On GPU Bus Power Reduction with 3D IC Technologies

Young-Joon Lee and Sung Kyu Lim
School of ECE, Georgia Institute of Technology, Atlanta, Georgia, USA
yjlee@gatech.edu, limsk@ece.gatech.edu

***Abstract*— The complex buses consume significant power in graphics processing units (GPUs). In this paper, we demonstrate how the power consumption of buses in GPUs can be reduced with 3D IC technologies. Based on layout simulations, we found that partitioning and floorplanning of 3D ICs affect the power benefit amount, as well as the technology setup, target clock frequency, and circuit switching activity. For 3D IC technologies using two dies, we achieved the total power reductions of up to 21.5% over a baseline 2D design.**

I. Introduction

The ever-increasing performance demand for mobile devices and slowly improving battery technology calls for highly power efficient designs. Today, consumers are expecting more engaging visual experience on mobile products such as cell phones and tablets that require cutting-edge graphics processing units (GPUs). To satisfy such needs, on top of the technology node scaling, industry is heading towards 3D ICs for better power efficiency. The through-silicon-via (TSV) technology enables multiple die stacking with low parasitic loss, and the fabrication technology is getting mature.

The core+memory stacking [1] is considered as a good 3D IC application for mobile application processors. In addition, 3D stacked DRAM chips are power-efficient [2]. Then, the next target for 3D stacking is to divide the core part, including GPUs, into multiple dies. In state-of-the-art GPUs, many small computation cores share a register file and cache memories through extensive use of buses. To further improve the power efficiency, it is essential to reduce the power consumed in these buses. The major contributions of this work are as follows:

- To the best of our knowledge, this is the first work to present the power benefit of 3D IC technology for GPUs. We demonstrate how floorplan affects design quality for 2D and 3D designs. For 3D, different partitioning methods are applied to study the trade-off between TSV count and power benefit.
- We demonstrate how 3D IC technology settings affect design quality. We compare three 3D IC technology options: face-to-back bonding with large TSVs and small TSVs, and face-to-face bonding.
- We analyze the power benefit amount based on detailed layouts and sign-off analysis. Factors affecting power benefit amount, such as clock period and circuit switching activity, are explored.

This research is supported by Intel Corporation under the Semiconductor Research Corporation (SRC), Task 2293.001.
978-3-9815370-2-4/DATE14/©2014 EDAA

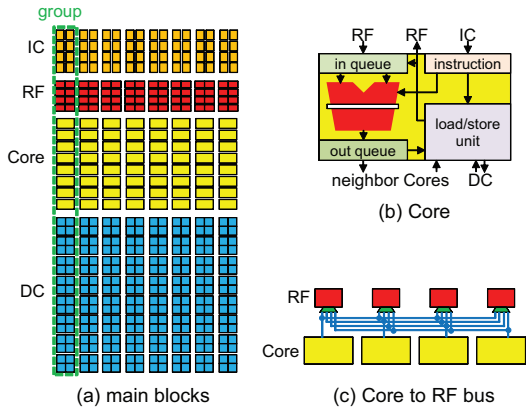

Fig. 2. Block diagram of our GPU. In (c), only 4 out of 8 Cores and RFs in a group are shown for the simplicity of illustration.

II. 2D Design

A. Overall GPU Architecture

Our GPU architecture consists of four basic modules as shown in Fig. 1: **Core**, register file (**RF**), instruction cache (**IC**), and data cache (**DC**). In Fig. 2(a), the entire 64-core architecture is divided into 8 *groups*. A group contains 8 Cores and one eighth of the entire RF, IC, and DC. Per each cycle, all Cores in a group access an instruction from IC and data from RF or DC within the group.

As shown in Fig. 2(b), our Core is based on a simple architecture with a computation path and a load/store path. Instructions sent from IC control the computation block and the load/store unit. The data for computation is delivered to the in queue, either from RF or directly from load/store unit. The output of computation is temporarily stored in out queue, which may be sent back to RF, DC, or forwarded to neighbor Cores through dedicated forwarding paths.

The Cores in the group shares the RF and DC. For example, as shown in Fig. 2(c), each Core has the access to any RF banks in the group, and vice versa. The bus architecture is a switch matrix implemented by multiplexers (MUXes). The bus width of the connections including data and control is as follows: core-to-RF = 249, core-to-IC = 73, core-to-DC = 79, and core-to-core = 66. Our GPU architecture is summarized in Table I. The baseline clock frequency is 667MHz.

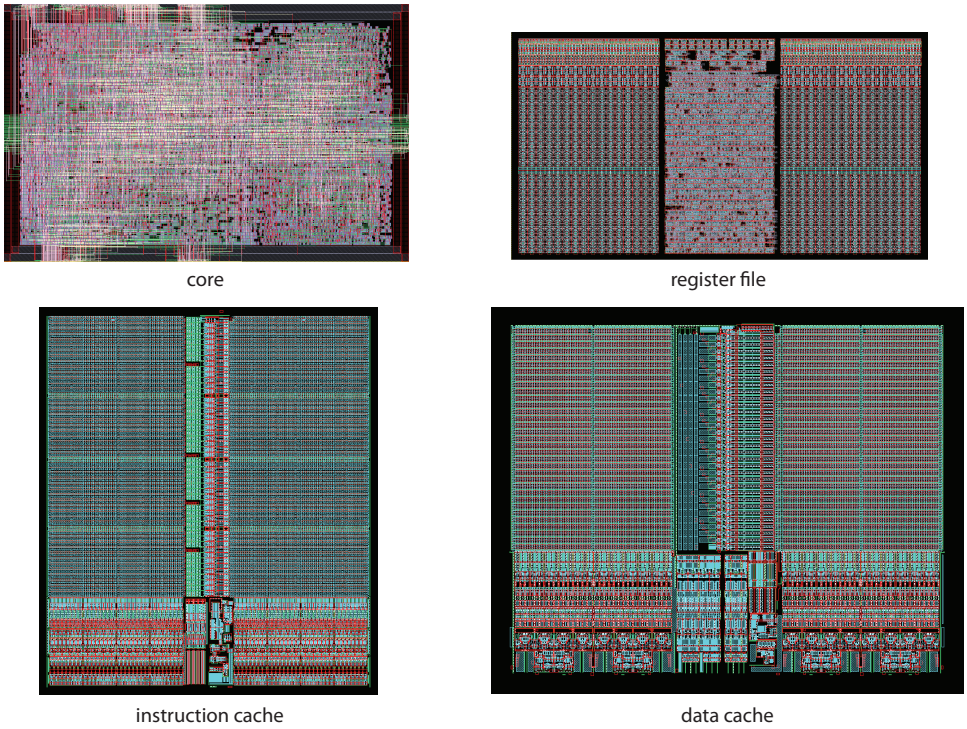

Fig. 1. GDSII layouts of our GPU modules.

TABLE I
SUMMARY OF OUR GPU ARCHITECTURE.

Core	a pipelined computation unit, a load/store unit, and a forwarding logic
RF	total 8KB, divided into 64 sub-banks, 4 read / 3 write ports per each bank
IC	total 96KB, divided into 64 sub-banks
DC	total 256KB, divided into 256 sub-banks

B. 2D Floorplan

In this research, we perform so called RTL-to-GDSII flow for all designs. We customize Synopsys 28nm library [3] for our GPU designs. We synthesize the register-transfer-level (RTL) code into a netlist and perform floorplan, placement, pre-route optimization, routing, and post-route optimization, followed by full-chip static timing analysis (STA) and power analysis. When the design is big with lots of blocks, floorplan affects overall design quality much. The aforementioned buses connect blocks far apart, thus it is essential to find a good floorplan. We explore the design space of floorplan for our baseline 2D design. Our GPU design contains lots of macro blocks with rather regular connectivity. For such a design, it is customary to perform manual floorplans. We performed floorplans using the automated floorplanner in a commercial layout tool, however the design quality was unsatisfactory. Considering the buses among blocks, we created three floorplans that are shown in Fig. 3:

- Tile-Based (**2D-TB**): A Core, an RF, an IC, and four sub-banks of DC are clustered to form a tile. Then, the tiles are placed in 8x8-grid fashion.
- Semi-Clustered (**2D-SC**): The Cores, RFs, DCs, and ICs in a group are clustered together. The Cores, RFs, and ICs are placed in a column, while DCs are placed on the left side.
- Fully Clustered (**2D-FC**): All Cores, RFs, DCs, and ICs are clustered together by the kind.

Layout results for the floorplan options are summarized in Table II. The 'bus' means the MUXes and buffers/inverters for inter-block buses. Memory macros mean RFs, ICs, and DCs. The 'buffer' means buffers and inverters. All designs closed the timing. Since the buses connect all Cores to all RFs (and ICs and DCs), it is better to group the blocks of a same kind together to reduce the wirelengths of buses. In 2D-TB, since all blocks of a same kind are spread apart, the bus length becomes longer. In 2D-FC, all buses were laid out in the column direction, causing severe routing congestions. Thus, we widened the floorplan to resolve the congestion. Compared with 2D-FC, in 2D-SC, Core-DC buses lie in the row direction and hence mitigate congestions. From the wirelength, buffer count, and net power of buses, it is clear that 2D-SC is the best floorplan. Note that in 2D-SC, the net power of buses (bus-net) is 33.4% of the total power, which is a significant amount; by folding the buses into 3D, we mainly reduce this bus-net power. All Cores and memory macros (RFs, ICs, and DCs) consume 25.7% and 33.7% of total power, respectively.

The power breakdown for 2D-SC is summarized in Table III. About 40.8% of the total power is used for buses, which is a significant amount. This is because the size (length and width) of the bus structures in our GPU is considerably large. Thus, it is important to reduce the bus power. Interestingly,

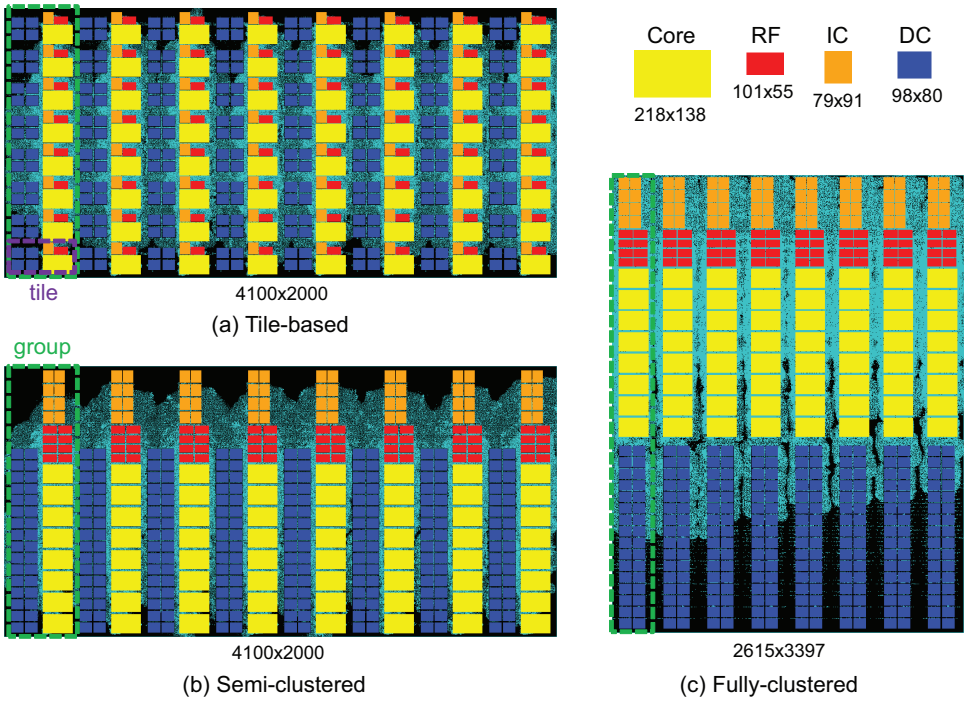

Fig. 3. Floorplans for 2D. Skyblue dots represent standard cells for buses. Numbers represent dimensions in μm.

TABLE II

LAYOUT COMPARISON AMONG 2D FLOORPLANS.

		2D-TB	2D-SC	2D-FC
footprint (mm^2)		8.20	8.20	8.88
wirelength (m)	total	53.62	46.47	53.87
	all buses	43.96	36.80	44.20
	all Cores	9.66	9.66	9.66
cells (thousand)	bus-MUX	118	118	118
	bus-buffer	252	206	237
	all Cores	415	415	415
power (W)	total	1.855	1.679	1.935
	bus-cell	0.062	0.050	0.079
	bus-net	0.695	0.561	0.755
	bus-leakage	0.093	0.074	0.101
	all memory macros	0.565	0.565	0.565
	all Cores	0.439	0.431	0.436

TABLE III

POWER BREAKDOWN FOR 2D-SC.

	total power			bus power		
	all Cores	memory macros	buses	cell	net	leakage
percent	25.6%	33.6%	40.8%	7.3%	81.9%	10.8%

about 81.9% of the bus power is consumed for net (switching) power. This justifies our focus on the bus power reduction, because the reduced bus lengths in 3D IC would lead to a significant total power reduction.

III. 3D DESIGN

A. Underlying 3D IC Technology

In this research, we design the GPU in 3D ICs using two dies. As shown in Fig. 4, there are two kinds of die bonding styles for 3D ICs: face-to-back (**F2B**) and face-to-face (**F2F**).

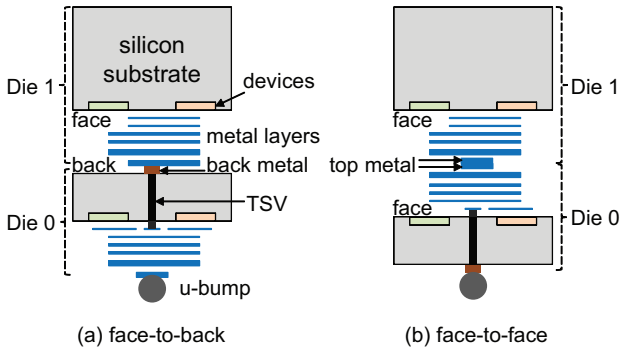

Fig. 4. Die bonding styles for TSV-based 3D ICs.

In F2B bonding, the top metal pad of a die is bonded to the back metal pad of an adjacent die. The 3D connection density is limited by the TSV pitch. In contrast, for two die designs, F2F bonding does not require TSVs for inter-die communications, because the inter-die connections are made by top metal pads.

Our 3D interconnect settings are summarized in Table IV. Note that we have two settings for TSVs: TSV-large and TSV-small. Resistance values include contact resistance, and the capacitance of TSV is calculated by the modeling [4] with the oxide liner thickness of 0.1 and 0.05μm for TSV-large and TSV-small. The diameter and pitch of our F2F pad is twice the minimum top metal (M8) width and pitch. Comparing the pitch numbers, we see that the F2F bonding provides the highest 3D interconnect density.

TABLE IV

3D INTERCONNECT SETTINGS. TSV PITCH INCLUDES KEEP-OUT ZONE.

name	diameter (μm)	height (μm)	pitch (μm)	R (Ω)	C (fF)
TSV-large	3.0	20	5.016	0.5	10.0
TSV-small	1.0	10	1.672	0.5	3.1
F2F pad	0.448	0.380	0.912	0.1	0.2

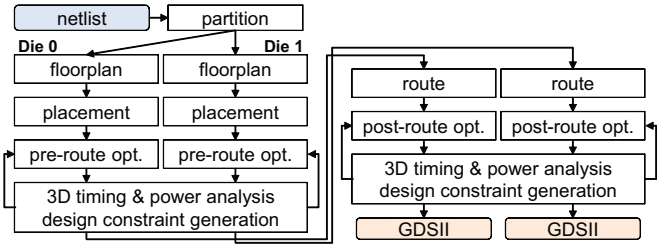

Fig. 5. Our 3D IC design and analysis flow.

B. Design and Analysis Flow

Our 3D IC design and analysis flow is summarized in Fig. 5. To implement the target design on two dies, first we partition the netlist into two subdesigns. The memory blocks and Cores are partitioned manually (see Section III-C), and the standard cells for buses are partitioned by a partitioner. Once the netlist is partitioned, we start the layout. Currently, there is no commercial layout tool that optimizes multiple die designs altogether. Thus, we perform layout steps (floorplan, placement, routing, optimization) die by die. The timing and power analysis is performed for the whole design [5], and further optimizations may be run to iteratively improve design quality. Note that the optimization engine cannot see the entire 3D path, therefore optimization capability is limited [6]. Finally, we obtain the GDSII files of the dies.

C. 3D Partitioning and Floorplan

With 2D-SC design as the 2D baseline, we partition the design into two parts and assign them to dies. We consider three partitioning options:

- Group-Cut (**3D-GC**): The groups are cut into the left and the right half. The left half is assigned to Die 0 and the right half to Die 1. Since only Core-Core buses (between neighboring Cores) cross the cut line, this cut results in the minimum number of 3D connections.
- Group-Fold (**3D-GF**): Each group is folded into 3D. Per each group, a half of the Cores/RFs/ICs/DCs are assigned to Die 0, and the rest to Die 1. All the buses become 3D; the number of 3D connections become large. The wirelengths of all buses become much shorter because of the 3D folding.
- Cluster-Cut (**3D-CC**): The cut is done in Core/RF/IC/DC cluster level. All Cores and ICs are assigned to Die 0, and all RFs and DCs are to Die 1. By placing RFs under Cores (in 3D manner), the Core-RF bus length becomes shorter, which may be helpful considering the bus width.

Using an automatic partitioner, we also partition the netlist of MUXes, considering connections to the blocks. Assuming F2B bonding with TSV-small, we perform floorplan and layout for the aforementioned partition methods. Our partition and floorplan options are shown in Fig. 6.

TABLE V

LAYOUT RESULTS FOR 3D FLOORPLAN OPTIONS.

		3D-GC	3D-GF	3D-CC
footprint (mm^2)		4.10	4.10	4.10
wirelength (m)	all buses	31.61	21.83	34.95
#TSVs		538	22,631	20,810
cells (thousands)	bus-buffer	199	134	198
power (W)	total	1.673	1.386	1.703
	bus-cell	0.058	0.030	0.054
	bus-net	0.533	0.294	0.547
	bus-leakage	0.076	0.045	0.069

TABLE VI

LAYOUT RESULTS WITH VARIED 3D IC TECHNOLOGY SETTINGS.

		TSV-large	TSV-small	F2F
footprint (mm^2)		4.10	4.10	4.10
wirelength (m)	all buses	27.08	21.83	20.62
cells (thousands)	bus-buffer	157	134	119
power (W)	total	1.504	1.386	1.341
	bus-cell	0.038	0.030	0.028
	bus-net	0.389	0.294	0.252
	bus-leakage	0.053	0.045	0.040

Layout results for the 3D floorplan options are summarized in Table V. Although the three options use the same footprint area, significant differences are observed in the wirelength and power of buses. As expected, the number of TSVs is the highest for 3D-GF and the smallest for 3D-GC. The 3D-GF provides the best design in terms of bus wirelength, bus power, and total power. A shorter bus wirelength leads to smaller buffer count and lower bus power. Comparing 3D-GF vs. 2D-SC (in Table II), we observe that footprint, bus wirelength, bus buffer count, and bus power reduces by 50.0%, 40.7%, 34.8%, and 46.0%, respectively. The total power reduces by 17.5%.

IV. IMPACT STUDY WITH 3D DESIGNS

A. Impact of 3D IC Technology Settings

To see the impact on design quality, for the best 3D floorplan option (3D-GF), we apply different 3D IC technology settings: F2B bonding with TSV-large (**TSV-large**) and TSV-small (**TSV-small**), and F2F bonding (**F2F**). Layout results with the varied 3D IC technology settings are summarized in Table VI. Compared with TSV-small, the increased wirelength for TSV-large is due to the increased TSV area overhead. As shown in Fig. 7(b), since TSVs consume significant area, cells/buffers for buses are more spread across the chip, and the nets connecting to the TSVs take more detours. In TSV-large and TSV-small, TSVs occupy 13.9% and 1.5% of the total silicon area. The wirelength and buffer count of buses increase by 24.0% and 17.2%, respectively. As a result, the total power increases by 8.5%, which is significant.

The layout results of F2F are better than those of TSV-small, because F2F pads do not occupy silicon space and can be placed over macros and standard cells. Compared with TSV-small, the wirelength, buffer count, and power of

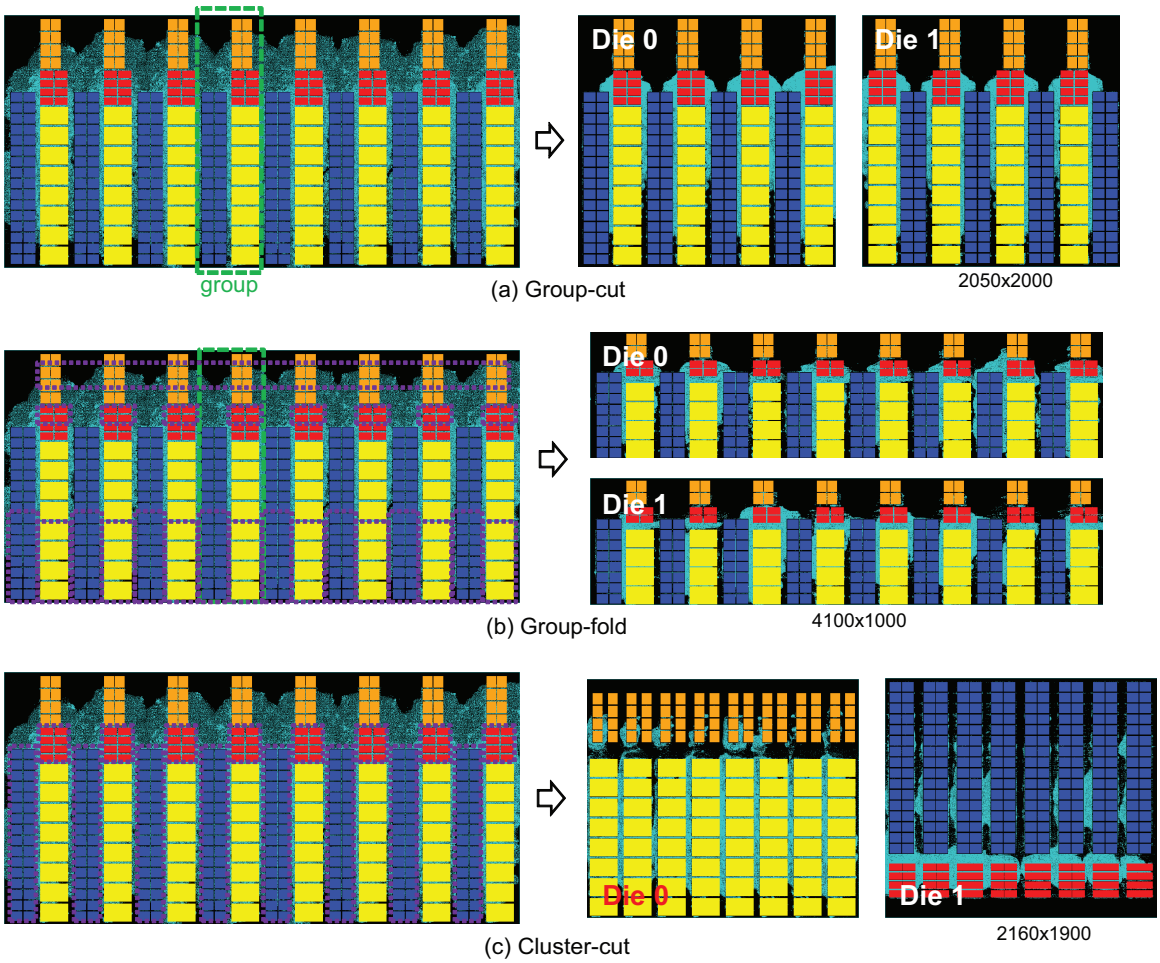

Fig. 6. Partition and floorplan options for 3D. The objects inside the purple dotted boxes are assigned to Die 1, and the rest to Die 0. Numbers represent dimensions in μm.

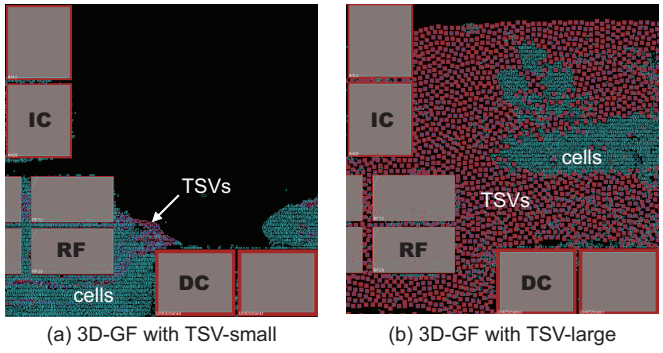

Fig. 7. Layout zoom-in shots for 3D-GF with (a) TSV-small and (b) TSV-large.

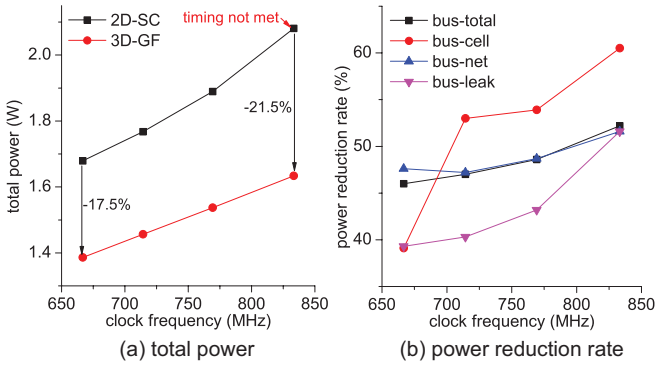

Fig. 8. Power vs. clock frequency for 2D-SC and 3D-GF designs.

buses in F2F reduce by 5.5%, 11.2%, and 13.3%, respectively. The total power reduces by 3.3%. From these results, we observe that 3D IC technology settings affect overall design quality significantly, especially when lots of 3D connections are required.

B. Impact of Clock Frequency

Clock frequency also affects the power benefit amount. As the target clock period becomes shorter, it is harder for 2D to meet the timing than 3D, because bus wires are longer in 2D. The timing critical path is from a RF to a F/F in a Core. For 2D-SC and 3D-GF (with TSV-small), we reduce the clock period from 1.5ns (=667MHz, baseline) down to 1.2ns

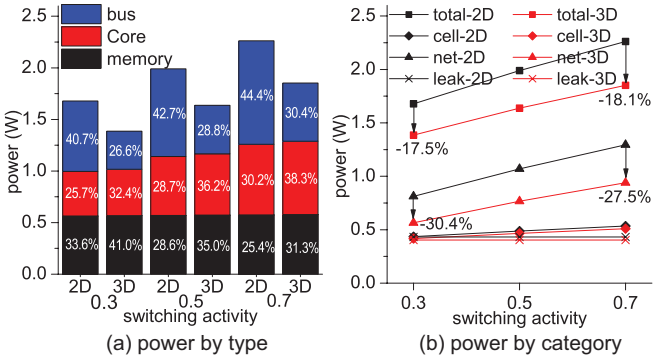

Fig. 9. Power vs. circuit switching activity. The '2D' and '3D' mean 2D-SC and 3D-GF designs. The percentage values in (a) mean the percentage of the power consumed by the type. The percentage values in (b) mean the power reduction rate for 3D.

(=833MHz) to observe the power-delay trade-off for 2D and 3D. For each target clock period, we perform the layout and optimizations to close the timing.

As shown in Fig. 8(a), as clock frequency becomes faster, total power increases faster with 2D. At 833MHz, the layout optimizer failed to close the timing for 2D, and total power reduction is 21.5%, which is larger than the baseline case. More buffers and larger cells would be required for 2D to close the timing, which would increase the power benefit amount. We also show the power reduction rates of buses in Fig. 8(b). Most of the bus power is net power, thus bus total power reduction curve closely follows bus net power reduction curve. Note that bus cell and leakage power also reduces more at faster clock speeds, because larger (and more leaky) cells/buffers are needed to close the timing for 2D.

C. Impact of Circuit Switching Activity

Our power analysis is based on static probability calculations. With the switching activity factor of primary inputs and sequential outputs set to 0.2 and 0.3, the switching activities of cells and nets are propagated throughout the whole netlist. In reality, depending on the GPU workload characteristics, the circuit activity may vary, which changes the power consumptions of Cores, memories, and buses. Thus, for the 2D-SC and 3D-GF designs at baseline clock frequency, we simulated different workload scenarios by changing the switching activity factor of sequential outputs to 0.5 and 0.7 to see the impact on the power benefit of 3D ICs.

As shown in Fig. 9(a), as the switching activity increases, the total power increases, mainly because the power consumptions of buses (and Cores) increase. The power consumption of memory stays about the same. In this scenario, since the wirelength and the power of buses are reduced with 3D, we expect to see larger power reductions with higher switching activities. Indeed, as shown in Fig. 9(b), the total power reduction increases from 17.5% to 18.1% as switching activity increases. Although the net power reduction rate reduces from 30.4% to 27.5%, the amount of net power (and the *percentage* out of total power) increases. Thus, the total power reduction rate increases.

V. Lessons Learned

We first learn that he floorplan majorly affects the design quality in terms of bus wirelength, congestion, and power consumption. And for 3D IC designs, partitioning style and technological setup also affect the design quality much. For successful adoption of 3D ICs in complex digital designs with lots of macros and complex buses, design automation tools that consider various partition and floorplan options with proper technology handling are highly anticipated.

Second, the power benefit of 3D ICs changes with target clock frequency and 3D IC technology setup. For faster clock frequency, the power reduction rate increases. In addition, when the 3D interconnect count is high, TSVs may occupy significant amount of silicon space, which may incur design inefficiency in terms of wirelength, buffer count, power consumptions, etc. This dependency of power benefit amount on the 3D IC technology may suggest how practical the technology is for the target application.

Lastly, our timing/power optimization flow for 3D IC designs (see Section III-B) is not optimal. We perform the timing/power optimization die by die with the design constraints on the die boundary ports (TSVs or F2F pads). The optimization engine cannot see the entire 3D path, therefore optimization capability is limited [6]. In addition, the design constraints on the die boundary ports are in reduced (or simplified) models and cannot accurately represent the actual design context. True 3D design automation tools with multi-die optimization capability are necessary for high performance and low power designs.

VI. Conclusions

In this paper, we demonstrated how the power consumption of buses in GPUs can be reduced with 3D IC technologies. To maximize the power benefit of 3D ICs, it is important to find a good partition and floorplan solution. The power benefit amount also depends on the 3D IC technology setup, target clock frequency, and circuit switching activity. With 3D IC technologies, the total power of our GPU can be reduced by up to 21.5%.

References

[1] B. Black *et al.*, "Die Stacking (3D) Microarchitecture," in *Proc. Annual IEEE/ACM Int. Symp. Microarchitecture*, 2006, pp. 469–479.

[2] U. Kang *et al.*, "8Gb 3D DDR3 DRAM Using Through-Silicon-Via Technology," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2009, pp. 130–131, 131a.

[3] Synopsys, "32/28nm Generic Library." [Online]. Available: http://www.synopsys.com

[4] G. Katti *et al.*, "Electrical Modeling and Characterization of Through Silicon via for Three-Dimensional ICs," *IEEE Trans. on Electron Devices*, vol. 57, no. 1, pp. 256–262, Jan. 2010.

[5] Y.-J. Lee and S. K. Lim, "Timing Analysis and Optimization for 3D Stacked Multi-Core Microprocessors," in *Proc. IEEE Int. 3D Systems Integration Conf.*, 2010, pp. 1–7.

[6] Y.-J. Lee, *et al.*, "Slew-Aware Buffer Insertion for Through-Silicon-Via-Based 3D ICs," in *Proc. IEEE Custom Integrated Circuits Conf.*, 2012, pp. 1–8.