INVITED: RTL-to-GDS Tool Flow and Design-for-Test Solutions for Monolithic 3D ICs

Heechun Park[1], Kyungwook Chang[1], Bon Woong Ku[1], Jinwoo Kim[1], Edward Lee[1], Daehyun Kim[1], Arjun Chaudhuri[2], Sanmitra Banerjee[2], Saibal Mukhopadhyay[1], Krishnendu Chakrabarty[2], and Sung Kyu Lim[1]

[1] Department of Electrical and Computer Engineering, Georgia Institute of Technology

[2] Department of Electrical and Computer Engineering, Duke University

ABSTRACT

Monolithic 3D IC overcomes the limitation of the existing through-silicon-via (TSV) based 3D IC by providing denser vertical connections with nano-scale inter-layer vias (ILVs). In this paper, we demonstrate a thorough RTL-to-GDS design flow for monolithic 3D IC, which is based on commercial 2D place-and-route (P&R) tools and clever ways to extend them to handle 3D IC designs and simulations. We also provide a low-cost built-in-self-test (BIST) method to detect various faults that can occur on ILVs. Lastly, we present a resistive random access memory (ReRAM) compiler that generates memory modules that are to be integrated in monolithic 3D ICs.

1 INTRODUCTION

Monolithic 3D ICs (M3D) [1] are introduced to maximize the benefit of 3D ICs with their nanoscale inter-layer-vias (ILVs) as inter-die connection. M3D offers massive inter-die connections that are not possible with TSVs that suffer from significant area and electrical coupling overhead. A key technological advance required to fully benefit from M3D is design tools. Unfortunately, no commercial vendor offers them in the market as of today.

Recently it has been demonstrated that the high integration density in M3D makes ILVs vulnerable to various defects [2]. Therefore, ILV testing is necessary for effective defect screening and quality assurance. Although ILVs can be tested together with logic and memory, we still need design-for-test (DfT) method for ILVs exclusively for defect isolation and yield learning.

Resistive Random Access Memory (ReRAM), which enables high-bandwidth logic-memory integration, has emerged as an attractive candidate for on-chip non-volatile memory (NVM) technology. There have been significant efforts in developing ReRAM device technologies as well as ReRAM macros [3, 4]. However, there is a lack of EDA methods and tools for automatic ReRAM macro generation.

In this paper, we first discuss a RTL-to-GDS design flow for monolithic 3D ICs. This flow fully benefits from the capabilities of commercial 2D IC tools that have been developed and enhanced during the last 50 years. Yet, they provide clever ways to extend the features to handle multiple tiers of active devices and interconnects in M3D. We also discuss a DfT method that detects various ILV faults at low resource overhead. Lastly, we present a ReRAM module generator that is proven to be useful in design space exploration (DSE) of M3D architectures. For simplicity, we focus on 2-tier M3D in this paper.

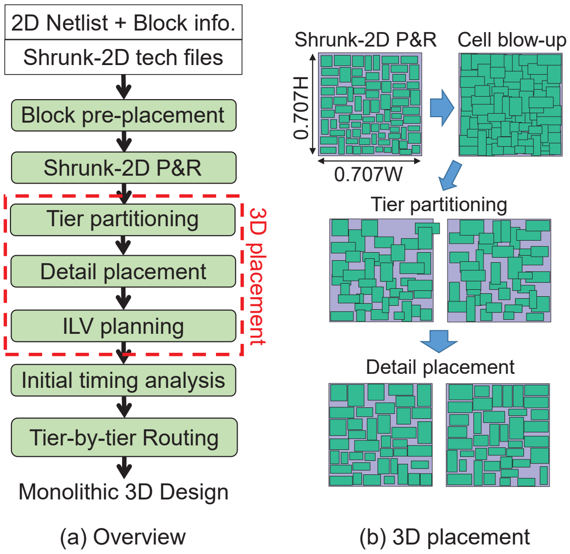

Figure 1: (a) Shrunk-2D flow [5], (b) illustration of Shrunk-2D flow.

2 MONOLITHIC 3D PHYSICAL DESIGN

2.1 Leveraging Commercial 2D IC Tools

Shrunk-2D flow introduced in 2014 [5] is the first commercial-quality RTL-to-GDS flow for monolithic 3D ICs that has inspired a few follow-up works [6, 7]. Figure 1(a) shows an overview. It starts from a 2D synthesized netlist, and macro block technology files if the design contains pre-defined macro blocks (e.g. memory modules). For initial pseudo-3D design named as *shrunk-2D* design, we need to prepare the scaled library exchange format (LEF) file, that standard cell and metal width/height are scaled down to 0.707 for a 2-tier 3D design.

In the floorplan stage, each macro block is pre-placed to each corresponding tier. We set a partial placement blockage on the area

DAC '19, June 2–6, 2019, Las Vegas, NV, USA

ACM ISBN 978-1-4503-6725-7/19/06…$15.00
https://doi.org/10.1145/3316781.3323486

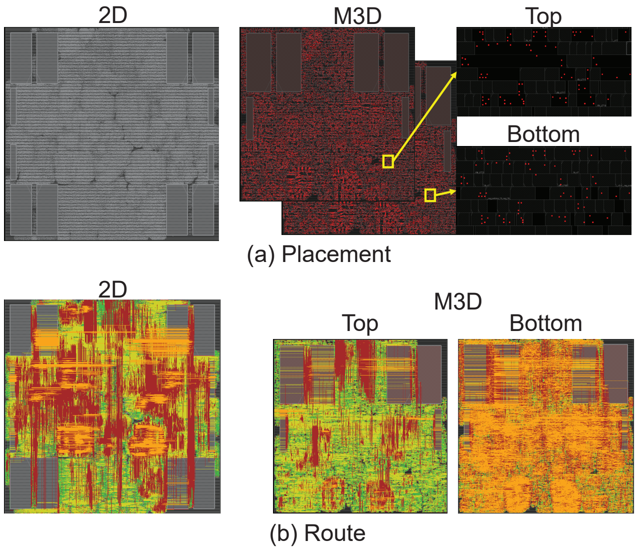

Figure 2: GDS layout of 2D and M3D designs of Rocket2 benchmark

of placed macro blocks, to enable the cell placement on the other tier with the same x-y area where the block is placed. With the generated floorplan and shrunk technology files, we use a commercial 2D placement and route (P&R) tool to generate a shrunk-2D P&R result, where all cells are shrunk into half size and placed into a 2D die with 50% footprint, which is the planar footprint of the final 3D design.

With the shrunk-2D P&R results, 3D placement is performed with in-house tools, as in Figure 1(b). Firstly, standard cells in the design are blown up to their original sizes. We can consider this status as a commercial-tool-based 3D P&R result, in which all cells are projected to a single tier. Then, these cells are separated into two different tiers with bin-based Fiduccia-Mattheyses (FM) min-cut partitioning[8], which divides the footprint into square bins and performs FM min-cut algorithm on each bin. Remaining cell overlaps are legalized in the tier-by-tier detail placement stage, which uses the refinement function of commercial 2D P&R tool. Lastly, we use a modified LEF files to import both tiers vertically in commercial 2D P&R tool , and perform routing to find efficient spots for ILVs. Then, we are able to generate placement information with ILV locations in separate verilog and design exchange format (DEF) files for each tier.

After 3D placement, the monolithic 3D design is finalized by tier-by-tier routing. We first perform initial timing analysis to generate design constraints file for each tier, and apply tier-by-tier routing to each tier with commercial 2D P&R tool.

2.2 Design Comparison

We compare monolithic 3D vs. 2D IC designs in terms of power-performance-area (PPA) metrics using three design benchmarks: RISC-V Rocketcore design with 2 cores (Rocket2), AES-128, and AVC-Nova. TSMC 28*nm* PDK is used for the entire experiments. Netlists are generated from synthesis of benchmark RTLs using

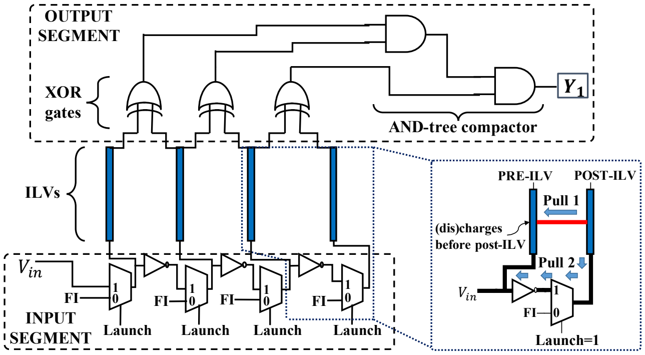

Figure 3: Illustration of the BIST architecture.

Synopsys Design Compiler, P&R is performed with Cadence Innovus Implementation System, and final analysis is performed by Cadence Tempus Timing Signoff Solution.

Figure 2 shows the layout of the Rocket2 in both designs. In monolithic 3D design, all external I/O ports are connected with the top tier. Red dotted squares are the location of ILVs, which has the size of 100*nm* × 100*nm*. Each ILV connects the bottom-most metal layer of top tier and the topmost metal layer of bottom tier, thereby overlap between ILV and cells is not allowed in the top tier.

Table 1 summarized design comparison between 2D and monolithic 3D designs. Monolithic 3D design has 50% smaller footprint and 14%~20% of wirelength reduction compared to 2D design, with the aid of vertical connections through ILVs. Shorter wirelength also leads to smaller net switching power, which results total power reduction. Monolithic 3D design flow is compelled to perform tier-by-tier final routing separately, which means it's impossible to perform accurate timing optimization throughout the full design. Nevertheless, their worst negative slacks are in the similar range, which is less than 10% of the target clock period.

3 DESIGN-FOR-TEST SOLUTIONS

Typical fault models for an ILV are shorts, opens, and stuck-at faults (SAFs). Shorts occur due to particle contamination, metal diffusion, etc. Opens occur when an ILV fails to land on a contact pad leading to a very high ILV resistance.

Methods such as [9] use dedicated scan elements (test points) for testing ILVs. These solutions require large test time and area since the number of test patterns and test points are directly proportional to the ILV count [10]. ATPG-based interconnect test methods [11] are likely to be less effective for ILV testing because I/O pins are available only on one layer in an monolithic 3D IC; the activated ILV faults must be propagated through multiple tiers and ILVs, adding to the propagation constraints for ATPG and impeding fault isolation for yield learning.

3.1 BIST Architecture

The proposed BIST architecture for testing shorts, opens, and SAFs in ILVs is shown in Figure 3. On the output side of the ILVs, we insert 2-input XOR gates between adjacent ILVs. There are $(N-1)$ XOR gates for a set of N ILVs. The XOR outputs feed a space compactor which is an optimally balanced AND tree with $(N-1)$ inputs and a

Table 1: Design comparison of 2D and Monolithic 3D (M3D) P&R results

	Rocket2, 800MHz			AVC-Nova, 730MHz			AES-128, 3.44GHz		
	2D	M3D	Δ(%)	2D	M3D	Δ(%)	2D	M3D	Δ(%)
Footprint (mm^2)	0.746	0.373	-50.01	0.286	0.142	-50.21	0.210	0.105	-50.2
Std. cell count	279983	278622	-0.49	132725	132299	-0.32	109878	109634	-0.22
Std. cell area (mm^2)	0.363	0.362	-0.30	0.190	0.189	-0.79	0.124	0.121	-2.71
Wirelength (mm)	4712.46	4045.65	**-14.15**	2336.97	1954.92	**-16.35**	1494.13	1196.88	**-19.89**
ILV count	0	111395		0	56989		0	44782	
Total Pwr. (mW)	371.90	366.25	**-1.52**	122.50	120.31	**-1.79**	269.59	258.96	**-3.95**
Net Switching Pwr. (mW)	82.70	78.00	-5.68	23.55	22.32	-5.20	67.65	61.57	-8.99
Cell Internal Pwr. (mW)	224.84	224.26	-0.26	63.46	63.20	-0.41	174.66	171.86	-1.60
Leakage Pwr. (mW)	64.36	63.99	-0.58	35.50	34.79	-1.99	27.29	25.52	-6.48
Worst Neg. Slack(ps)	-54.27	-49.47		-60.48	-135.57		-14.88	-26.87	

1-bit output signature Y_1. We can determine whether there is a fault in the given set of ILVs by observing Y_1. Test data is fed to the ILV inputs from an input source V_{in}, which provides complementary signals to adjacent ILVs in the test mode via an inverter chain. A 2:1 multiplexer is present at ILV's input to switch between functional input (*FI*) and test mode based on the *Launch* signal.

The ILVs are tested in two clock cycles by switching V_{in}. The test patterns to the ILVs are "101..." in the first cycle and "010..." in the second cycle. It can be shown that a set of ILVs contains no hard faults if and only if Y_1 is 1 in both clock cycles. The manner in which the ILVs are driven in the test mode leads to a deterministic hard-short behavior; this is illustrated in the inset of Figure 3. If two ILVs are shorted, the ILV appearing first (pre-ILV) in the path of the incoming test signal from V_{in} will drive the other ILV (post-ILV) because of the lower resistance path through the hard short.

3.2 Dual-BIST Architecture

The BIST design of Section 3.2 is also susceptible to SAFs, which may mask ILV fault(s). To minimize the masking probability, a parallel propagation path is added from the ILV outputs to a second 1-bit signature Y_2. The physical structure of this parallel path to Y_2 (BIST-B) is identical to that of the path from the XOR inputs to Y_1 (BIST-A). The XOR and AND gates in BIST-B are replaced with their logical duals in BIST-A. With this "dual-BIST" architecture, it can be shown that ILV fault(s) cannot be masked by a single BIST fault. Moreover, the masking probability due to multiple BIST faults is negligible.

3.3 Overhead for BIST

We evaluated the impact of BIST on the PPA metrics of four monolithic 3D benchmarks (Table 2). For clear comparison of BIST impact, we applied the same design flow of Section 2 with block-level approach, to manipulate the number of ILVs used in the benchmarks. The dual BIST is inserted tier-wise to generate 3D BISTed designs.

Table 3 presents the power consumption and area overheads of the BISTed designs (BI) relative to the non-BISTed designs (N-BI). The impact of BIST on the circuit's PPA is minimal.

The dual-BIST solution requires less area compared to a baseline DfT scheme where scan flops are added at both ends of an ILV for controllability and observability; Table 4 shows the comparison.

Table 2: M3D benchmarks used for PPA comparison.

Name	Footprint (μm×μm)	Cell count	ILV count	max freq
Rocket2	700×700	301,219	1,200	350MHz
AVC-Nova	420×420	134,547	317	366MHz
AES-128 (I)	350×350	102,278	425	2273MHz
AES-128 (II)	350×350	90,233	426	1250MHz

Table 3: Impact of BIST on PPA.

Circuit	Metric	N-BI	BI	Overhead (%)
Rocket2	Cell area (μm^2)	382037.6	389785.6	2.03
	Power (mW)	227.3607	232.5544	2.28
AVC-Nova	Cell area (μm^2)	196841.5	199038.7	1.12
	Power (mW)	87.46587	90.79729	3.81
AES-128 (I)	Cell area (μm^2)	132049.9	134771.2	2.06
	Power (mW)	178.5411	194.3433	8.85
AES-128 (II)	Cell area (μm^2)	103632.4	106324	2.6
	Power (mW)	110.96	119.42	7.63

Table 4: Area comparison of baseline DfT and dual-BIST.

Benchmark	Baseline DfT (μm^2)	Dual-BIST (μm^2)	Reduction (%)
Rocket2	12288	7748	36.9
AVC-Nova	3246.1	2197.2	32.3
AES-128 (I)	4352	2721.3	37.5
AES-128 (II)	4362.2	2691.6	38.3

4 RERAM MODULE GENERATOR

Resistive Random Access Memories(ReRAM) have emerged as attractive candidates for on-chip non volatile memory (NVM) technology. A ReRAM device encodes bits as resistance of the device, namely, a high-resistance (HRS) and a low-resistance (LRS). The application of voltage above a threshold results in a change of the resistance from HRS to LRS (SET) or LRS to HRS (RESET). There have been significant effort in developing ReRAM device technologies as well as custom design of ReRAM macros[3, 4, 12]. Likewise, analysis tools for ReRAM architecture and their impact on processor design have also been developed [13, 13]. However, there is a lack of EDA methods and tools for physical design of ReRAM sub-arrays, and generation of large ReRAM macros.

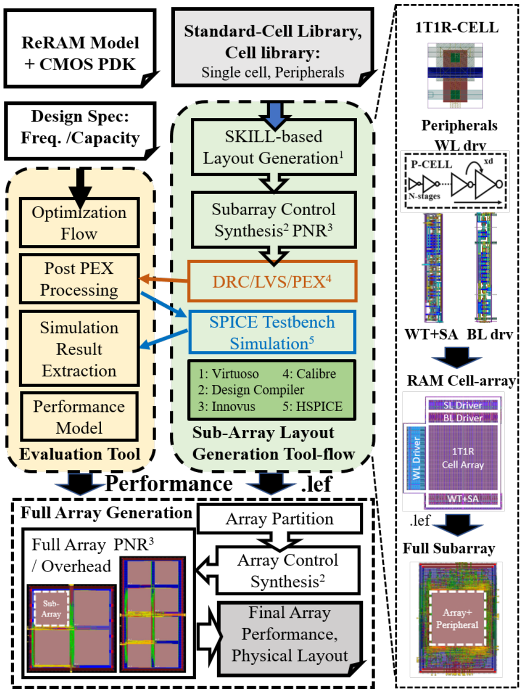

Figure 4: Top level view of tool architecture and sample layouts.

4.1 Generator Architecture

A complete ReRAM module generator will need to automatically generate physical layouts of ReRAM macros by partitioning them into optimized ReRAM sub-arrays of different dimensions (i.e. $N_{row} \times N_{column}$), and accurately evaluate their performance/power characteristics to support fast design space exploration.

We present a EDA methodology for automated generation, design space exploration, and optimization of ReRAM memory arrays. The proposed approach successfully combines circuit-level design of cells and peripherals with automated layout of custom circuits to enable physical design of ReRAM sub-arrays. The auto-generated ReRAM sub-arrays are coupled with auto-generated (logic synthesis and P&R) connectivity between sub-arrays to design a large array. The generation tool is coupled with an evaluation tool for fast circuit level and layout accurate ReRAM performance/power analysis to enable design space exploration and optimization. The tool structure is as shown in Figure 4.

4.2 Array Generation Results

The developed ReRAM module generator is demonstrated for 1T-1R ReRAM cells considering 65nm CMOS technology and open-source ReRAM models [14]. The experimental results show that our tool can generate physical design of ReRAM sub-arrays for different cell topology and sub-array dimensions including read/write circuits. The proposed tool allows us to optimize the sub-array configurations (i.e. number or rows versus columns) for different target sub-array sizes, as well as select optimal partitions for large ReRAM arrays. Table 5 summarizes performance and area for generated arrays of different 8KB and 32KB capacity.

Table 5: Design Results for ReRAM module generator

Sub-Array	Total Capacity 8kB			
	Area (mm^2)	R.E*(pJ)	W.E*(pJ)	Top.E.**(pJ)
64x64	0.605	22.10	36.44	19.13
128x128	0.279	11.25	38.40	5.30
256x256	0.125	13.19	65.91	1.31
	Total Capacity 32kB			
64x64	2.421	64.14	78.48	61.17
128x128	1.073	30.08	57.24	24.14
256x256	0.530	19.01	71.73	7.13

5 CONCLUSIONS

In this paper, we presented a RTL-to-GDS tool flow, a low-cost BIST solution for ILVs, and an ReRAM module generator, all of which are essential in building high-quality and practical designs for monolithic 3D IC (M3D) technologies. Further research is also required to improve these tools and/or overcome their limitations to make M3D an indefensible option for future applications in cloud and edge computing.

6 ACKNOWLEDGEMENT

This research is funded by the DARPA ERI 3DSOC Program under Award HR001118C0096.

REFERENCES

[1] S. Wong et al., "Monolithic 3D Integrated circuits," in *VLSI-TSA*, 2007.
[2] A. Koneru *et al.*, "A Design-for-Test Solution Based on Dedicated Test Layers and Test Scheduling for Monolithic 3D Integrated Circuits," in *TCAD*, 2018.
[3] C.-C. Chou *et al.*, "An N40 256K×44 Embedded RRAM Macro with SL-precharge SA and Low-voltage Current Limiter to Improve Read and Write Performance," in *ISSCC*, 2018.
[4] M. F. Chang *et al.*, "19.4 Embedded 1Mb ReRAM in 28nm CMOS with 0.27-to-1V Read using Swing-Sample-and-Couple Sense Amplifier and Self-Boost-Write-Termination Scheme," in *ISSCC*, 2014.
[5] S. Panth *et al.*, "Design and CAD Methodologies for Low Power Gate-level Monolithic 3D ICs," in *IEEE Int. Symp. on Low Power Electronics and Design*, 2014.
[6] K. Chang *et al.*, "Cascade2D: A Design-Aware Partitioning Approach to Monolithic 3D IC with 2D Commercial Tools," in *IEEE Int. Conf. on Computer-Aided Design*, 2016.
[7] B. W. Ku, K. Chang, and S. K. Lim, "Compact-2D: A Physical Design Methodology to Build Commercial-Quality Face-to-Face-Bonded 3D ICs," in *IEEE Int. Symp. on Physical Design*, 2018.
[8] C. M. Fiduccia and R. M. Mattheyses, "A Linear-Time Heuristic for Improving Network Partitions," in *DAC*, 1982.
[9] R. Pendurkar *et al.*, "Switching Activity Generation with Automated BIST Synthesis for Performance Testing of Interconnects," in *TCAD*, 2001.
[10] A. Jutman, "Shift Register Based TPG for At-Speed Interconnect BIST," in *ICM*, 2004.
[11] D. Erb *et al.*, "Multi-Cycle Circuit Parameter Independent ATPG for Interconnect Open Defects," in *VTS*, 2015.
[12] S.-S. Sheu *et al.*, "A 4Mb embedded SLC resistive-RAM macro with 7.2 ns read-write random-access time and 160ns MLC-access capability," in *ISSCC*, 2011.
[13] X. Dong *et al.*, "NVsim: A Circuit-level Performance, Energy, and Area Model for Emerging Nonvolatile Memory," *TCAD*, 2012.
[14] H. Li *et al.*, "Variation-Aware, Reliability-Emphasized Design and Optimization of RRAM using SPICE Model," in *DATE*, 2015.