Design Methodologies for Low-Power 3-D ICs With Advanced Tier Partitioning

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 considered as the driving force for 3-D ICs, yet there have been few thorough design studies on how to reduce power in 3-D ICs. In this paper, we discuss computer-aided design techniques and design methodologies to reduce power consumption in 3-D IC designs using a commercial grade CPU core (OpenSPARC T2 core). To demonstrate power benefits in 3-D ICs, four design techniques are explored: 1) 3-D floorplanning; 2) metal layer usage control for intrablock-level routing; 3) dual-Vth design; and 4) functional unit block (FUB) folding. The benefits and challenges of multiple FUB folding are also discussed. Finally, the through-silicon via technology scaling impact on FUB folding and 3-D power benefit is examined. With the aforementioned methods combined, our 2-tier 3-D designs provide up to 52.3% reduced footprint, 27.9% shorter wirelength, 35.4% decreased buffer cell count, and 27.8% power reduction over the 2-D counterpart under the same performance.**

***Index Terms*— 3-D integrated circuits, block folding, low power, physical design methods.**

I. Introduction

POWER reduction has been one of the most critical design considerations for IC designers. Minimizing both dynamic and leakage power is imperative to meet power budgets for portable devices (low power applications) as well as server farms (high power applications). The power efficiency also directly affects ICs 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, 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.

There are quite a few works on analytical placement engines for TSV-based 3-D ICs. Luo *et al.* [3] proposed both local and global smoothing techniques for the 3-D area density functions. This placer reduced the wirelength more than 20% compared with prior works. Hsu *et al.* [4] presented a global and detailed 3-D placement based on a weighted-average wirelength model. This work demonstrated 10% wirelength reduction along with 21% TSV count reduction compared with a force-directed 3-D placer [5]. However, all these works lack detailed wirelength and power comparisons with 2-D ICs.

In this paper, we present four physical design techniques that are shown to significantly reduce power consumption in 3-D ICs. Our study is based on the OpenSPARC T2 core design database [6] and PDK that are available to the academic community. This T2 core contains 13 distinctive blocks that have different design characteristics. We present 3-D design impacts on the entire T2 core as well as representative blocks. We build GDSII-level 2-D and 2-tier 3-D layouts and analyze and optimize designs using the sign-off CAD tools.

First, CAD tools developed for this work is presented. Then, we discuss how to rearrange functional unit blocks (FUBs) into 3-D to reduce power. Next, we study how the number of intrablock-level routing layer used affects routing congestion and power consumption in 2-D and 3-D designs differently. We also examine the impact of dual-Vth (DVT) design technique on 2-D and 3-D power consumptions. Then, we demonstrate the effectiveness of FUB folding, i.e., partitioning an FUB into two sub-FUBs and stacking them, in achieving power savings in the 3-D design. In addition, we further examine the benefits and challenges of multiple FUB foldings on 3-D design and its power consumption. Then, we discuss the impact of TSV technology scaling on FUB folding and 3-D power reduction.

II. New 3-D IC CAD Tool Development

In this section, new physical design tools developed in our study, mixed-size 3-D placer and 3-D clock tree router, are presented.

A. Mixed-Size 3-D Placer

A TSV-based 3-D placer based on a system of supply/demand of placement space was presented in [5], but it lacks the capability to handle hard macros such as SRAMs and register files. This capability can be added by treating a hard macro as a large cell, which demands some placement space. However, as observed in a similar 2-D placer [7], this leads

Manuscript received August 19, 2016; revised December 19, 2016; accepted January 24, 2017. Date of publication March 2, 2017; date of current version June 23, 2017. This work was supported by Intel Corporation through Semiconductor Research Corporation under Grant ICSS Task 2293.

M. Jung is with Intel Labs, Intel, Santa Clara, CA 95054 USA (e-mail: moongon.jung@intel.com).

T. Song, Y. Peng, and S. K. Lim are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: 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/TVLSI.2017.2670508

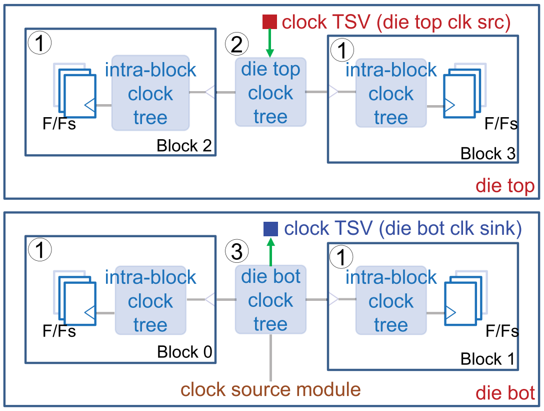


Fig. 1. 3-D CTS flow. 1—Intrablock CTS. 2—Die top CTS. Clock TSVs are clock sources of die top. 3—Die bottom CTS. Die top clock tree information is utilized during die bottom CTS through clock TSVs.

TABLE I

CLOCK TREE COMPARISON: 2-D VERSUS 3-D (WITHOUT FOLDING)

	# clk buffers	max skew (ps)	clock power (mW)
2D	10.9K	182.7	31.5
3D	9.2K	174.8	25.8

to large whitespace regions in the vicinity of the hard macros called halos. Spindler *et al.* [7] solved this issue by reducing the demand of the hard macros.

However, we observe that this tactic is insufficient for extremely large hard macros such as memory macros in load/store unit (LSU), for which halos still exist. Instead, in this paper, we set both the supply and the demand of the regions the hard macros occupy to zero. This is essentially a hole in the supply/demand map and it works well for hard macros of all sizes.

B. 3-D Clock Tree Router

Our 3-D CTS flow is briefly described in Fig. 1. The goal is to use a commercial 2-D clock router to build 3-D clock trees while minimizing clock skew.

We first build an intrablock level clock tree for each block with the given clock skew and slew constraints. Then, we construct clock trees in die top. Clock TSV pads at M9 are clock roots in this case. In this step, the basic clock tree information of each block, such as the minimum and maximum clock latency, is transferred to the clock root pin of each block. This information is then utilized for skew control between blocks during the die top CTS. Finally, we perform die bottom CTS where the clock source block is located. In this case, clock TSV pads at M1 are now clock sinks and the clock tree information of die top is passed on these pads. With this 3-D specific clock tree information annotated on clock TSV pads, we can fully utilize existing commercial 2-D CTS and routing engines. During the die bottom CTS, the clock tree data of die top and blocks in die bottom are taken into account for both 2-D and 3-D clock skew control.

The 2-D and 3-D CTS results are summarized in Table I. The maximum clock skew is close to 12% of clock period for both cases. We also observe that our 3-D design uses 15.6% less clock buffer and hence 18.0% less clock power than the 2-D design. The distance between the clock source block and clock pin of each block decreases in 3-D, which in turn reduces the clock buffer count. Although TSV *RC* adds additional latency on clock tree, since only one clock TSV used per clock domain in our implementation, this impact on latency, skew, and clock buffer count is negligible. In addition, this clock source to each block's clock pin distance varies more in 2-D than 3-D due to the larger footprint area. This forces 2-D design to use stronger clock buffers to balance clock skew across blocks.

III. 3-D FLOORPLANNING BENEFITS

In this section, we explain how we implement both 2-D and 3-D block-level designs in detail. Then, based on our layout simulations, we compare several critical design metrics such as footprint area, wirelength, and power consumption of 3-D designs with the traditional 2-D designs under the same performance, i.e., isoperformance comparison.

A. 2-D Design

The OpenSPARC T2 core consists of 13 FUBs including two integer execution units (EXUs), a floating point and graphics unit (FGU), five instruction fetch units (IFUs), and an LSU [6]. Each FUB is synthesized with a 28-nm cell library. In our implementation, top-level logic cells, i.e., cells outside FUBs, are grouped during synthesis to form an additional block. Thus, a total of 14 FUBs are floorplanned, and special care is taken to use both connectivity and data flow between FUBs to minimize interblock wirelength.

With a given target timing constraint, cells and memory macros are placed in each FUB. Note that we only utilize regular-Vth (RVT) cells as a baseline. Then, we perform a static timing analysis (STA) on the placed 2-D T2 core and obtain a new timing constraint for I/O pins of each FUB. With this new timing constraint, we perform FUB-level and core-level timing optimizations (buffer insertion and gate sizing) as well as power optimizations (gate sizing). We improve the design quality through each design step: placement, preroute optimization, CTS, post-CTS optimization, routing, and postroute optimization. The 2-D placement result is shown in Fig. 2(a). Note that intrablock and interblock routing utilize up to M5 and M9, respectively. Thus, top four metal layers can be used for over-the-block wiring.

B. 3-D Design

The T2 core netlist is partitioned into two dies considering the area balance between dies and connectivity between FUBs. Then, the 3-D floorplanner in [8] is employed with an objective of minimizing interblock wirelength. In addition, two dies are assumed to be bonded in a face-to-back style. Note that TSV arrays are treated as additional blocks in this flow, and hence all TSVs can be placed outside FUBs only. The TSV diameter, height, resistance, capacitance, and landing pad size are 3 μm, 18 μm, 43 mΩ, 8.35 fF, and 3.3×3.3 μm^2, respectively, unless otherwise specified. The total number of TSVs is 2979 in this design and TSVs occupy only 1.1% of the total silicon area. The 3-D placement result is shown in Fig. 2(b).

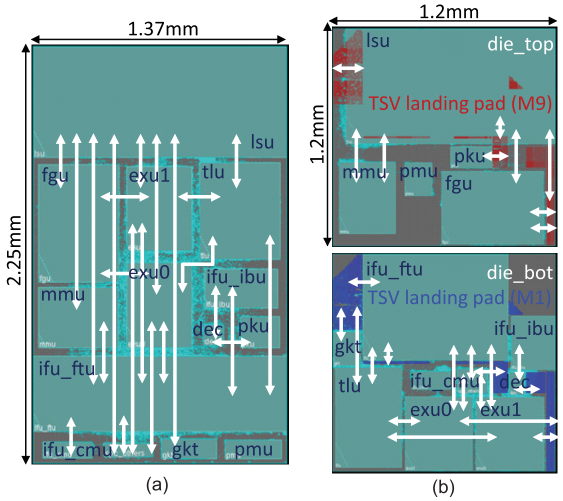


Fig. 2. 2-D and 3-D placement results. (a) 2-D design. (b) 3-D design with 2979 TSVs. The cyan dots are core-level buffers. The blue and red rectangles are TSV landing pads at M1 and M9, respectively. The white arrows represent major interblock connections.

TABLE II

COMPARISON BETWEEN 2 D AND 3 D DESIGNS WITH A TARGET CLOCK PERIOD OF 1.5 ns. NUMBERS IN PARENTHESES ARE (INTRABLOCK/INTERBLOCK) BREAKDOWN

	2D	3D	diff
footprint (mm^2)	3.08	1.47	-52.3%
utilization (%)	67.8	66.8	-1.0%
# cells (×1000)	504.8 (483.8/21.0)	481.0 (458.8/22.2)	-4.7%
# buffers (×1000)	209.5 (188.5/21.0)	186.0 (163.8/22.2)	-11.2%
Wirelength (m)	23.3 (18.6/4.7)	20.2 (17.6/2.6)	-13.3%
Total power (mW)	**408.6 (358.9/49.7)**	**350.5 (320.6/29.9)**	**-14.2%**
Cell power (mW)	79.5 (73.4/6.1)	67.9 (62.2/5.7)	-14.6%
Net power (mW)	181.8 (150.9/30.9)	154.5 (137.3/17.2)	-15.0%
Leakage power (mW)	147.3 (134.6/12.7)	128.1 (121.1/7.0)	-13.0%

Our RTL-to-GDSII tool chain 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. The netlists and the extracted parasitic files are used for 3-D STA, followed by the timing and power optimization with the timing constraints from the 3-D timing results [9].

C. 2-D Versus 3-D Floorplanning

We now compare our 2-D and 3-D designs with a target clock period of 1.5 ns (=667 MHz), as shown in Table II. Note that our designs run much slower than UltraSPARC T2, a commercial product version of OpenSPARC T2, which runs at 1.4 GHz [10]. This is mainly because some custom memory blocks in T2 core such as a content-addressable memory are synthesized with cells, since a general memory compiler cannot afford this kind of memories. Unfortunately, these synthesized memories are much larger and run slower than the memory macros generated by a memory compiler. Also, note that we fully redesigned 2-D blocks used in the 3-D design based on floorplan and I/O timing constraint updates.

First, interestingly, the footprint area reduction in the 3-D design is more than 50%. This is largely related to the buffer count reduction in the 3-D design because of shorter wirelength and hence better timing. Note that the silicon area utilization, i.e., area occupied by cells, memory macros, and TSVs (3-D only), for 2-D and 3-D designs is 67.8% and 66.8%, respectively, which supports a fair comparison.

TABLE III

CELL SIZE USAGE (%) COMPARISON BETWEEN 2-D AND 3-D DESIGNS. X0 IS THE SMALLEST CELL SIZE

	X0	X1	X2	X4	X8	X16	X32
2D (%)	11.3	44.4	25.9	6.7	8.2	1.8	1.7
3D (%)	13.2	51.8	21.3	6.1	6.3	0.8	0.5

Second, we observe 11.2% total buffer count reduction and 13.3% total wirelength decrease in the 3-D design. However, counterintuitively, interblock level buffers (=22.2k) in the 3-D design are more than the 2-D (=21.0k) even with the much shorter interblock wirelength. As we optimize the design iteratively in FUB level and core level, buffers can be inserted either inside or outside FUBs to optimize paths. Additionally, to drive 3-D nets with a large TSV capacitance, buffers need to be inserted. Thus, although interblock level buffers are deployed more in the 3-D design, we save a significant number of buffers in the intrablock level. In addition, we see 5.4% intrablock wirelength reduction in the 3-D design mainly because of the intrablock level buffer counter reduction.

Third, most importantly, the 3-D design reduces power consumption over the 2-D counterpart by 14.2%. We see that cell (14.6%) and leakage (13.0%) power reduction are far more than the cell count decrease (4.7%) in the 3-D design. As shown in Table III, the 3-D design utilizes less larger cells than the 2-D case thanks to better timing, i.e., more positive timing slack in paths. With the positive slack, we can downsize cells in the 3-D design if this change still satisfies 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 input pin capacitance of the loading side, and hence the net power is defined as the sum of wire and pin power. Thus, the wire power reduction is directly from shorter wirelength, and the pin power decrease is from the smaller cell size as well as the reduced cell count.

IV. JUDICIOUS METAL LAYER USAGE

A. Different Routing Resource Demand of 2-D and 3-D

So far, each FUB is routed using five metal layers to reserve sufficient routing resources for interblock level routing that utilizes all nine metal layers. In this setting, four high metal layers can be used for over-the-block interconnections.

However, as shown in Fig. 3, the interblock routing demand is quite different between 2-D and 3-D designs. In the 2-D case, a large number of over-the-block wires are required, and this increases both total and average wirelength. Thus, more high metal layers (or global metal layers) are necessary to complete interblock routing. On the other hand, many wires in the 3-D design are connected to nearby TSVs, and this reduces over-the-block wiring demand significantly.

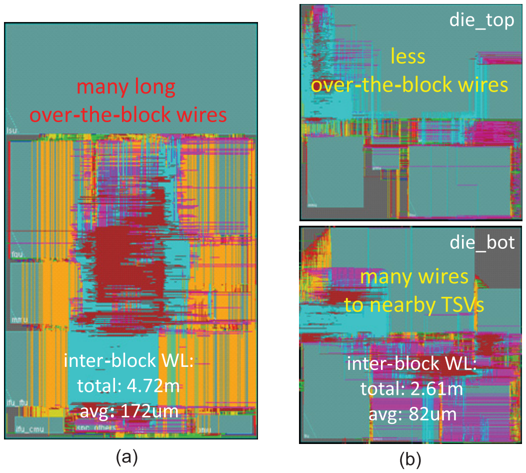


Fig. 3. 2-D and 3-D interblock routing results. Intrablock routing uses up to M5 and interblock routing uses up to M9. (a) 2-D design. (b) 3-D design.

TABLE IV

IMPACT OF METAL LAYER USAGE IN INTRABLOCK LEVEL ROUTING ON POWER CONSUMPTION FOR 2-D AND 3-D DESIGNS. POWER IS NORMALIZED TO THE CASE OF INTRABLOCK ROUTING UP TO M5

	intra-blk M5	intra-blk M6	intra-blk M7
2D power	1.0	1.026	1.021
3D power	1.0	0.971	0.960

Additionally, interblock distance within a die is reduced with the reduced footprint area. As a result, the 3-D design achieves a huge reduction in both total (44.7%) and average (52.3%) interblock wirelengths over the 2-D design. Therefore, this 3-D design may not need four high metal layers for over-the-block wiring.

B. Impact of Intrablock Metal Layer Usage on Power

Next, we investigate whether we can further save power in the 3-D design by allowing more metal layers for intrablock level routing. The key idea here is to reduce the amount of coupling capacitance inside FUBs by relaxing routing congestions with more metal layers and hence to reduce net power consumption. Three cases are studied in this paper: 1) intrablock routing up to M5 (baseline); 2) M6; and 3) M7.

The total power consumption of these three cases is shown in Table IV. All power numbers are normalized to the baseline 2-D and 3-D. As more metal layers are available for intrablock routing (less high metal layers for over-the-block wiring in interblock routing), the 3-D design further reduces power. For example, in the case of intrablock routing up to M7, the total wirelength and wire capacitance reduce by 1.4% and 4.5%, respectively, compared with the baseline. Note that the wire capacitance reduction is much more than the wirelength decrease, which indicates less routing congestion inside FUBs. This results in 6.8% net power and 4.0% total power saving.

However, in the 2-D case, the opposite trend is observed largely because of the increase in both interblock wirelength and buffer count. Moreover, the 2-D design with intrablock routing up to M7 does not even close the target timing, and thus the power number is not reliable.

TABLE V

IMPACT OF INTRABLOCK METAL LAYER USAGE ON INTRABLOCK AND INTERBLOCK DESIGN METRICS IN THE 3-D DESIGN. THE TARGET CLOCK PERIOD IS 1.5 ns AND NUMBERS IN PARENTHESES ARE THE DIFFERENCE WITH RESPECT TO THE CASE OF INTRABLOCK ROUTING UP TO M5

		intra-blk M5	intra-blk M6	intra-blk M7
Wirelength (m)	intra block	17.6	17.3 (-1.5%)	17.1 (-3.0%)
	inter block	2.6	2.7 (+2.7%)	2.8 (+8.8%)
	total	20.2	20.0 (-1.0%)	19.9 (-1.4%)
# buffers (×1000)	intra block	163.8	149.5 (-8.7%)	145.2 (-11.4%)
	inter block	22.2	22.9 (+3.2%)	25.9 (+16.7%)
	total	186.0	172.4 (-7.3%)	171.0 (-8.1%)
Power (mW)	intra block	320.6	309.2 (-3.6%)	303.7 (-5.3%)
	inter block	29.9	31.1 (+4.0%)	32.7 (+9.4%)
	total	**350.5**	**340.3 (-2.9%)**	**336.4 (-4.0%)**

The impact of intrablock metal layer usage on intrablock and interblock design metrics of the 3-D design is shown in Table V. We see that the 3-D design with more intrablock metal layers achieves power reduction by improved intrablock level wirelength and buffer count that overwhelm the degraded interblock level metrics.

V. DUAL-VTH BENEFITS FOR 3-D ICS

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. In this section, we employ high-Vth (HVT) cells to examine their impact on power consumption in 2-D and 3-D designs. Each HVT cell shows around 30% slower, yet 50% lower leakage and 5% smaller cell power consumption than the RVT counterpart.

A. HVT Cell Usage

To examine the 3-D power benefit under different performances, we implement five designs for both 2-D and 3-D cases: target clock periods are 1.5, 1.8, 2, 2.5, and 3 ns. In all cases, we used a DVT cell library. As shown in Fig. 4, 3-D designs always use more HVT cells than 2-D, and the HVT cell usage increases as the target period decreases. Even in the fastest case (1.5 ns), the HVT cell usage in the 3-D design is 91.2%, while that in the 2-D design is only 69.6%. Thus, better timing in 3-D designs translates to higher HVT cell usage and this further reduces leakage power.

B. Power Benefit in 3-D With DVT Design

As shown in Fig. 4, with a DVT design method, 3-D designs benefit more in power reduction for faster cases. This is directly related to the HVT cell usage. At a 1.5-ns clock period, the 3-D design reduces power consumption by 22.5%. As target clock period becomes slower, 2-D designs also heavily utilize HVT cells and reduce the total power consumption notably, which decreases the 3-D power benefit. Still, the DVT design method provides higher power improvement to 3-D designs than RVT-only cases for all target performances.

We observe that the DVT design technique reduces power notably for both 2-D (8.6%) and 3-D (14.0%) designs compared with the RVT only design at a 1.5-ns clock period. However, in the 2-D case, the power saving is solely from

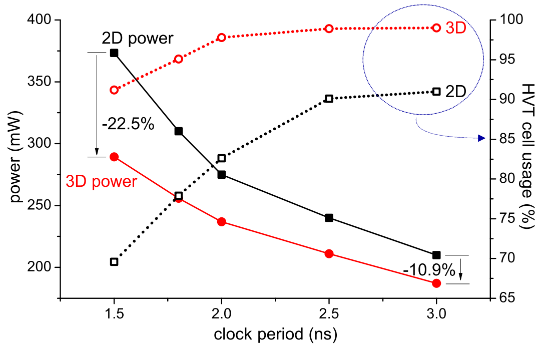


Fig. 4. Power versus delay curves and HVT cell usage for 2-D (intrablock routing up to M5) and 3-D (intrablock routing up to M7) DVT designs.

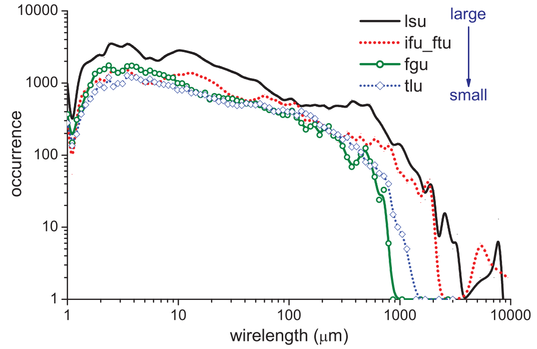


Fig. 5. Wirelength distributions of top four largest FUBs in T2 core.

leakage power reduction (29.9%). By employing weak HVT cells, the 2-D design uses 7.6% more buffers and 6% longer wirelength than the RVT counterpart, which worsens the cell and net power by 0.9% and 4.5%, respectively.

On the other hand, although the 3-D DVT design uses slightly more buffers (1.5%) and longer wirelength (0.5%) than the 3-D RVT design, the cell power decreases by 0.9% since the HVT cell power is slightly lower than the RVT cell and net power remains similar. Most importantly, the leakage power decreases by 35.1%. Thus, the 3-D design benefits more from the DVT design, especially for faster cases.

VI. Folding Functional Unit Block

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

A. Which Block to Fold?

For the FUB folding to provide power saving, certain criteria need to be met. First, the target FUB needs to contain a large number of long wires so that wirelength decrease and hence net power reduction in the folded FUB can be non-negligible. In general, large blocks tend to contain many long wires. Wirelength distributions of top four largest FUBs in the T2 core are shown in Fig. 5. We observe that top two largest FUBs, the LSU and IFU_FTU, are outstanding.

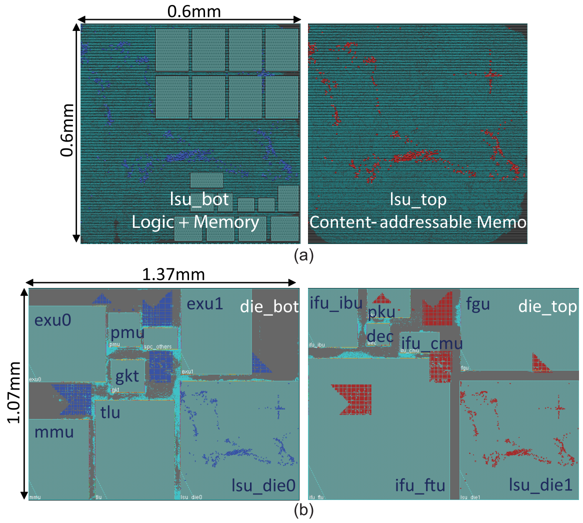


Fig. 6. Placement results of a folded FUB and a 3-D block-level design with the folded FUB. (a) Folded LSU block (#TSV: 596). (b) 3-D design with the folded LSU [#TSV: 2411 (1815 + 596)].

Second, the target FUB is required to consume high enough portion of the total system power. Otherwise, the power saving from the FUB folding could be negligible in the system level. In our implementations, the LSU and IFU_FTU consume around 28% and 23% of the total T2 core power, respectively. Third, the net power portion of the target FUB needs to be high. If the FUB is cell and leakage power dominant, the wirelength reduction of the folded FUB may not reduce the total power notably. The net power portions of the LSU, FGU, and TLU are about 33%, 47%, and 43%, respectively, while that of IFU_FTU is only 17%. Therefore, in this T2 core case, the LSU is the best choice for folding.

B. FUB Folding Impact on Power

The LSU block is partitioned into two dies and designed with an in-house mixed-size 3-D placer, as shown in Fig. 6(a). The DVT design technique is also applied. This folded LSU block reduces the footprint, buffer count, and wirelength by 50.8%, 9.7%, and 7.1%, respectively, compared with the 2-D LSU block. In addition, the HVT cell usage in the folded LSU is 96.8%, while that in the 2-D LSU is 79.7%. More importantly, the total power of LSU is reduced by 5.4% largely due to the decreased net (9.2%) and leakage (4.9%) power.

Detailed comparisons among 2-D, 3-D without FUB folding (*3-D w/o folding*), and 3-D with FUB folding (*3-D w/ folding*) designs are shown in Table VI. In *3-D w/ folding*, the total power reduces by 25.4% compared with the 2-D design and by 3.7% compared with *3-D w/o folding*.

Interestingly, FUB folding helps both the folded block and the overall floorplan for power saving. The interblock wirelength decreases significantly because of better 3-D floorplanning with smaller FUBs, i.e., the largest FUB is divided into two. In this design, the interblock wirelength decreases by 27.6%, which in turn reduces interblock buffers by 6.9% compared with *3-D w/o folding*. As a result, the interblock power reduces by 23.6% compared with *3-D w/o folding*.

TABLE VI

COMPARISON AMONG 2-D (INTRABLOCK ROUTING UP TO M5), *3-D w/o folding* (INTRABLOCK ROUTING UP TO M7), AND *3-D w/ folding* DESIGNS WITH A TARGET CLOCK PERIOD OF 1.5 ns. THE DVT DESIGN TECHNIQUE IS APPLIED TO ALL CASES. NUMBERS IN PARENTHESES ARE THE DIFFERENCE WITH RESPECT TO THE 2-D DESIGN

	2D	3D w/o folding	3D w/ folding
footprint (mm^2)	3.08	1.47 (-52.3%)	1.47 (-52.3%)
utilization (%)	69.2	67.5 (-1.7%)	67.1 (-2.1%)
# cells (×1000)	532.3	471.9 (-11.3%)	450.9 (-15.3%)
# buffers (×1000)	225.5	173.6 (-23.0%)	157.5 (-30.2%)
# HVT cells (×1000)	370.4	430.4 (+16.2%)	444.8 (+20.1%)
Wirelength (m)	24.7	20.0 (-19.0%)	18.4 (-25.5%)
Total power (mW)	**373.3**	**289.3 (-22.5%)**	**278.6 (-25.4%)**
Cell power (mW)	80.2	66.0 (-17.7%)	65.6 (-18.2%)
Net power (mW)	189.9	142.7 (-24.9%)	136.3 (-28.2%)
Leakage power (mW)	103.2	80.6 (-21.9%)	76.7 (-25.7%)

VII. FOLDING MULTIPLE FUBS

In Section VI, we observed that one FUB folding (= LSU) improves the overall design metrics and hence the 3-D power consumption. In this section, we further examine benefits and challenges of folding multiple FUBs on the system-level design. For this experiment, the IFU_FTU, TLU, FGU, and EXU are folded, and these are in top five FUBs in terms of area and power consumption in the T2 core.

A. Partition Matters for FUB Folding

The IFU_FTU, TLU, FGU, and EXU are partitioned into two dies. Traditional min-cut partitioner is first used, while memory macros are manually assigned to two dies. However, as the partitioner does not consider the TSV area overhead in die bottom, the partitioning result often leads to unbalanced die area. Thus, several cut size targets and area ratio areas are tried during partitioning to balance area between dies.

Manual partition is also tried using the connectivity information. For example, TLU block is largely divided into two parts: control units and remaining logic and memory macros. The control units occupy 35% of the block area and are tightly connected. Thus, it is natural to assign the entire control units in one die (= control units are assigned to die bottom along with TSVs in our case). Fig. 7 shows that the manual partition provides a slightly better wirelength, buffer count, and power consumption than the min-cut partitioner based folding. This indicates that there is still room for improvement in FUB folding quality with a better partitioning scheme.

The comparison between 2-D and 3-D folded FUBs are listed in Table VII. By 3-D folding, the power consumption of the TLU, FGU, and EXU reduces by 7.3%, 2.8%, and 3.9%, respectively. Although the IFU_FTU is the second largest in size and power consumption, the folded IFU_FTU consumes a little more power than the 2-D counterpart. This IFU_FTU block is dominated by memory macros and net power portion is low (17%), and hence there are few 3-D partitioning options. In addition, the TSV area overhead is non-negligible (10.9%), and this in turn worsens wirelength and power. This clearly shows the importance of target folding block selection.

B. 3-D Core Design With Multiple Folded FUBs

The folded FUBs are integrated in 3-D core-level designs. Two more 3-D T2 core cases are designed: *two FUBs folded* (LSU and TLU, two best power saving FUBs by folding) and *five FUBs folded* (LSU, TLU, FGU, and two EXUs). The 3-D placement results are shown in Fig. 8. As folded FUBs should be aligned in both dies, 3-D floorplanning options are reducing as more FUBs are folded.

Note that folded FUBs in die bottom uses up to M7, while that in die top utilizes up to M9 to connect to TSV landing pads at M9. Folded FUBs in die top become routing blockages, i.e., over-the-block wiring is impossible over the folded FUBs in die top. This can deteriorate the interblock routing quality especially in case many FUBs are folded. Therefore, folded FUBs are placed along the die boundary, as shown in Fig. 8, so that interblock routing can be mostly done in the middle of the die and hence be less affected by the routing blockage effect of folded FUBs.

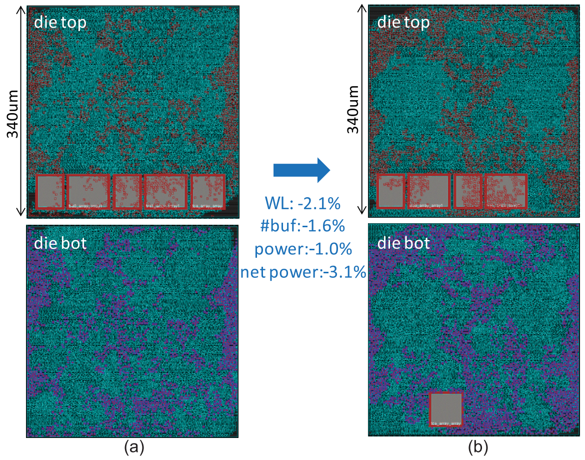


Fig. 7. Impact of FUB (= TLU) partition on design quality. (a) Using min-cut partitioner (#TSV: 2016). (b) Manual partition (die bot: control units and die top: remaining cells, #TSV: 2633).

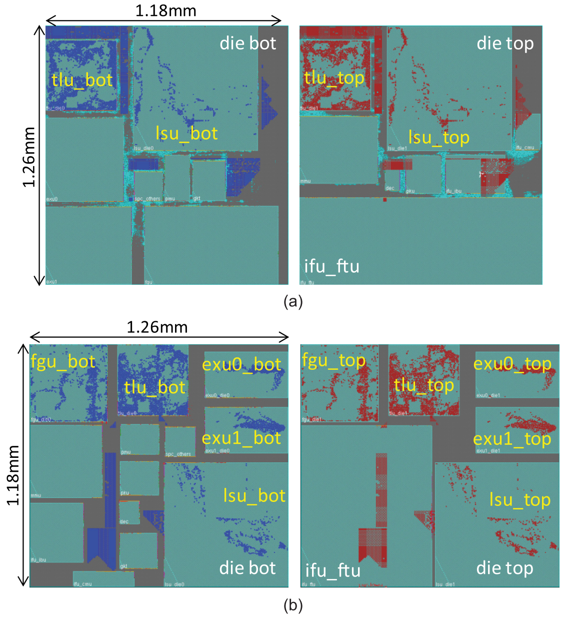


Fig. 8. 3-D design with multiple FUBs folded. Folded FUBs are labeled in yellow. (a) Two FUBs (LSU and TLU) are folded [#TSV: 5272 (2043 + 3229)]. (b) Five FUBs folded [#TSV: 6562 (1331 + 5231)].

TABLE VII
COMPARISON BETWEEN 2-D AND 3-D FOLDED FUBs

	TLU			FGU			EXU			IFU_FTU		
	2D	3D	diff	2D	3D	diff	2D	3D	diff	2D	3D	diff
footprint (mm^2)	0.20	0.12	-42.3%	0.23	0.14	-40.0%	0.15	0.09	-42.0%	0.53	0.31	-41.5%
# cells (×1000)	51.0	50.8	-0.5%	56.2	57.1	+1.5%	15.8	14.5	-8.4%	68.9	66.2	-3.9%
# buffers (×1000)	18.0	16.9	-4.8%	16.3	16.9	+3.7%	6.1	4.8	-21.8%	23.3	20.8	-11.0%
Wirelength (m)	1.89	1.72	-9.1%	1.61	1.48	-7.9%	0.68	0.56	-18.3%	3.19	3.23	+1.5%
# TSV		2633			1098			452			2949	
Total power (mW)	**36.5**	**33.9**	**-7.3%**	**37.8**	**36.7**	**-2.8%**	**43.8**	**42.0**	**-3.9%**	**66.1**	**66.2**	**+0.2%**
Cell power (mW)	8.3	8.4	+0.5%	8.2	8.2	+0.5%	13.4	13.3	-0.7%	18.3	18.8	+2.7%
Net power (mW)	18.0	15.5	-13.9%	19.2	18.2	-5.2%	5.8	4.5	-22.6%	22.1	22.3	+0.9%
Leakage power (mW)	10.2	10.0	-2.0%	10.4	10.3	-1.0%	24.6	24.3	-1.2%	25.7	25.1	-2.3%

TABLE VIII
IMPACT OF MULTIPLE FOLDED FUBs ON INTRABLOCK- AND INTERBLOCK-LEVEL DESIGN METRICS IN T2 CORE WITH A TARGET CLOCK PERIOD OF 1.5 ns. NUMBERS IN PARENTHESES ARE THE DIFFERENCE WITH RESPECT TO THE *3-D w/o Folding* (ZERO FUB FOLDED)

		0 FUB folded	1 FUB folded (LSU)	2 FUBs folded (LSU, TLU)	5 FUBs folded (LSU, TLU, FGU, 2 EXUs)
Wirelength (m)	intra block	17.1	16.3 (-4.7%)	16.2 (-5.3%)	16.0 (-6.4%)
	inter block	2.9	2.1 (-27.6%)	2.8 (-3.4%)	3.9 (+34.5%)
	total	20.0	18.4 (-8.0%)	19.0 (-5.0%)	19.9 (-0.5%)
# buffers (×1000)	intra block	157.6	142.6 (-9.5%)	128.9 (-18.2%)	121.4 (-23.0%)
	inter block	16.0	14.9 (-6.9%)	24.3 (+51.9%)	30.7 (+91.9%)
	total	173.6	157.5 (-9.3%)	153.2 (-11.8%)	152.1 (-12.4%)
Power (mW)	intra block	261.3	252.6 (-3.3%)	240.6 (-7.9%)	227.8 (-12.8%)
	inter block	**28.0**	**21.4 (-23.6%)**	**28.2 (+0.7%)**	**41.2 (+47.1%)**
	total	289.3	274.0 (-5.3%)	268.8 (-7.1%)	269.0 (-7.0%)

The 3-D design results so far are summarized in Table VIII. *3-D w/o folding (zero FUB folded)* is the baseline for this comparison. Surprisingly, *two FUBs folded* shows almost the same power saving as *five FUBs folded*. Although, intrablock power improves as more FUBs are folded, interblock power degrades significantly largely due to increased wirelength and buffer count. In *five FUBs folded*, the interblock wirelength, buffer count, and power increases by 34.5%, 91.9%, and 47.1% compared with *zero FUB folded*. Even though interblock-level power is only around 10% of the T2 core power, this becomes the decisive factor that determines the overall 3-D power benefit. Thus, the number of folded blocks and their 3-D floorplan need to be carefully evaluated to optimize 3-D power benefits.

VIII. TSV SCALING IMPACT ON FUB FOLDING

It is shown that the right amount of TSVs placed at the right spots in 3-D IC layouts is essential to achieve shorter wirelength, smaller critical path delay, and lower power consumption. This also indicates that if the area and capacitance overhead of TSVs themselves become smaller, less design effort is necessary to achieve the aforementioned goals. This is the main motivation behind recent efforts in reducing the size of TSVs [11]. For example, the TSV diameter in recent research reaches 0.7 μm [12]. According to the recent research and ITRS predictions, the TSV diameter will reach the submicrometer domain within the next few years.

In this section, the TSV scaling impact on FUB folding quality is discussed. Table IX shows the TSV technology setup used in this experiment. TSV-large is the one used up to this point, and TSV-small is a submicrometer (0.5-μm diameter) TSV that shows negligible capacitance (=0.43 fF).

TABLE IX
TSV TECHNOLOGY SETUP. *RC* NUMBERS ARE CALCULATED BASED ON [13]

	diameter	height	pitch	R	C
TSV-large	$3\mu m$	$18\mu m$	$6\mu m$	$43m\Omega$	$8.35fF$
TSV-small	$0.5\mu m$	$3\mu m$	$1\mu m$	$513m\Omega$	$0.43fF$

A. FUB Folding With Advanced TSV Technology

In Section VII, the TSV area overhead is discussed as one of critical factors that affects the FUB folding quality. For IFU_FTU, the TSV area overhead is 10.9% of die area, and power consumption in folded case is worse than the 2-D counterpart. Using TSV-small with the same partitioning scheme (= same number of TSVs), the area, wirelength, buffer count, and power of folded IFU_FTU reduces by 3.6%, 7.1%, 10.6%, and 6.0%, respectively, compared with the TSV-large case. Note that the TSV area overhead is only 0.7% in the TSV-small case.

The TSV scaling also increases flexibility in the die partitioning as the TSV size overhead can be ignored during partitioning. In the TLU folding, only one memory macro block was assigned to die bottom along with control units to balance area between dies. With TSV-small, all memory macros are placed in die bottom, and this increases the TSV count by 23.3%. Still, the TSV area overhead is only 2.3%, and this partitioning improves the area, wirelength, buffer count, and power of folded TLU by 11.4%, 9.9%, 17.2%, and 8.6% compared with the TSV-large case. Note that 2633 TSVs in TSV-large case attributed 12.7% of wire capacitance, while 3249 TSVs in TSV-small only contribute 1.0% of wire capacitance. Thus, finer 3-D connections are available for FUB folding without much penalty in TSV-small case.

TABLE X

IMPACT OF INTERBLOCK TSV PITCH ON INTERBLOCK-LEVEL DESIGN METRICS. THE FIVE FUBs FOLDED CASE IS USED

	pitch $1\mu m$	pitch $3\mu m$	diff
Wirelength (m)	3.03	3.02	-0.3%
# buffers (×1000)	21.7	20.2	-6.9%
Power	31.1	30.3	-2.6%

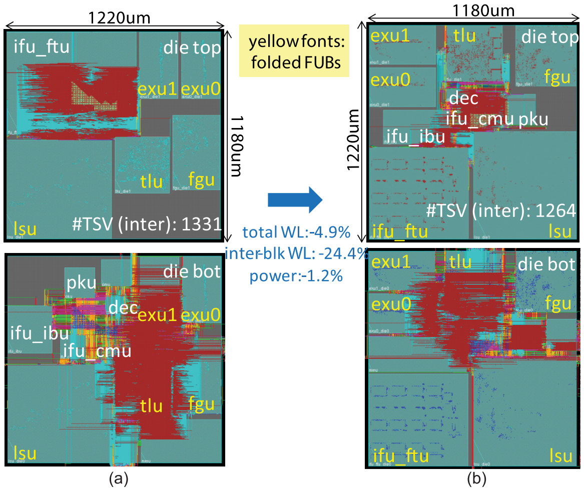


Fig. 9. Core design with folded FUBs using TSV-small. (a) Five FUBs folded excluding IFU_FTU. (b) Six FUBs folded including IFU_FTU.

B. Careful Interblock-Level Design

The TSV scaling provides improved design metrics in folded FUBs in all examined aspects. However, interblock-level routing that utilizes many global wires becomes challenging with the smaller TSV pitch (=1 μm) unless the interconnect scaling is provided as well. Especially, the routing in die top where folded FUBs act as routing blockages is harder. Unlike the TSV-large case, severe routing congestion is observed nearby interblock TSVs. This increases the interblock wirelength and power by 23.4% and 21.9%, respectively, compared with the TSV-large case, which diminishes the power gain obtained from FUB folding with TSV-small.

The pitch of TSV-small is increased to 3 μm to improve the interblock routing congestion. By relaxing interblock TSV pitch, the interblock-level power consumption reduces by 2.6% as shown in Table X. Although the wirelength reduction is only 0.3%, the buffer count decreases by 6.9%. This indicates the coupling capacitance reduction due to less routing congestion and hence reduces buffer count.

C. Overall Comparison

As TSV scaling helps reduce power consumption in the folded IFU_FTU, the folded IFU_FTU block is also integrated in the 3-D design with TSV-small. *Five FUBs folded* and *six FUBs folded* designs are shown in Fig. 9. Note that interblock TSV pitch of 3 μm is used for TSV-small case.

By folding IFU_FTU, no connections are needed between IFU_FTU pins and TSV landing pads at M9 in die top as all IFU_FTU pins are placed in die bottom. This relieves interblock routing demand in die top where folded FUBs act as routing blockages as shown in Fig. 9(b). As a result, the interblock wirelength reduces by 24.4% in the *six FUBs folded* case compared with the *five FUBs folded* case. This further reduces the total power by 1.2%.

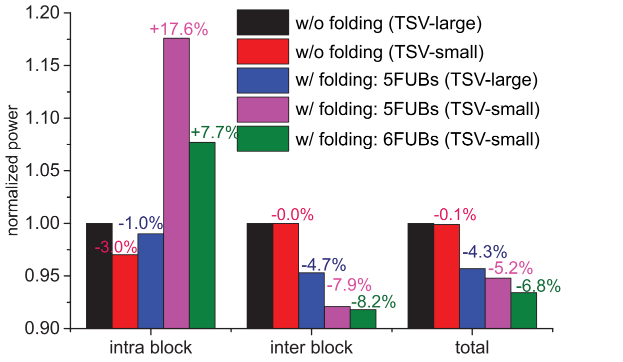


Fig. 10. TSV scaling impact on 3-D power. A target clock period of 1.5 ns is used. FUB folding with TSV-small improves intrablock-level power, while worsens interblock-level power.

The overall power comparison between 3-D designs with a different number of folded FUBs and TSV sizes is presented in Fig. 10. *3-D w/o folding* with TSV-large is the baseline. First, TSV-small mostly reduces intrablock power by folding more FUBs with smaller area and less TSV area/capacitance overhead. In addition, finer 3-D connections enable better 3-D partitioning and power reduction as discussed in TLU folding. With *six FUBs folded* using TSV-small, the total power consumption reduces by 6.8% compared with *3-D w/o folding* using TSV-large and by 27.8% compared with the 2-D design.

Second, although the interblock TSV pitch in TSV-small case is increased to mitigate the interblock routing congestion issue, yet interblock power increases significantly compared with *3-D w/o folding* and *3-D w/ folding* using TSV-large. Thus, 3-D floorplan and TSV placement scheme that resolve these issues need to be developed given process technology. Third, TSV scaling does not affect much in *3-D w/o folding* as there are not many TSVs deployed. In our design, there are only 2979 TSVs, while the *six FUBs folded* case with TSV-small contains 10 060 TSVs (intrablock 8796, interblock 1296).

Thus, the TSV scaling is a very important factor that improves 3-D power benefit. This is largely due to negligible TSV area/capacitance overhead and hence better FUB folding opportunities. However, design challenges such as interblock routing congestion are carefully considered to fully benefit the newer TSV technology.

IX. CONCLUSION

In this paper, the power benefit of 3-D ICs was demonstrated with an OpenSPARC T2 core. Four design techniques were explored to optimize power in 3-D IC designs: 1) 3-D floorplanning; 2) intrablock-level metal layer usage control; 3) DVT design; and 4) FUB folding. The impact of multiple FUB foldings on 3-D power benefit was also investigated. TSV scaling provided the 3-D power benefit with less TSV overhead and hence finer 3-D connections for FUB folding.

With the aforementioned methods combined, the total power saving of 27.8% has been achieved against the 2-D counterpart.

References

[1] B. Black *et al.*, "Die stacking (3D) microarchitecture," in *Proc. Annu. Int. Symp. Microarchit.*, 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] G. Luo, Y. Shi, and J. Cong, "An analytical placement framework for 3-D ICs and its extension on thermal awareness," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 32, no. 4, pp. 510–523, Apr. 2013.

[4] M.-K. Hsu, V. Balabanov, and Y.-W. Chang, "TSV-aware analytical placement for 3-D IC designs based on a novel weighted-average wirelength model," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 32, no. 4, pp. 497–509, Apr. 2013.

[5] D. H. Kim, K. Athikulwongse, and S. K. Lim, "A study of through-silicon-via impact on the 3D stacked IC layout," in *IEEE/ACM Int. Conf. Comput.-Aided Des. Dig. Tech. Papers*, Nov. 2009, pp. 674–680.

[6] Oracle. *OpenSPARC T2*, accessed on Jan. 2013. [Online]. Available: http://www.oracle.com

[7] P. Spindler, U. Schlichtmann, and F. M. Johannes, "Kraftwerk2—A fast force-directed quadratic placement approach using an accurate net model," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 27, no. 8, pp. 1398–1411, Aug. 2008.

[8] D. H. Kim, R. Topaloglu, and S. K. Lim, "Block-level 3D IC design with through-silicon-via planning," in *Proc. Asia South Pacific Design Autom. Conf.*, Feb. 2012, pp. 335–340.

[9] M. B. Healy *et al.*, "Design and analysis of 3D-MAPS: A many-core 3D processor with stacked memory," in *Proc. IEEE Custom Integr. Circuits Conf.*, Sep. 2010, pp. 1–4.

[10] 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. 108–109.

[11] D. H. Kim, S. Kim, and S. K. Lim, "Impact of nano-scale through-silicon vias on the quality of today and future 3D IC designs," in *Proc. Int. Workshop Syst. Level Interconnect Predict.*, Jun. 2011, pp. 1–8.

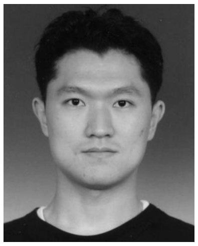
[12] M. Koyanagi, T. Fukushima, and T. Tanaka, "High-density through silicon vias for 3-D LSIs," *Proc. IEEE*, vol. 97, no. 1, pp. 49–59, Jan. 2009.

[13] G. Katti, M. Stucchi, K. D. 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.

Moongon Jung (S'11–M'15) received the B.S. degree in electrical engineering from Seoul National University, Seoul, South 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 the Intel Labs at Intel Corp., Santa Clara, CA, USA, as a Research Scientist in 2014, and is involved in the design methodologies for future technologies. His current research interests include computer-aided design for VLSI circuits, especially on physical design methods for low power 3-D ICs and thermomechanical reliability analysis and optimization of TSV-based 3-D ICs. His research on thermomechanical reliability of 3-D ICs was featured as a Research Highlight in the Communication of the ACM in 2014.

Dr. Jung's works were nominated for the Best Paper Award at DAC 2011 and DAC 2012, and the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN in 2013.



Taigon Song (S'09) received the B.S. degree in electrical engineering from Yonsei University, Seoul, South Korea, in 2007, and the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South 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 On-Line Electric Vehicle Business Department, KAIST, 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 the 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 14.

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, 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 has authored 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). He led the Cross-Center Theme on 3-D Integration for the Focus Center Research Program of Semiconductor Research Corporation during 2010–2012. His current research interests include the architecture, design, test, and EDA solutions for 3-D ICs. His research was featured as a Research Highlight in the Communication of the ACM in 2014.

Dr. Lim was a recipient of the National Science Foundation Faculty Early Career Development Award in 2006. He was on the Advisory Board of the ACM Special Interest Group on Design Automation (SIGDA) during 2003–2008, 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 during 2007–2009, and the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS since 2013. 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.