% and 5.4% reduction in the core/core 3-D case compared with the 2-D counterpart. In addition, interblock wirelength

[2]Our designs run slower than OpenSPARC T2 that runs at 1.4 GHz [5]. This is mainly because some custom memory blocks are synthesized with cells, since a general memory compiler cannot afford these kinds of memories. Unfortunately, these synthesized memories are much larger and run slower than the memory macros generated by a memory compiler.

reduces by 15.6% (core/cache) and 17.8% (core/core), which is a direct consequence of 3-D floorplanning.

Second, most importantly, the 3-D designs reduce power consumption over the 2-D counterpart by 10.3% (core/cache) and 9.1% (core/core). We see that cell (15.6%) and leakage (9.9%) power reduction are far more than the cell count decrease (2.4%) in the core/cache 3-D design. This is because the 3-D design utilizes more smaller cells than the 2-D because of better timing, i.e., more positive timing slack in paths. With the positive slack, cells can be downsized in the 3-D design if this change still meets the timing constraint during power optimization stages.

This smaller cell size in the 3-D design also helps reduce net power consumption. The load capacitance of a driving cell is defined as the sum of wire capacitance and the input pin capacitance of the loading side, and hence the net power is defined as the sum of wire and pin power. Therefore, the wire power reduction is directly from reduced wirelength, and the pin power decrease is from the smaller cell size as well as the reduced cell count.

Third, the core/cache 3-D stacking case shows 1.2% smaller power consumption than the core/core case, which is essentially a negligible difference. This also indicates that there is not much room to further reduce power by 3-D floorplans only, since there are not many floorplan options for the T2 design that contains multiple large same-size blocks that need to be placed in a specific way.

We choose the core/cache case as a baseline 3-D block-level design as it consumes a little less power than the core/core case. In addition, this 3-D design will be better in terms of the thermal coupling between dies as SPCs (higher power density) and L2Ds (lower power density) are stacked, while SPCs are stacked in the core/core case.

III. Block Folding Strategies

So far, the block-level designs are implemented for both 2-D and 3-D designs. Thus, even in the 3-D designs, each block is located in the same die. In addition, TSVs are always outside blocks and used only for interblock connections. In this section, we examine the impact of block folding, i.e., partitioning a single block into two sub-blocks and connect them with TSVs for intrablock connections, on power consumption.

A. Block Folding Criteria

For the block folding to provide power saving, certain criteria need to be met. First, the target block is required to consume a high enough portion of the total system power. Otherwise, the power saving from the block folding could be negligible in the system level. Blocks that consume more than 1% of the total system power are listed in Table III. Note that the total power portion of SPC, L2D, and L2T is the average of corresponding eight blocks. Thus, SPC, L2D, and L2T are outstanding target blocks. In addition, RTX and CCX consume high power as a single block, and hence could provide a nonnegligible power benefit if folded.

Second, the net power portion of the target block needs to be high. If the block is cell and leakage power dominant, the wirelength reduction of the folded block may not reduce the total power noticeably. Therefore, SPC and CCX are attractive blocks to fold. L2D shows a relatively low net power portion compared with other blocks as L2D is the memory (and its power) dominated design that contains 512 kB (3216-kB memory macros in our implementation). Third, the target block needs to contain many long wires so that wirelength decrease, and hence net power reduction in the folded block can be maximized. In this paper, we define long wires as wires longer than the 100× standard cell height. We observe that SPC, RTX, and CCX have a large number of long wires.

TABLE III
2-D Design Characteristics of Folding Candidate Blocks. Long Wires Are Defined as Wires Longer Than 100× Standard Cell Height. CPU Clock Runs at 500 MHz and I/O Clock at 250 MHz

Block	Total power portion	Net power portion	# long wires	Remark
SPC	5.8%	55.1%	27.7K	CPU clock, 8X
RTX	3.6%	44.4%	27.5K	I/O clock
CCX	2.8%	57.6%	12.4K	CPU clock
L2D	2.1%	29.2%	6.5K	8X
L2T	1.8%	48.5%	6.0K	8X
RDP	1.7%	48.9%	5.2K	I/O clock
TDS	1.3%	43.1%	4.8K	I/O clock
DMU	1.1%	40.7%	5.4K	I/O clock

TABLE IV
L2T Die Partitioning Schemes

Part #	die bot	die top	# TSV
1	small macros syn' mem, logic	large macros syn' mem, logic	1014
2	small macros, logic	large macros, syn' mem	1950
3	syn' mem, logic	all macros, logic	2451
4	small macros, syn' mem	large macros, logic	4120
5	large macros, logic	small macros, syn' mem	5073

In this paper, we fold five blocks: 1) SPC; 2) CCX; 3) L2D; 4) L2T; and 5) RTX. In the following sections III-B and III-C, we discuss the block folding methodologies for SPC and L2T. Each block shows the distinctive folding characteristics. We tried both manual (based on design information such as connectivity between submodules) and automated (min-cut partitioner with cut size control) partitioning methods for each block folding, and selected a better case.

B. Folding L2T Block

The L2T consists of memory macros, synthesized memory blocks, and control logic cells, and each of them occupies about one third of the total area. Partitioning options that we examined are listed in Table IV. Note that memory macros are divided into two groups based on their connectivities, i.e., tightly connected macros form a group. In partition #1, after splitting memory blocks, logic cells are partitioned using a min-cut partitioner, which leads to the smallest number of TSVs among five cases. On the other hand, in partition #5, where the largest number of TSVs are used, the silicon area occupied by TSVs is as high as 10%, as shown in Fig. 2. All these partitions are determined considering the area balance between dies including the TSV area.

The die partitioning impact on the 3-D design quality is shown in Fig. 3. We observe that partitioning cases with

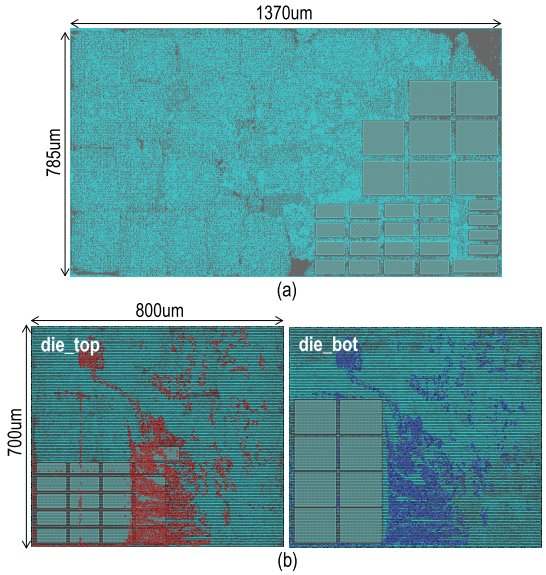


Fig. 2. L2T 2-D and 3-D layouts. (a) 2-D design. (b) 3-D design of partition #5 in Table IV (#TSV: 5073). The total TSV area is 10%.

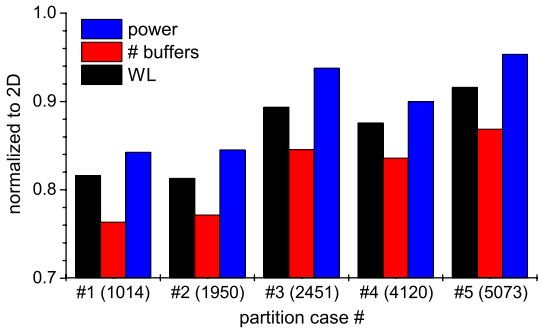


Fig. 3. Impact of L2T die partitioning on wirelength, buffer count, and power. All numbers are normalized to the 2-D.

a larger number of TSVs tend to lose the 3-D power benefit. For example, partition #1 (#TSV: 1014) shows 15.7% power saving, while partition #5 (#TSV: 5073) achieves only 4.7% power reduction compared with the 2-D. In these cases, the large TSV area overhead results in the increase in footprint area, wirelength, buffer usage, and hence power consumption. However, we cannot generalize that more 3-D connections degrade the 3-D design quality as this highly depends on 3-D interconnect elements.

C. Second-Level Folding SPC Block

In case of SPC, we employ our block folding strategy, one step further: we fold functional unit blocks (FUBs) inside a SPC that contains 14 FUBs, including two integer execution units, a floating point and graphics unit, five instruction fetch units, and a load/store unit (LSU). This SPC is the highest power consuming block in T2.

We apply the same block folding criteria discussed in Section III-A, and based on this, six FUBs are folded, as shown in Fig. 4. We call this second-level folding. With this second-level folding, we obtain 9.2% shorter wirelength, 10.8% less

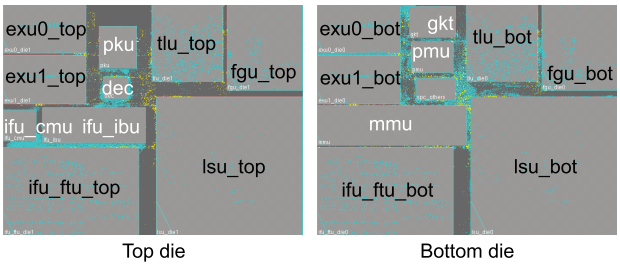


Fig. 4. Second-level folding of an SPC. Six FUBs shown in black text are folded (#F2F via: 10251).

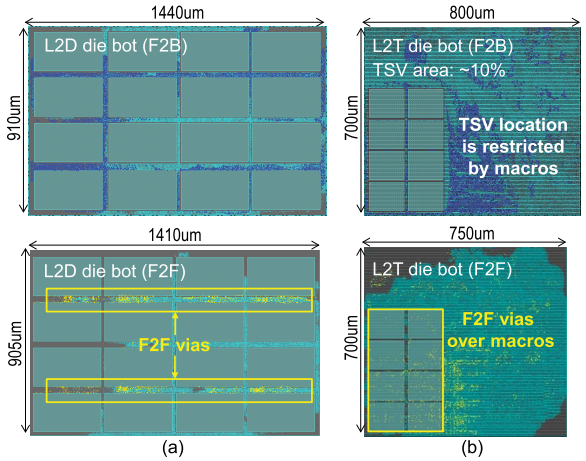


Fig. 5. Bonding style impact on 3-D placement. Blue rectangles: TSV landing pads at M1. Yellow dots: F2F vias. (a) L2D bottom die. (b) L2T bottom die.

buffer counts, and 5.1% reduced power consumption than the SPC without the second-level folding, i.e., a block-level 3-D design of the SPC. Additionally, our 3-D SPC achieves 21.2% power saving over the 2-D SPC.

D. Block Folding in F2F Bonding

So far, we discussed 3-D designs based on F2B bonding using TSVs. In this section, we examine how F2F bonding style utilizing F2F vias for 3-D connections affects the 3-D block folding design quality and power.

F2F vias do not consume silicon area, and hence a 3-D footprint area can be further reduced, as shown in Fig. 5. For example, the folded L2D and L2T with F2F bonding reduce footprint by 2.6% and 6.3%, respectively, compared with F2B bonding cases. In the folded L2D case, as shown in Fig. 5(a), all F2F vias are located on the horizontal channels between memory macros to connect memory I/O pins and logic cells right below them. On the other hand, TSVs are spread out all over the place because of their size and pitch. This affects cell placement as well, and hence degrades wirelength and power. For the same 3-D partition, the folded L2D with F2F bonding shows 11.1% shorter wirelength, 3.9% less buffer count, and 4.1% less power consumption than the F2B case.

In addition, F2F via locations are not restricted by cells and macros. In the folded L2T case, as shown in Fig. 5(b), F2F vias are found over large memory macros. However, TSVs are ousted from memory macro area, which increases wirelength.

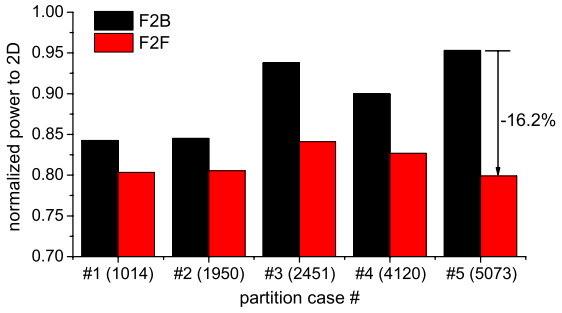


Fig. 6. Bonding style impact on power in L2T folding. Numbers in parentheses are the number of TSVs/F2F vias.

The five partitioning cases for L2T are implemented in both F2B and F2F bonding styles. Power comparisons between both bonding styles are shown in Fig. 6. First, F2F wins over the F2B bonding style in all cases. This is the combined effect of reduced footprint, better 3-D connection points, shorter wirelength, less buffer usage, and better timing. Second, F2F bonding cases show larger power savings over the F2B cases in partition cases with more 3-D connections. In particular, partition #5 that shows the smallest 3-D power benefit in F2B now achieves the best power saving with the F2F bonding. Compared with the F2B case, the F2F case reduces power by 16.2%. In this specific case, the 3-D design quality in the F2B bonding is degraded largely by TSV area overhead, not by the partition. Third, more 3-D connections in the F2F style do not necessarily mean better power saving. Although partitions #3 and #4 show much better power saving than the F2B cases, these power savings are still less than partitions #1 and #2. This emphasizes the importance of die partitioning again.

IV. Full-Chip Assembly With Folded Blocks

So far, we discussed the impacts of block folding along with bonding styles on 3-D power savings. In this section, we integrate all these folded blocks into 3-D T2 full-chip and examine its impact on the system-level power.

A. 3-D Floorplan With Folded Blocks

Based on the block folding criteria in Section III-A, SPC, CCX, L2D, L2T, and RTX are folded. Unlike other four blocks, RTX runs at I/O clock frequency (=250 MHz). In addition, almost all signals to/from RTX are connected with MAC, TDS, and RDP that form a network interface unit (NIU) with RTX. Thus, the impact of RTX folding is limited to the RTX and NIU. In this paper, we implement two 3-D designs with folded blocks: 1) T2 with folded SPCs, CCX, L2Ds, and L2Ts and 2) T2 with all five types of blocks folded.

In each case, we build two designs using either F2B or F2F bonding style. Note that there is a difference in routing layer usage in folded blocks depending on the bonding style. For the F2B bonding, the die bottom of folded blocks uses up to M7 (TSV landing pad at M1) as other unfolded blocks, while the die top utilizes up to M9 (TSV landing pad at M9). Thus, M8 and M9 can be used for over-the-block routing

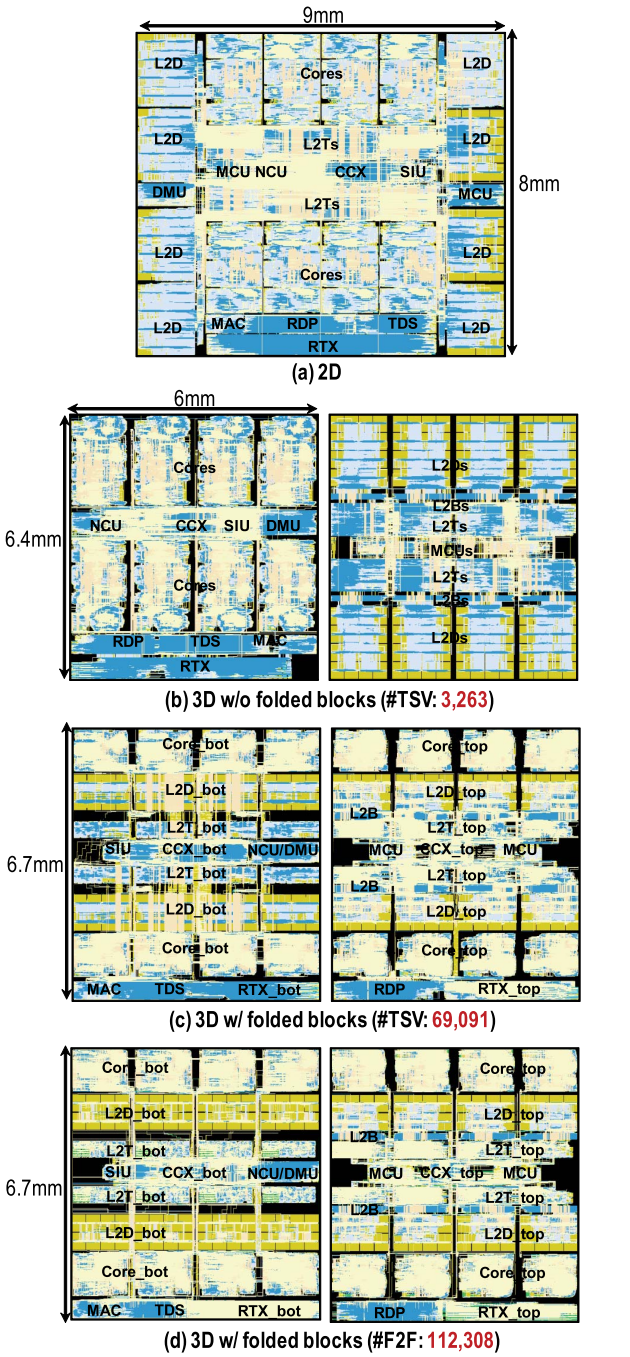


Fig. 7. GDSII layouts of four design styles of OpenSPARC T2 (full-chip). We compare (a) 2-D design (9×7.9 mm^2), (b) core/cache stacking (6×6.4 mm^2, #TSV = 3 263), (c) block folding with TSVs (6×6.6 mm^2, #TSV = 69 091), and (d) block folding with F2F (6×6.6 mm^2, #F2F = 1 12 308).

including folded blocks in the die bottom. The only exception is SPC that uses up to M9 for both dies, as this block requires most routing resources. This is why SPCs are placed on the top and the bottom of the chip, as shown in Fig. 7(c). Otherwise, these SPC blocks will act as interblock routing blockages.

In the F2F bonding case, since F2F via is on the top of M9, all nine metal layers are used for routing in folded blocks. Thus, folded blocks are interblock routing blockages for both

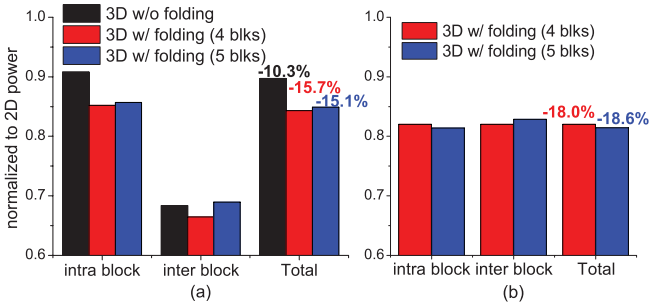


Fig. 8. Block folding impact on 3-D full-chip power. (a) F2B bonding case. (b) F2F bonding case.

dies, as shown in Fig. 7(d). For this reason, although this F2F bonding achieves more power saving than the F2B case in block folding, interblock design quality could be degraded.

In both bonding style cases, we place CCX in the center. There are ∼300 wires between CCX and each SPC (or L2T). Thus, wires between CCX and L2T are much shorter than those between CCX and SPC. All other control units (SIU, NCU, DMU, and MCU) are placed in the center row as well. Finally, NIU blocks are placed in the bottom-most part of the chip, as most of the connections are confined in NIU.

B. Bonding Style Impact: F2B Versus F2F

As discussed in Section III-D, block folding schemes (or die partitioning) are largely affected by bonding styles. For example, in L2T folding, partition #1 (#TSV: 1014) is the best for F2B, while partition #5 (#F2F via: 5073) shows the lowest power for F2F. We choose the best case for each block folding depending on the bonding style, and integrate these folded blocks, as shown in Fig. 7.

3-D T2 full-chip power normalized to 2-D power is shown in Fig. 8. First, we see more 3-D power benefit with block folding (up to 18.6%) compared with the pure block-level 3-D design (10.6%). We also observe most of power saving is from intrablock level (=folded blocks). Note that the interblock power is only ∼5% of the total power.

Second, interblock power savings are worse in the F2F cases. This is because all folded blocks act as interblock routing blockages in the F2F bonding style, which increases interblock wirelength and buffer count. For example, in four types of block folded case, interblock wirelength and buffer count are 19.8 m and 97.6k in F2B, respectively, while 22.5 m and 122.3k in F2F. Third, however, power savings in the intrablock level with F2F bonding overwhelm the loss in the interblock level, and hence F2F cases show a better power benefit than F2B cases.

Last, as shown in Fig. 8(a), folding more blocks does not always lead to more power saving. The RTX folding reduces an interblock power benefit in both bonding styles. Without RTX folding, connections between RTX and other NIU blocks are directly made by TSVs (or F2F vias). Thus, there are not many long horizontal wires. However, with RTX folding, long horizontal wires are unavoidable between RTX and MAC, for example, as shown in Fig. 7(d). In the F2B case, interblock wirelength increases by 9.6% compared with the four block types folded case. This in turn degrades the intrablock design quality, as shown in Fig. 8(a). Therefore, interblock connections and its impact need to be considered when selecting blocks to fold.

TABLE V
COMPARISON BETWEEN 2-D, 3-D WITHOUT BLOCK FOLDING (CORE/CACHE, F2B), AND 3-D WITH BLOCK FOLDING (FIVE TYPES OF BLOCKS FOLDED, F2F) DESIGNS. DVT DESIGN TECHNIQUE IS APPLIED TO ALL CASES. NUMBERS IN PARENTHESES ARE DIFFERENCE AGAINST THE 2-D EXCLUDING HVT CELL COUNT THAT SHOWS % OF TOTAL CELL COUNT

	2D	3D w/o folding	3D w/ folding
footprint (mm^2)	71.1	38.4 (-46.0%)	40.8 (-42.6%)
Wirelength (m)	339.7	321.3 (-5.5%)	309.6 (-8.9%)
# cells ($\times 10^6$)	7.41	7.09 (-4.3%)	6.83 (-7.8%)
# buffers ($\times 10^6$)	2.89	2.37 (-17.9%)	2.23 (-22.8%)
# HVT cells ($\times 10^6$)	6.50 (87.8%)	6.38M (90.0%)	6.42 (94.0%)
# TSV/F2F via	0	3,263	165,044
Total power (W)	**8.240**	**7.113 (-13.7%)**	**6.570 (-20.3%)**
Cell power (W)	1.770	1.394 (-21.2%)	1.175 (-33.6%)
Net power (W)	4.467	3.966 (-11.2%)	3.806 (-14.8%)
Leakage power (W)	2.003	1.753 (-12.4%)	1.589 (-24.2%)

C. Overall Comparison With Dual-Vth Cells

Up to this point, both 2-D and 3-D designs utilize only RVT cells. However, industry has been using multi-Vth cells to further optimize power, especially for leakage power, while satisfying a target performance. We employ high-Vth (HVT) cells to examine their impact on power consumption in 2-D and 3-D designs. Each HVT cell shows ∼30% slower, yet 50% lower leakage, and 5% smaller cell power consumption than the RVT counterpart.

We now compare three full-chip T2 designs: 1) 2-D IC; 2) 3-D IC without folding (core/cache stacking, F2B bonding); and 3) 3-D IC with block folding (five types of blocks folded, F2F bonding), all with a dual-Vth (DVT) cell library. Detailed comparisons are shown in Table V. We first observe a higher HVT cell usage in 3-D designs, especially for the 3-D with folding case (94% of cells are HVT). This is largely due to better timing in 3-D designs, and this helps further reduce power in 3-D ICs. The 2-D DVT design reduces the power by 9.5% and the 3-D with folding by 11.4% compared with the corresponding RVT only design, which again shows the benefit of 3-D designs.

Most importantly, the 3-D with folding case with F2F bonding reduces the total power by 20.3% compared with the 2-D and by 10% compared with the 3-D without folding case. This clearly demonstrates the powerfulness of block folding along with its bonding style in 3-D designs for power reduction.

V. BLOCK FOLDING IMPACT ON THERMAL

Thermal is one of critical issues in 3-D ICs and has been actively researched. Block folding enhances power saving in 3-D ICs by reducing wirelength and buffer count. However, there are few studies on the thermal impact of block folding

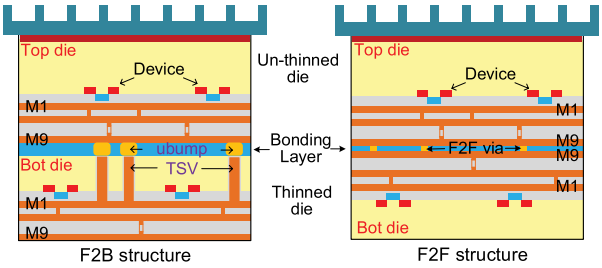


Fig. 9. Thermal structure of F2B and F2F bonding.

and 3-D bonding styles if any. In this section, the thermal impact of bonding styles, i.e., F2B and F2F, is studied in detail.

A. Thermal Analysis Flow

The structures of F2B and F2F are shown in Fig. 9. In the F2B structure, a BCB layer is often used as an adhesive between dies, since it provides a cost-effective solution to form a strong and reliable bonding. However, the thermal conductivity of BCB is very low, and this limits vertical heat flow. On the other hand, F2F bonding uses a direct copper bonding without adhesive. The background material of the bonding layer is SiO_2 that has about five times larger thermal conductivity than BCB. Both improve the thermal conductivity of F2F bonding layer.

The 3-D IC thermal analysis tools, such as 3-D-ICE [8], take a floorplan to compute the thermal conductivity of each layer. This is useful for early stage thermal estimation. To accurately assess the thermal impact, we first build a mesh structure where each layer contains thousands of thermal cells. Then, the layout information, including cells, wires, and TSVs, is extracted from GDSII file, and then the thermal conductivity of each thermal cell is computed based on the material portion inside the cell. A detailed power distribution map is then used 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.

Even though block folding reduces the overall power consumption, this increases the maximum power density, especially when high-power-density modules, such as a core, are folded. For the nonfolded design, this problem can be mitigated by a thermal-aware floorplan, so that the hot spots of each die do not overlap. However, as tiers of a folded block have to be placed at the same XY-location, the maximum power density is still much higher than that of nonfolded designs even with power reduction considered. The power maps for both folded and nonfolded designs are shown in Fig. 10, and the power density in folded case increases by 72% in the core area than the nonfolded case.

B. Thermal Results: Block Folding in F2B Bonding

Thermal analysis results of F2B bonding are summarized in Table VI, and thermal maps of die bottom, farther die from heatsink, and hence hotter die, are shown in Fig. 11. Interestingly, the block folding does not worsen thermal results even with the increased maximum power density. In all cases, the maximum temperature of a folded design is in a similar

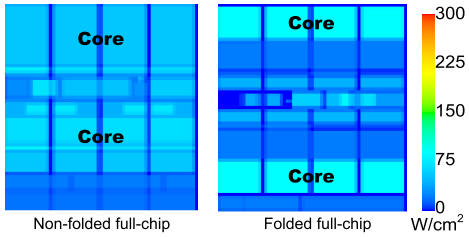


Fig. 10. Power map (gcc) comparison: nonfolded versus folded (F2B).

TABLE VI
IMPACT OF BLOCK FOLDING (F2B) ON THERMAL. 2-D TEMPERATURE RANGE IS INCLUDED FOR COMPARISON

Benchmark	Folded?	Temperature range (°C)		
		2D	die bottom	die top
gcc	no	36.3 - **45.4**	53.6 - **61.7**	52.9 - 57.6
	yes		51.1 - **59.2**	50.6 - 57.3
spice	no	37.0 - **46.2**	54.6 - **63.2**	53.9 - 58.8
	yes		52.1 - **60.7**	51.6 - 58.6

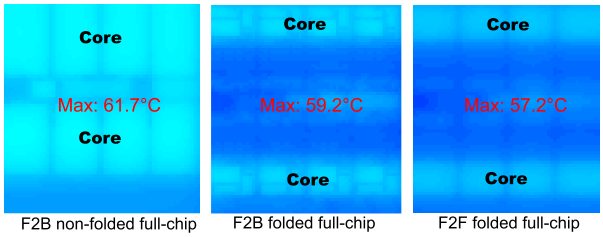


Fig. 11. Full-chip level die bottom temperature map (gcc) comparison.

range of its nonfolded counterpart. First, with block folding, half of the high-power-density blocks are moved to die top, which is closer to heatsink. Second, the large number of TSVs used in block folding increases thermal conductivity. This improves a vertical heat flow and helps heat dissipation.

TSV locations in the nonfolded and folded cases affect the thermal conductivity and hence temperature, as shown in Fig. 12. In the nonfolded design, TSVs are placed outside of blocks, which introduce longer paths between heat sources and TSVs. This results in a higher intradie temperature variation, where the functional blocks are hot spots, while TSV farms are cooler spots. However, in the block folding case, as TSVs are placed inside each block, lateral heat dissipation paths become shorter, which helps cool the block more evenly. Moreover, since any signal TSVs are paired with microbumps, they further improve the thermal benefit, i.e., the thermal conductivity of the bonding layer increases. Finally, the overall power consumption decreases by block folding, and this leads to an average temperature reduction for both dies.

C. Thermal Results: Block Folding in F2F Bonding

A direct copper bonding is used in F2F stacking instead of a BCB layer. This leads to a background thermal conductivity improvement in the bonding layer. In addition, a thinner bonding layer in the F2F structure has less limitation on vertical heat flow than F2B. Since both metal layers and the F2F bonding layer use the same background material, the F2F bonding layer is no longer the bottleneck for vertical heat flow.

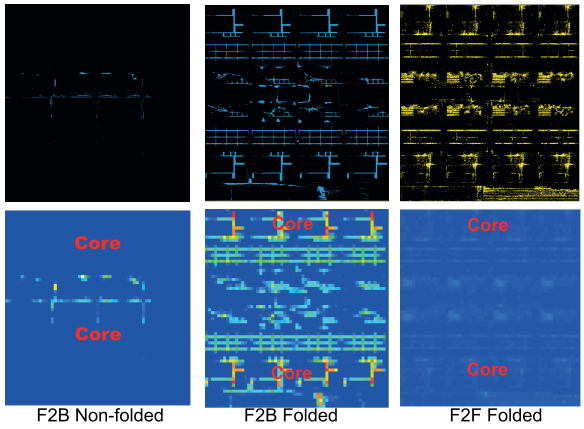


Fig. 12. Full-chip level bonding layer thermal conductivity map comparison. Red region has a thermal conductivity of 80 W/m/K. Top: TSV and F2F via locations are shown as blue and yellow dots.

TABLE VII
IMPACT OF BLOCK FOLDING (F2F) ON THERMAL

Benchmark	Bonding	Temperature range (°C)	
		die bottom	die top
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

F2F vias are much smaller than TSVs, which is good for the 3-D IC design perspective, but not for thermal. F2F vias introduce less copper into the bonding layer than microbumps. As shown in Fig. 12, the F2B designs show better thermal conductivity where TSVs are located than the F2F case.

The thermal analysis results of F2F bonding are summarized in Table VII, and the thermal maps are shown in Fig. 11. First, we observe that whether blocks are folded or not, and F2F bonding cases show a lower maximum temperature than their F2B counterparts, although the difference is not significant. This is because of better vertical heat flow and lower power consumption in F2F bonding.

VI. IMPACT OF POWER DISTRIBUTION NETWORK

So far, both 2-D and 3-D designs are built without PDN. However, PDN occupies a considerable amount of routing resources, and its impact on overall design quality is nongligible. In this section, based on layout simulations, the impact of PDN on 3-D power benefit is discussed.

The details of the PDN are described in Table VIII. PDN is planned in the initial design stage before placement and routing. The PDN width/pitch is chosen considering the alignment with routing tracks, and PDN is not planned on M1 and M2. This is because standard cells already contain VDD/VSS lines on M1, and the PDN on M2 acts as placement blockages.

A. PDN Impact on Full-Chip 3-D IC Quality

In 3-D ICs, the location of TSVs and F2F vias is affected by PDN. Unless backside RDL is used, a TSV must satisfy two constraints. First, TSV location should not overlap with the standard cells/memory macros as well as with M1 wires in die bottom. Second, the landing pad of a TSV in die top should not overlap with the top metal PDN (M9). In case of F2F bonding, a F2F via must be placed in an empty space on both dies where there is no top metal PDN (M9). How PDN affects the F2F via location.

TABLE VIII
PDN SPECIFICATIONS. # TRACKS SHOW THE MAX NUMBER OF SIGNAL WIRES THAT CAN FIT IN BETWEEN TWO ADJACENT P/G WIRES

	Local	Intermediate	Global		
	M3	M4 - M6	M7	M8	M9
width/pitch nm	56/152	112/228	224/456		
PDN density (%)	10.5	14.9	18.0	21.4	24.9
PDN width (nm)	208	340	2048		
PDN pitch (nm)	1,976	2,280	11,400	9,576	8,208
# tracks	11	8	20	16	13

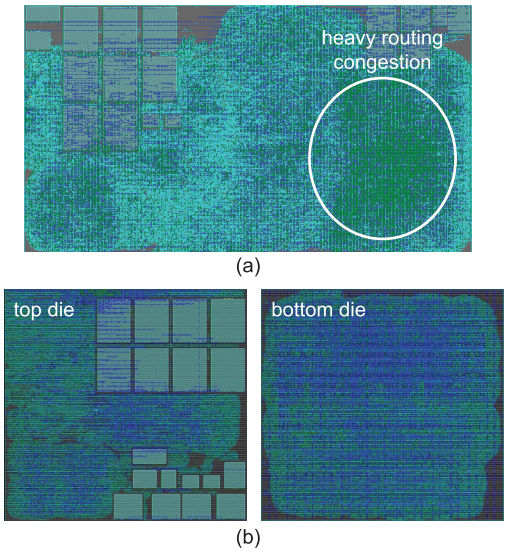


Fig. 13. Congestion map showing the impact of PDN on routing in LSU. Green area: routing demand exceeds the capacity (high congestion). Blue area: routing demand and capacity are same. (a) 2-D LSU, w/PDN, #DRV = 99. (b) 3-D folded LSU, w/PDN, #DRV = 0 for both dies.

In terms of signal net routing, both 2-D and 3-D designs suffer from the reduced routing resources due to the PDN. For example, in a 2-D LSU inside the T2 core, heavy routing congestion is observed because of the PDN [see Fig. 13(a)]. Hence, routing design rule violations (DRVs) increase. Blocks that require more routing resources will suffer more from PDN, which results in higher congestion and DRVs. The same happens in 3-D designs as well. However, shorter wirelength in 3-D IC helps reduce the impact of PDN on the routing congestion as the routing demand also reduces. As shown in Fig. 13(b), the routing congestion problem is much reduced in 3-D, and hence no DRVs.

The entire T2 is redesigned with PDN. Power comparison results with and without PDN are shown in Fig. 14. First of all, the 3-D power saving still holds with PDN. The PDN mostly increases net power largely due to increased wirelength as well as wire capacitance including coupling capacitance to PDN. However, as 3-D requires less routing resources than 2-D, the impact of PDN on 3-D design is slightly less than the 2-D counterpart. That is why 3-D power reduction is also slightly improved with PDN, especially for block folding cases where more wirelength reduction is achieved.

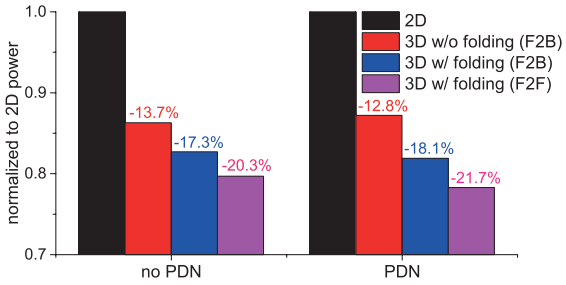


Fig. 14. PDN impact on full-chip power.

VII. Conclusion

In this paper, the power benefit of 3-D ICs was demonstrated with an OpenSPARC T2 chip. To further enhance the 3-D power benefit on top of the conventional 3-D floorplanning method, block folding methodologies and bonding style impact were explored. We demonstrated more 3-D power reduction with F2F bonding than F2B. In addition, the block folding, especially in F2F bonding case, was shown to improve thermal issue slightly compared with nonfolded case. Last, we demonstrated that PDN does not hurt the 3-D power benefit. With aforementioned methods, the total power saving of 21.7% has been achieved against the 2-D counterpart.

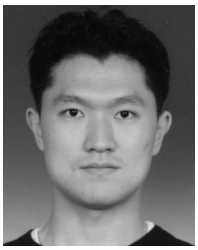
References

[1] B. Black *et al.*, "Die stacking (3D) microarchitecture," in *Proc. 39th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Dec. 2006, pp. 469–479.
[2] U. Kang *et al.*, "8 Gb 3-D DDR3 DRAM using through-silicon-via technology," *IEEE J. Solid-State Circuits*, vol. 45, no. 1, pp. 111–119, Jan. 2010.
[3] Oracle. *OpenSPARC T2*. [Online]. Available: http://www.oracle.com, accessed 1, 2013.
[4] Synopsys. *32/28 nm Generic Library*. [Online]. Available: http://www.synopsys.com, accessed 1, 2013.
[5] U. G. Nawathe *et al.*, "An 8-core 64-thread 64b power-efficient SPARC SoC," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2007, pp. 108–590.
[6] G. Katti, M. Stucchi, K. De Meyer, and W. Dehaene, "Electrical modeling and characterization of through silicon via for three-dimensional ICs," *IEEE Trans. Electron Devices*, vol. 57, no. 1, pp. 256–262, Jan. 2010.
[7] D. H. Kim, R. O. Topaloglu, and S. K. Lim, "Block-level 3D IC design with through-silicon-via planning," in *Proc. Asia South Pacific Design Autom. Conf.*, Jan./Feb. 2012, pp. 335–340.
[8] A. Sridhar, A. Vincenzi, D. Atienza, and T. Brunschwiler, "3D-ICE: A compact thermal model for early-stage design of liquid-cooled ICs," *IEEE Trans. Comput.*, vol. 63, no. 10, pp. 2576–2589, Sep. 2014.

Moongon Jung (S'11–M'15) received the B.S. degree in electrical engineering from Seoul National University, Seoul, Korea, in 2003, the M.S. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 2009, and the Ph.D. degree from the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2014.

He joined Intel Labs, Intel Corporation, Santa Clara, CA, USA, as a Research Scientist, in 2014. He is involved in design methodologies for future technologies. His current research interests include computer-aided design for very large scale integration circuits, in particular, on physical design methods for low power 3-D ICs and thermomechanical reliability analysis and optimization of TSV-based 3-D ICs.

Dr. Jung was a nominee for the best paper awards at DAC in 2011 and 2012, and the IEEE Transactions on Computer-Aided Design in 2013. His research on thermomechanical reliability of 3-D ICs was featured as a Research Highlight in the Communication of the ACM in 2014.



Taigon Song (S'09) received the B.S. degree in electrical engineering from Yonsei University, Seoul, Korea, in 2007, and the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 2009. He is currently pursuing the Ph.D. degree with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

He was an Electromagnetic Interference Engineer with the Online Electric Vehicle Business Department in 2010. His current research interests include low power design methodologies for 3-D ICs, silicon interposer design and co-analysis, TSV-to-TSV/face-to-face coupling in 3-D ICs, chip-package-PCB co-analysis on power integrity, and thermal analysis of 3-D ICs with integrated voltage regulators.

Yarui Peng (S'12) received the B.S. degree from Tsinghua University, Beijing, China, in 2012, and the M.S. degree from the Georgia Institute of Technology, Atlanta, GA, USA, in 2014, where he is currently pursuing the Ph.D. degree with the School of Electrical and Computer Engineering.

His current research interests include physical design and analysis for 3-D ICs, including parasitic extraction and optimization for signal integrity, thermal, and power delivery issues.

Mr. Peng was a recipient of the Best in Session Award in SRC TECHCON in 2014.

Sung Kyu Lim (S'94–M'00–SM'05) received the B.S., M.S., and Ph.D. degrees from the Computer Science Department, University of California at Los Angeles, Los Angeles, CA, USA, in 1994, 1997, and 2000, respectively.

He joined the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2001, where he is currently a Professor. He led the Cross-Center Theme on 3-D Integration for the Focus Center Research Program with Semiconductor Research Corporation, Durham, NC, USA, from 2010 to 2012. He has authored the books entitled *Practical Problems in VLSI Physical Design Automation* (Springer, 2008), and the *Design for High Performance, Low Power, and Reliable 3-D Integrated Circuits* (Springer, 2013). His current research interests include the architecture, design, test, and EDA solutions for 3-D ICs.

Dr. Lim was a recipient of the National Science Foundation Faculty Early Career Development Award in 2006. He was also a recipient of the best paper award from SRC TECHCON'11, TECHCON'12, and ATS'12. His work was also nominated for the best paper award at ISPD'06, ICCAD'09, CICC'10, DAC'11, DAC'12, ISLPED'12, ISPD'14, and DAC'14. He was on the Advisory Board of the ACM Special Interest Group on Design Automation (SIGDA) from 2003 to 2008, and also a recipient of the ACM SIGDA Distinguished Service Award in 2008. He was an Associate Editor of the IEEE Transactions on Very Large Scale Integration Systems from 2007 to 2009, and has been an Associate Editor of the IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems since 2013. His research was featured as a Research Highlight in the Communication of the ACM in 2014.