On Enhancing Power Benefits in 3D ICs: Block Folding and Bonding Styles Perspective

Moongon Jung, Taigon Song, Yang Wan, Yarui Peng, and Sung Kyu Lim
School of ECE, Georgia Institute of Technology, Atlanta, GA, USA
moongon@gatech.edu, limsk@ece.gatech.edu

ABSTRACT

Low power is widely considered as a key benefit of 3D ICs, yet there have been few thorough design studies on how to maximize power benefits in 3D ICs. In this paper, we present design methodologies to reduce power consumption in 3D ICs using a large-scale commercial-grade microprocessor (OpenSPARC T2). To further improve power benefits in 3D ICs on top of the traditional 3D floorplanning, we study the impact of block folding and bonding styles. We also develop an effective method to place face-to-face vias for our 2-tier 3D design for power optimization. With aforementioned methods combined, our 3D designs provide up to 20.3% power reduction over the 2D counterpart under the same performance.

Categories and Subject Descriptors

B.7.2 [**Performance and Reliability**]: Performance Analysis and Design Aids

General Terms

Design

Keywords

3D IC, block folding, bonding style, power benefit

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

Most of previous works showed 3D power benefit by 3D floorplanning. In this work, we present 3D block folding methods to further reduce power in 3D ICs on top of the traditional 3D floorplanning. We also study impacts of bonding styles, i.e., face-to-back (F2B) and face-to-face (F2F), on 3D power consumption. Our study is based on the OpenSPARC T2 (an 8-core 64-bit SPARC SoC) design database and a Synopsys 28nm PDK with nine metal layers that are both available to the academic community. We build GDSII-level 2D and 2-tier 3D 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 3D to reduce power. Next, we explore block folding methods, i.e., partitioning a block into two sub-blocks and bonding them, to achieve power savings in the 3D design. We employ a mixed-size 3D placer for block folding. Then, we study how bonding styles affect the folded design quality. For the F2F bonding case, we develop an efficient method to place face-to-face vias for our 2-tier 3D design utilizing existing commercial CAD tools with in-house scripts. Lastly, we demonstrate system-level 3D power benefits by assembling folded blocks in different bonding scenarios. Additionally, the impact of dual-Vth design technique on 2D and 3D designs is presented.

2. SIMULATION SETTINGS

2.1 Benchmark Design

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 banks (L2D), eight L2-cache tags (L2T), eight L2-cache miss buffers (L2B), and a cache crossbar (CCX). Each block is synthesized with 28nm cell and memory macro libraries. 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 PLL (analog block) in a clock control unit (CCU) is replaced by ideal clock sources. Thus, a total of 46 blocks are floorplanned. For the 2D design, we try to follow the original T2 floorplan [3] as much as possible as shown in Figure 8(a). In addition, special cares are taken to use both connectivity and data flow between blocks to minimize inter-block wirelength.

2.2 3D IC Design Flow

Our RTL-to-GDSII tool chain for 3D IC design is based on commercial tools and enhanced with our in-house tools to handle TSVs

This work is supported by Intel Corporation through Semiconductor Research Corporation (ICSS Task 2293).

DAC '14, Jun 01-05 2014, San Francisco, CA, USA
Copyright 2014 ACM 978-1-4503-2730-5/14/06 ...$15.00.
http://dx.doi.org/10.1145/2593069.2593167.

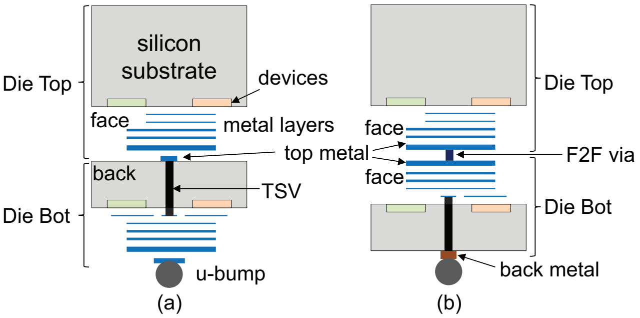


Figure 1: Die bonding styles. (a) face-to-back. (b) face-to-face.

Table 1: 3D 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

and 3D stacking. With initial design constraints, the entire 3D netlist is synthesized. The layout of each die is done separately based on the 3D floorplanning result. With a given target timing constraint, 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 3D static timing analysis using Synopsys PrimeTime to obtain new timing constraints for each block's I/O pins as well as die boundaries (= TSVs).

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-CTS (clock tree synthesis), post-CTS, and post-route optimizations. Note that we utilize all nine metal layers for SPC design that requires most routing resources among all blocks, but seven layers for all other blocks. Thus, top two metal layers can be utilized for over-the-block routing in the chip-level design.

2.3 Die Stacking Technology

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

Our 3D interconnect settings are summarized in Table 1. TSV resistance and capacitance values are calculated based on the model in [4]. We assume that TSV diameter is much larger than F2F via size as manufacturing reliable sub-micron TSVs is challenging. Additionally, the physical size of F2F via can be made comparable to the top metal dimension, around twice the minimum top metal (M9) width in our setup.

3. 3D FLOORPLANNING BENEFITS

3.1 3D Floorplan Options

The T2 chip contains eight copies of SPARC cores (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 3D floorplan cases to examine their impact on power as shown in Figure 8(b) and (c): (1) core/cache stacking: all cores are in one die and all L2D blocks are in another die, and (2) core/core stacking: four cores and L2-cache blocks are located in each die.

Table 2: Comparison between 2D and 3D block-level designs with a target clock frequency of 500MHz. Numbers in parentheses are differences against the 2D 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%)

We use the F2B bonding style only for 3D block-level designs as a baseline. The 3D floorplanner in [5] is modified to handle user-defined floorplans, and then used to determine TSV locations with an objective of minimizing inter-block wirelength. TSV arrays are treated as additional blocks in this flow, hence all TSVs can be placed outside blocks only.

3.2 2D vs. 3D Floorplanning

We now compare our 2D and 3D block-level designs with a target CPU clock frequency of 500MHz that is the highest performance that our 2D design achieves.[1] Design metrics in 2D and 3D designs are shown in Table 2. First, we observe 16.3% buffer count and 5.0% wirelength reduction in the core/cache 3D stacked design and 15.2% and 5.4% reduction in the core/core 3D case compared with the 2D counterpart. In addition, inter-block wirelength reduces by 15.6% (core/cache) and 17.8% (core/core), which is a direct consequence of 3D floorplanning.

Second, most importantly, the 3D designs reduce power consumption over the 2D 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 3D design. This is because the 3D design utilizes more smaller cells than the 2D thanks to better timing, i.e., more positive timing slack in paths. With the positive slack, cells can be downsized in the 3D design if this change still meets the timing constraint during power optimization stages.

This smaller cell size in the 3D 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, 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 3D 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 3D floorplans only, since there are not

[1] Our designs run slower than OpenSPARC T2 that runs at 1.4GHz [3]. This is mainly because some custom memory blocks 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.

Table 3: 2D design characteristics used for block folding candidate selection. Long wires are defined as wires longer than 100X standard cell height. CPU clock runs at 500MHz and I/O clock at 250MHz.

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

many floorplan options for the T2 design that contains multiple large same-size blocks that need to be placed in a specific way.

4. BLOCK FOLDING BENEFITS

So far, block-level designs are implemented for both 2D and 3D designs. Thus, even in 3D designs, each block is located in the same die. In addition, TSVs are always outside blocks and used only for inter-block 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 intra-block connections, on power consumption.

4.1 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 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 3. 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 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 relatively low net power portion compared with other blocks, as L2D is the memory (and its power) dominated design that contains 512KB (32 16KB 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 our study, we define long wires as wires longer than 100X standard cell height. We observe that SPC, RTX, and CCX have a large number of long wires.

In this work, we fold five blocks: SPC, CCX, L2D, L2T, and RTX. In the following sections, we discuss block folding methodologies for SPC, CCX, and L2D.Each block shows distinctive folding characteristics. Before this, we briefly explain our mixed-size 3D placer that is employed for block folding.

4.2 CAD Tool Need: Mixed-size 3D Placer

A TSV-based 3D placer, based on a system of supply/demand of placement space was presented in [6], but it lacks the capability to handle hard macros. 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 2D placer [7], this leads to large whitespace regions in the vicinity of the hard macros called halos. The authors of [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 banks in L2D, 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.

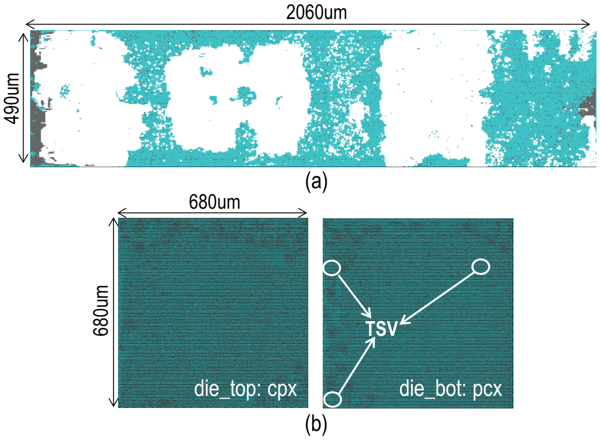


Figure 2: CCX 2D and 3D layouts. (a) 2D design. CPX is highlighted with white color. (b) 3D design (# TSV=4).

4.3 Folding CCX Block

In T2, eight cores use the cache crossbar (CCX) to exchange data in eight L2-cache banks. This CCX is divided into two separate modules, the processor-to-cache crossbar (PCX) and the cache-to-processor crossbar (CPX). There are no signal connections between these two blocks except clock and a few test signals. The PCX occupies 48% of the block area and utilizes 48% of the CCX I/O pins, and the CPX uses the rest of them. Thus, the natural way to fold this CCX is placing the entire PCX block in one die and the CPX in another die along with related I/O pins.

The 2D and 3D CCX layouts are shown in Figure 2. Interestingly, in the 2D design, we see that PCX (and CPX) block is separated into several groups. The PCX has eight sources (SPCs) and nine targets (eight L2-cache banks and I/O bridge). Depending on the target core and L2-cache bank locations in the chip-level floorplan, PCX I/O pin locations are determined, which in turn attracts connected cells. Because of this, the PCX block is not fully gathered, which degrades cell-to-cell wirelength significantly.

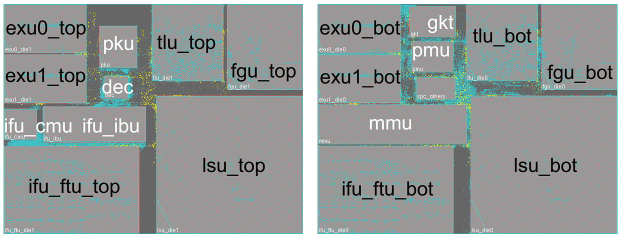
However, folding CCX eliminates this problem and hence cell-to-cell wirelength decreases by 31.7% compared with the 2D. The folded CCX leads to 54.6% reduced footprint, 28.8% shorter wirelength, 62.5% less buffer count, and 32.8% power reduction over the 2D counterpart. Note that only four signal TSVs are used in this 3D design, and this is due to the unique characteristics of CCX. We also examine whether different 3D partitions with more 3D connections can provide better power savings. However, as we increase the TSV count up to 6,393, largely due to the area overhead by TSVs (13.3%), the 3D power benefit reduces down to 23.4%.

4.4 Folding L2D Block

The single L2-cache data bank contains 512KB memory array. This L2D is further divided into four logical sub-banks. In our implementation, each sub-bank group is partitioned into eight blocks of size 16KB each. This L2D is a memory macro dominated design, and hence there are not many 3D partitioning options to bal-

Table 4: Comparison between 2D and 3D L2D designs.

L2D	2D	3D	diff
footprint (mm^2)	2.54	1.31	-48.4%
Wirelength (m)	3.41	3.19	-6.4%
# cells ($\times 10^6$)	53.1	42.2	-20.5%
# buffers ($\times 10^6$)	38.1	25.3	-33.5%
Total power (mW)	**172.9**	**164.0**	**-5.1%**
Cell power (mW)	25.8	24.6	-4.7%
Net power (mW)	50.5	44.5	-11.9%
Leakage power (mW)	96.6	94.9	-1.8%



Top die	Bottom die

Figure 3: Second-level folding of a SPARC core. 6 FUBs shown in black text are folded (# F2F via: 10,251).

ance area after folding. Thus, two sub-banks are placed in each die along with related logic cells.

Although, the buffer count and wirelength reduce by 33.5% and 6.4%, respectively in the folded L2D, their impact on the total power saving is not significant (5.1% reduction over 2D) as shown in Table 4. This is because both cell and leakage power are dominated by memory macros, which 3D folding cannot help unless these memory macros themselves are folded. Additionally the net power portion is only about 29% of the total power in 2D, and hence the small net power reduction in 3D does not lead to a noticeable total power reduction. Still, the footprint area reduction of 48.4% is nonnegligible and this might affect chip-level design quality.

4.5 Second-level Folding SPC Block

In case of SPARC core (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 (EXU), a floating point and graphics unit (FGU), five instruction fetch units (IFU), 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 4.1, and based on this six FUBs are folded as shown in Figure 3. We call this second-level folding. With this second-level folding, we obtain 9.2% shorter wirelength, 10.8% less buffers, and 5.1% reduced power consumption than the SPC without second-level folding, i.e. a block-level 3D design of the SPC. Additionally, our 3D SPC achieves 21.2% power saving over the 2D SPC.

5. FACE-TO-FACE BONDING BENEFIT

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

5.1 CAD Tool Need: F2F Via Placer

Several previous works discussed TSV-aware 3D placement algorithms [6, 8, 9] assuming F2B bonding. However, there is no existing work on how to decide F2F via locations in F2F bonding.

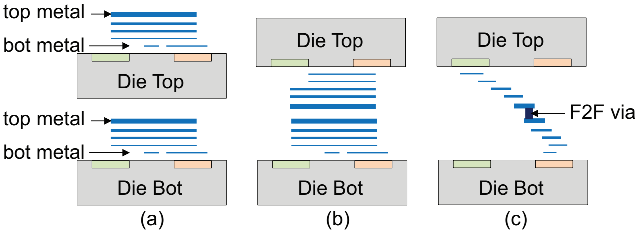


Figure 4: Finding F2F via locations by 3D net routing. (a) Run 3D placer assuming an ideal 3D interconnect. (b) Create 2D-like 3D design files (Verilog, LEF, and DEF). (c) Route 3D nets and extract F2F via locations.

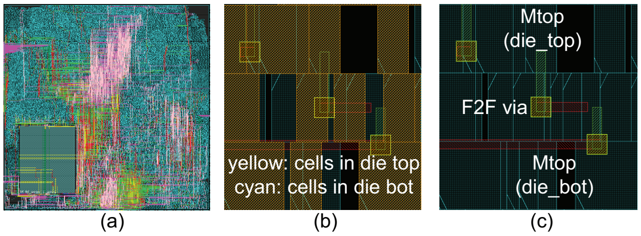


Figure 5: 3D net routing. (a) Layout shot after 3D net routing. (b) Close-up shot showing cells in both dies. (c) Close-up shot of 3D net routing showing F2F vias.

Unlike TSVs, F2F vias can be located above cells and macro blocks. Thus, 3D placement algorithms are not adequate for F2F via placement. In this section, we discuss how to find F2F via locations by 3D net routing using existing commercial CAD tools.

A simplified flow is shown in Figure 4. With a given die partitioning result, we first run the 3D placer assuming an ideal 3D interconnect element (TSV size = 0) and obtain netlist and DEF (design exchange format) files for both dies. Next, we create 2D-like 3D design files, i.e., netlist, DEF, and LEF (library exchange format), that can be fed into commercial 2D place and route tools (in our case, Cadence Encounter). For example, 3D LEF file contains the interconnect structure for F2F bonding as well as cells and memory macros in both dies as shown in Figure 4(b). For this, we modify metal layer and cell names such as M1_die_top, M1_die_bot, INVX1_die_top, and INVX1_die_bot.

Once all 3D design files are ready, we employ a commercial CAD tool to route 3D nets. In tool's perspective, these 3D nets are still 2D nets with cell pins located in either M1_die_top or M1_die_bot. Note that we exclude 2D net routing by modifying netlists: tying 2D nets to ground. By this, F2F via locations are not affected by 2D net routing and possible congestions. Layouts with 3D net routing and F2F via locations are shown in Figure 5. From this result, we extract F2F via locations for all 3D nets, and use these F2F via locations in each die design.

5.2 F2F Impact on Block Folding

F2F vias do not consume silicon area, and hence 3D footprint area can be further reduced as shown in Figure 6. 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 Figure 6(a), all F2F vias are located on 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

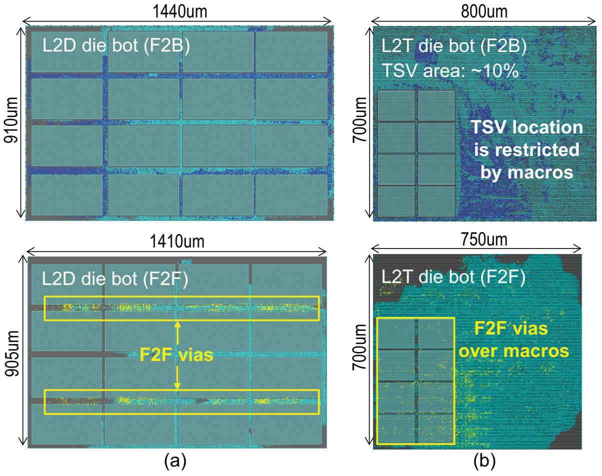


Figure 6: Bonding style impact on 3D placement. Blue rectangles are TSV landing pads at M1 and yellow dots are F2F vias. (a) L2D bottom die. (b) L2T bottom die.

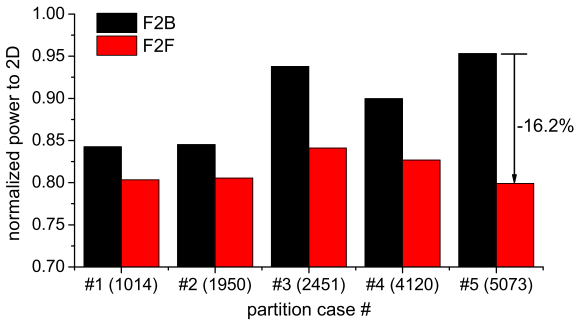


Figure 7: Bonding style impact on power in L2T folding. Numbers in parentheses are number of TSVs/F2F vias.

and power. For the same 3D 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 Figure 6(b), F2F vias are found over large memory macros. However, TSVs are ousted from memory macro area, which increases wirelength.

The five partitioning cases for L2T are implemented in both F2B and F2F bonding styles. Power comparisons between both bonding styles are shown in Figure 7. First of all, F2F wins over F2B bonding style in all cases. This is the combined effect of reduced footprint, better 3D 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 3D connections. Especially, the partition #5 that shows the smallest 3D power benefit in F2B now achieves the best power saving with F2F bonding. Compared with the F2B case, the F2F case reduces power by 16.2%. In this specific case, the 3D design quality in F2B bonding is degraded largely by TSV area overhead, not by the partition. Third, more 3D connections in F2F style does not necessarily mean better power saving. Although partition #3 and #4 show much better power saving than the F2B cases, these power savings are still less than partition #1 and #2. This emphasizes the importance of die partitioning again.

Table 5: Comparison between 2D, 3D without block folding (core/cache, F2B), and 3D with block folding (5 types of blocks folded, F2F) designs. Dual-Vth design technique is applied to all cases. Numbers in parentheses are difference against the 2D excluding HVT cell count which 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%)

6. FULL-CHIP WITH FOLDED BLOCKS

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

6.1 3D Floorplan with Folded Blocks

Based on the criteria on block folding discussed in Section 4.1, SPC, CCX, L2D, L2T, and RTX have been folded. Unlike other four blocks, RTX runs at I/O clock frequency (= 250MHz). 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 block and NIU. In this study, we implement two 3D designs as shown in Figure 8: (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 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 in top and bottom of the chip as shown in Figure 8(d). Otherwise, these SPC blocks will act as inter-block routing blockages.

In the F2F bonding case, since F2F via is on top of M9, all nine metal layers are used for routing. Thus, folded blocks in F2F bonding are routing blockages for both dies as shown in Figure 8(e). For this reason, although this F2F bonding achieves more power saving than the F2B case in block folding, inter-block design quality could be degraded.

In both bonding style cases, we place CCX in the center. There are about 300 wires between CCX and each SPC (or L2T). Thus, in this implementation, 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 connections are confined in NIU.

6.2 Full-chip Design Comparison

Up to this point, both 2D and 3D designs utilize only regular-Vth (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 2D and 3D designs. Each HVT cell shows around 30% slower, yet 50% lower leakage and 5% smaller cell power consumption than the RVT counterpart.

We now compare three full-chip T2 designs: 2D IC, 3D IC without folding (core/cache stacking, F2B bonding), and 3D 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 5.We first observe higher HVT cell usage in 3D designs, especially for the 3D with folding case (94.0% of cells are HVT). This is largely due to better timing in 3D designs, and this helps reduce power in 3D ICs further. The 2D DVT design reduces power by 9.5% and the 3D with folding by 11.4% compared with the corresponding RVT only design, which again shows the benefit of 3D designs.

Most importantly, the 3D with folding case with F2F bonding reduces the total power by 20.3% compared with the 2D and by 10.0% compared with the 3D without folding case. This clearly demonstrates the powerfulness of block folding along with its bonding style in 3D designs for power reduction.

7. CONCLUSIONS

In this paper, the power benefit of 3D ICs was demonstrated with an OpenSPARC T2 chip. To further enhance the 3D power benefit on top of the conventional 3D floorplanning method, block folding methodologies and bonding style impact were explored. We also developed an efficient method to find face-to-face via locations for 2-tier 3D ICs, and showed more 3D power reduction with F2F bonding than F2B. With aforementioned methods, the total power saving of 20.3% has been achieved against the 2D counterpart.

Note that the 3D power benefit will improve even more with faster clock frequency. With better timing in 3D, the discrepancy in cell size and HVT cell usage between 2D and 3D designs will increase, which in turn will enhance the 3D power saving. In addition, our future work will address thermal issues in various 3D design styles with different bonding styles, the impact of parasitics such as TSV-to-wire coupling capacitance on 3D power, and other sources of 3D power benefit loss.

8. REFERENCES

[1] B. Black et al. Die Stacking (3D) Microarchitecture. In *Proc. Annual Int. Symp. Microarchitecture*, 2006.

[2] U. Kang et al. 8 Gb 3-D DDR3 DRAM Using Through-Silicon-Via Technology. In *IEEE J. Solid-State Circuits*, 2010.

[3] U. Nawathe et al. An 8-Core 64-Thread 64b Power-Efficient SPARC SoC. In *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2007.

[4] G. Katti et al. Electrical Modeling and Characterization of Through Silicon via for Three-Dimensional ICs. In *IEEE Trans. on Electron Devices*, 2010.

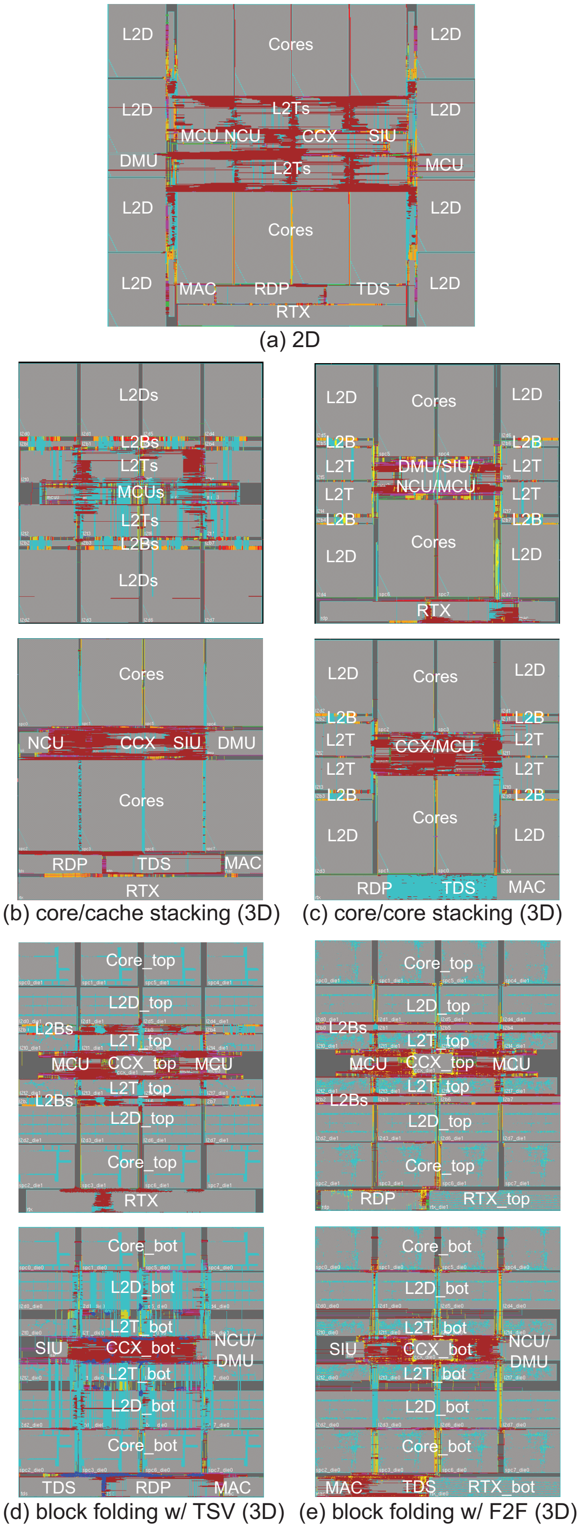
[5] D. Kim et al. Block-level 3D IC Design with Through-Silicon-Via Planning. In *Proc. Asia and South Pacific Design Automation Conf.*, 2012.

[6] D. Kim et al. A Study of Through-Silicon-Via Impact on the 3D Stacked IC Layout. In *Proc. IEEE Int. Conf. on Computer-Aided Design*, 2009.

[7] P. Spindler et al. Kraftwerk2 - A Fast Force-Directed Quadratic Placement Approach Using an Accurate Net Model. In *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 2008.

[8] G. Luo et al. An Analytical Placement Framework for 3-D ICs and Its Extension on Thermal Awareness. In *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 2013.

[9] M. Hsu et al. TSV-Aware Analytical Placement for 3-D IC Designs Based on a Novel Weighted-Average Wirelength Model. In *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 2013.



(a) 2D

(b) core/cache stacking (3D) (c) core/core stacking (3D)

(d) block folding w/ TSV (3D) (e) block folding w/ F2F (3D)

Figure 8: GDSII layouts of 5 design styles of OpenSPARC T2 (full-chip) we compare: (a) 2D design (9x7.9mm^2), (b) core/cache stacking (6x6.4mm^2, #TSV=3,263), (c) core/core stacking (6x6.4mm^2, #TSV=7,606), (d) block folding with TSVs (6x6.6mm^2, #TSV=69,091), (e) block folding with F2F (6x6.6mm^2, #F2F=112,308). Cyan dots inside blocks are intra-block TSVs or F2F vias.