Impact of Die Partitioning on Reliability and Yield of 3D DRAM

Woongrae Kim[1], Dae-Hyun Kim[1], Hee Il Hong[2], Linda Milor[1], and Sung Kyu Lim[1]
[1]School of ECE, Georgia Institute of Technology, Atlanta, GA, USA
[2]Samsung Electronics, Hwasung, Korea

***Abstract*—In this paper we present comparative study on reliability and yield analysis of 3D SDRAM designs built with two practical die partitioning styles, namely, cell/logic-mixed and cell/logic-split. In cell/logic-mixed partitioning, each die contains DRAM cells and peripheral logic components except for the last one that contains I/O logic. In our cell/logic-split style, each die contains DRAM cells and small amount of logic except the bottom die that is all logic including peripheral modules and I/O cells. Our simulation and analysis results provide useful design tradeoffs in terms of area, TSV count, reliability, power, performance, and yield.**

I. Introduction

3D DRAM, where multiple DRAM dies are vertically stacked and connected with through-silicon-vias (TSVs), is believed by many to be the first commercial product that will bring 3D stacking and TSV technologies to the mainstream market. The total memory capacity increases linearly by the number of tiers stacked under the same footprint. It is important to note that 3D DRAM is different from the Hybrid Memory Cube (HMC). In 3D DRAMs, the number of I/O signals is the same as the conventional DRAM such as DDR3. HMCs, on the other hand, communicate through the high parallel wide-I/O interface. Due to HMC's cost, yield, and testing issues with the additional TSVs in the logic die, 3D DRAMs are expected to be commercialized first and thus our focus.

In this paper we compare two practical design styles of 3D DRAM, namely, cell/logic-mixed (see Figure 1(a)) and cell/logic-split (see Figure 1(b)), that differ by how various modules in a DRAM system are partitioned into multiple dies. In our cell/logic-mixed partitioning, each die contains DRAM cells and peripheral logic components except for the last one that contains additional I/O logic. In our cell/logic-split style, each die contains DRAM cells and small amount of logic except the bottom die that is all logic including peripheral modules and I/O cells. Our goal is to compare these two styles in terms of area, power, performance, and yield.

The 3D DRAM from Samsung presented in [1] is based on cell/logic-mixed style. The cell/logic-split style resembles the die partitioning used in HMC. However, there is no existing work that shows how the actual design in the split style is done. In addition, to the best of our knowledge, there exists no comparative study that compares the quality of these two. In this paper we present our optimized cell/logic-split design. Our studies are based on GDSII layouts and sign-off quality timing, power, and reliability analysis.

II. Layouts of Two Partitioning Styles

In cell/logic-mixed partitioning style [1], the four dies are almost identical except for the bottom (= master die) that contains I/O pads and interface circuits (see Figure 2(a)). Each die contains 8Gb DRAM cells, 400 signal TSVs, and 100 P/G TSVs. The TSVs used in this 3D SDRAM are via-last type with 10um diameter and 60um pitch. Our designs are based on 20nm PDK. The data rate of 3D stacked DDR3 SDRAM is 1,600Mbps based on the burst length of 8. We identify

This work is supported by Samsung Electronics.

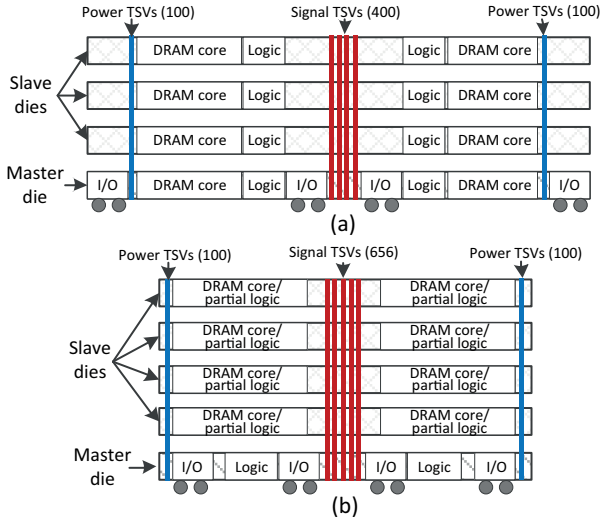

Fig. 1. Conceptual drawing of 3D stacked DDR3 SDRAM (a) 4-tier cell/logic-mixed design [1], (b) our 5-tier cell/logic-split design

two issues with this style. First, the large area of I/O pads and buffers is expected to become more serious with today's 20-30nm DRAM process technology since their size may not scale as DRAM cell technology. Second, the package bumps below I/O pads cause non-trivial reliability problem in DRAM cells. This is mainly caused by the coefficient of thermal expansion (CTE) mismatch among various materials in that area. This leads to a highly compressive stress on dies which contain DRAM cells [2]. Since DRAM cells contain significantly smaller feature sizes and are consequently much more vulnerable to mechanical reliability problems, we separate I/O pads and interface circuits and package bumps from the dies that contain DRAM cells in our cell/logic-split design.

Our cell/logic-split design incorporates 5 tiers of DRAM dies that altogether provide 32Gb of DDR3 memory (see Figure 1(b)). Each die (both slave and master) contains 656 signal TSVs that are located in the middle and 100 power/ground (P/G) TSVs on both the top and bottom. The bottom master die contains peripheral components, I/O pads/circuits, buffers, and serializer/deserializers (see Figure 1(b)). We move most peripheral circuits between Global I/O (GIO) drivers and I/O circuits to the bottom die so that we reduce chip area and reliability impact. We define the peripheral circuits as the *DQ Peripheral Unit* (DQPU). Each DQPU handles the communication between GIO drivers and one I/O pad for DQ. We also have an empty space available for extra logic such as DRAM controllers in the master die. We argue that using a better logic process technology in this logic-only master die, transistors with shorter channel lengths and low Vth can be used to optimize design quality further. Our related experiments show that with high-speed logic and reduced

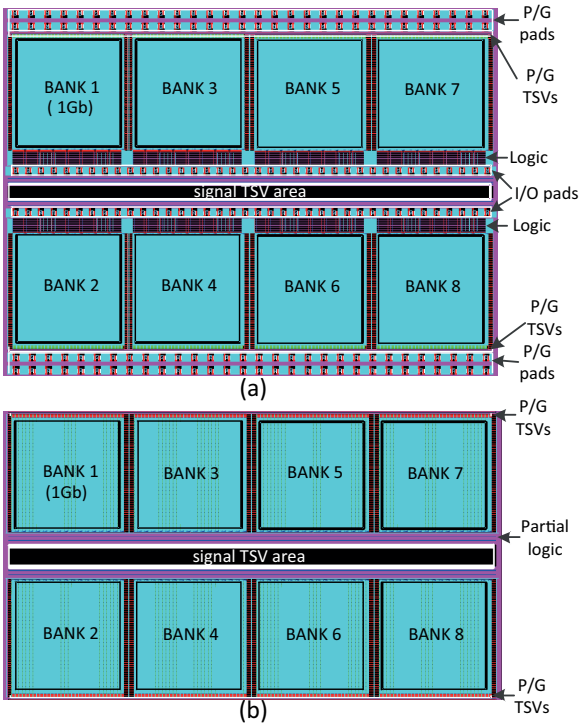

Fig. 2. Full-chip GDSII layouts. (a) master die of cell/logic-mixed design [1], (b) slave die of cell/logic-split design

RC parasitic effects on the data paths, we are able to reduce the size of the peripheral circuits significantly (up to 27%) while using lower supply voltage (1.3V) in the bottom die. The top four slave dies consist of DRAM core (cell arrays, sense-amps, decoders, and equalizers), partial logic, and GIO drivers (see Figure 2(b)). The logic portions are used to drive DRAM cell cores.

III. Design Optimization for TSV Reduction

In cell/logic-split design, all of the DQPUs and I/O pads are located in the bottom die. In this case, each DQ path between a DRAM bank and its DQPU requires a dedicated connection and be distinguished between read and write operations. Thus, 4096 *non-shared* TSVs (= 2 x 8 DQs x 8 burst length x 8 banks x 4 dies) are used in the master die, where 75% of them are "feed-through TSVs" that provide connections between the master and other slave dies. This high TSV usage poses challenges in area and reliability. We propose two solutions to tackle this problem.

- Bank-level DQPU Sharing: we share DQPUs between a pair of active and inactive bank. Note that we use more advanced process technology for the peripheral logic circuits in the master die of our cell/logic-split design. This allows our DQPUs in the master die to drive larger loads. In addition, we add switches in GIO drivers between a DQPU and its two banks so that we can disconnect the loads from the inactive bank and its data lines from the DQPU. Thus, our DQPUs only need to drive the loads from active banks. This bank-level sharing of DQPU also leads to a significant reduction in both DQ TSV and DQPU counts by 2x.
- Die-level DQPU Sharing: we share DQPUs among the DRAM banks in different tiers. In the original cell/logic-split design, each bank is connected to 64 DQPUs in the master die. However, there is only one die that is active during read/write operation. This means we can share a group of 64 DQPU sets among 4 banks in 4 slave dies so that we can disconnect 3 inactive dies using switches in that dies and drive only the one from active die. This leads to 4x saving in both DQPU and DQ TSV usage.

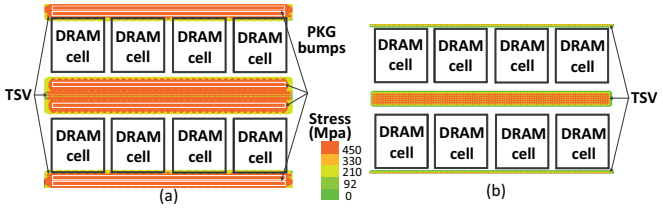

Fig. 3. Full-chip mechanical stress analysis in s11 direction with 20um Keep-Out-Zone. (a) master die of cell/logic-mixed design, (b) slave die of cell/logic-split design

TABLE I
Mechanical and Timing Reliability Comparison

Mechanical stress		
	Cell/logic-mixed	Cell/logic-split
Area over 450Mpa stress	36.8%	4.37%
Maximum stress	1350.4Mpa	688.1Mpa
Mobility variation		
	Cell/logic-mixed	Cell/logic-split
Area over 15% variation	34.8%	5.01%
Maximum variation	55.2%	37.7%

Thus, we reduce the total DQ TSV usage from 4,096 to 512 and DQPU usage from 2,048 to 256 with both solutions combined. This corresponds to 2x worse DQ TSV usage (512 for split design vs 256 for mixed design) and 1.64x worse total signal TSV usage (656 for split design vs 400 for mixed design). In case of DQPU saving, our split design uses 256 DQPUs in the entire 5-tier, whereas the mixed style uses 512 DQPUs in each die. Thus, the total DQPU count is 2048 in the mixed style, which leads to 8x saving with our split style.

IV. Simulation Results

We perform sign-off analysis using HSPICE and Synopsys PrimeTime for timing and power calculations. PrimeTime is built for 2D IC analysis, and we extended it to handle 3D IC as suggested in [3]. We also use the full-chip mechanical stress and mobility variation analysis tools presented in [2].

A. Mechanical Reliability Simulation

Figure 3(a) shows the simulation results of mechanical stress in the S11-direction for cell/logic-mixed design. The significant stress induced by CTE mismatch among package bumps, micro-bumps, and TSVs mostly affect the area nearby the TSV arrays located in the middle, top, and bottom of the die. This stress may cause serious structural damage such as cracks in the substrate and TSVs, delamination of TSV liner, and TSV protrusion [2]. Also, the mechanical stress decreases electron mobility of DRAM cell transistors near the top and bottom edges [4]. The variation of electron mobility introduces undesirable timing variations and may lead to read/write failures.

In Table I we show the comparison of mechanical reliability and mobility variation between cell/logic-mixed vs cell/logic-split design styles. We focus on the area with more than 450MPa mechanical stress and 15% mobility variation. We observe that our cell/logic-split design shows significantly lower mechanical stress and mobility variation impact (see Figure 3(b)). Since there are no package bumps under the substrate that contains DRAM cells in cell/logic-split

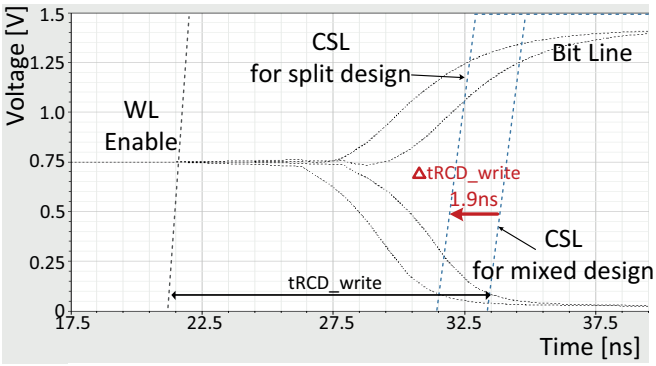

Fig. 4. HSPICE simulation comparison for write operation ($tRCD_{\mathrm{write}}$)

TABLE II
POWER ANALYSIS FOR DQ DATAPATH ELEMENTS

Write operation			
	Cell/logic-mixed	Cell/logic-split	Cell/logic-split
Slave/Master	1.5V/1.5V	1.5V/1.5V	1.5V/1.3V
8 DQPUs	10.58 mW	6.87 mW	5.22 mW
1 I/O SERDES	12.68 mW	10.88 mW	8.81 mW
8 GIO drivers	13.37 mW	13.67 mW	13.94 mW
Total	36.63 mW	31.42 mW	27.97 mW
Read operation			
	Cell/logic-mixed	Cell/logic-split	Cell/logic-split
Slave/Master	1.5V/1.5V	1.5V/1.5V	1.5V/1.3V
8 DQPUs	12.9 mW	7.14 mW	5.36 mW
1 I/O SERDES	12.2 mW	10.8 mW	8.39 mW
8 GIO drivers	14.6 mW	15.1 mW	15.1 mW
Total	39.70 mW	33.04 mW	28.85 mW

design, mechanical stress is only due to TSVs and micro bumps. This significantly alleviates mechanical stress and electron mobility variation compared with cell/logic-mixed design.

B. Performance and Power Simulation

Figure 4 shows HSPICE simulation results for both mixed and split designs under write operation. Using advanced logic process in the master die of split design, DQPUs in that die can be designed with transistors with shorter channel lengths and low Vth. This helps the DQPD units better handle the load even with bank-level and die-level DQPU sharing schemes described in Section III. All of these benefits lead to $tRCD_{\mathrm{write}}$ reduction of 1.9ns (15.6%) for our split design as shown in Figure 4.

Using a more advanced process technology in the master die of split design, we reduce the size of logic devices (up to 27%) and operate at a lower Vdd (= 1.3V). Our power analysis summarized in Table II shows that our device scaling and low supply voltage (1.3V) together improve the power consumption of DQPUs and I/O circuits for both read and write operations. This saving leads to the total power consumption reduction of 23.6% for write operation and 27.3% for read operation in our split design at 1.3V Vdd.

C. Yield and Cost Analysis

We use the Poisson yield model shown in Equation (1) in our yield analysis. We assume that the defect size distribution ($f_{\mathrm{i}}(r)$) and the defect density (D_{i}) are the same in each wafer [5]. $A_{\mathrm{i}}(r)$ is the critical area that is vulnerable to a short from a defect of radius r. Here, if there are different design rules for space, because typically $f_{\mathrm{i}}(r) = k/r^3$, the narrow spaces will dominate the calculation of $\int_0^\infty \mathrm{A_i(r)f_i(r)}\,\mathrm{d}r$, unless there is much more area with the wide spaces, or if the narrow space part has enough redundancy to tolerate defect.

TABLE III
COMPARISON OF AREA AND # OF MANUFACTURED CHIPS

	Mixed design	Split design
Area of each die	$42.99mm^2$	$33.10mm^2$
DRAM core area in a die	$23.52mm^2$	$23.52mm^2$
# of manufactured chips (12" wafer)	1,064	1,342

$$Y_{\mathrm{random}} = \exp(-D_{\mathrm{i}} \int_0^\infty \mathrm{A_i(r)f_i(r)}\,\mathrm{d}r) \quad (1)$$

In 3D SDRAM, the DRAM core area shown in Figure 1 has a significantly smaller feature size and is consequently much more vulnerable to defects. On the other hand, peripheral circuit parts with a larger feature size are much more tolerant to the defects. Since the DRAM core area is the same in each design style (see Table III), we can approximately make the assumption that $\int_0^\infty \mathrm{A_i(r)f_i(r)}\,\mathrm{d}r$ and yield are the same in each case [5].

Note that the total profit depends on the yield, the number of manufactured chips, bonding and mask costs, and cost of additional logic die [5]. Table III shows that the smaller footprint of split design leads an increase in the number of chips that can be manufactured per wafer. For a set of N wafers, the mixed design produces $1064NY$ good chips. Since each product requires 4 good chips, mixed design produces $266NY$ products. On the other hand, split design requires five chips, of which four will have the DRAM core. Hence, the same N wafers will produce $268.4N(4Y+1)$ good chips and $268.4NY$ good products. Hence, split design produces on average $2.4Y$ more good product per wafer. On the other hand, because the yield is higher for the master die for the split design, it becomes possible to allocate more wafers to the slave die. Specifically, for every M master die, we need $4M/Y$ slave die. Then, N wafers produce $1342N/(1+\frac{4}{Y})$ good products with the split design. Hence, as the yield drops below 100%, the number of good products produced per wafer increases.

V. CONCLUSIONS

In this paper we studied benefits of different partitioning styles in 3D SDRAM in terms of reliability, power, area, TSV count, performance, yield, and cost. The cell/logic split partitioning style outperforms or shows comparable results to the cell/logic-mixed style in area, reliability, power, and performance. On the other hand, the cell/logic-mixed style shows less TSV count and lower bonding costs. Our yield and cost analysis provides design guidelines on how to best optimize cell/logic partitioning to enhance profit.

REFERENCES

[1] U. Kang *et al.*, "8Gb 3D DDR3 DRAM Using Through-Silicon-Via Technology," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 1, pp. 111–119, 2010.

[2] M. Jung, D. Pan, and S. K. Lim, "Chip/Package Co-Analysis of Thermo-Mechanical Stress and Reliability in TSV-based 3D ICs," in *Proc. ACM Design Automation Conf.*, 2012, pp. 317–326.

[3] D. H. Kim *et al.*, "3D-MAPS: 3D Massively Parallel Processor with Stacked Memory," in *IEEE International Solid-State Circuits Conference*, 2012, pp. 531–538.

[4] J.-S. Yang, K. Athikulwongse, Y.-J. Lee, S. K. Lim, and D. Pan, "TSV Stress Aware Timing Analysis With Applications To 3D-IC Layout Optimization," in *Proc. ACM Design Automation Conf.*, 2010, pp. 803–806.

[5] L. Milor, "A Survey of Yield Modeling and Yield Enhancement Methods," *IEEE Trans. on Semiconductor Manufacturing*, vol. 26, no. 2, pp. 196–213, 2013.