Advances in Design and Test of Monolithic 3-D ICs

Arjun Chaudhuri and Sanmitra Banerjee
Duke University

Krishnendu Chakrabarty
Duke University

Heechun Park, Jinwoo Kim, Gauthaman Murali, Edward Lee, Daehyun Kim, Sung Kyu Lim, and Saibal Mukhopadhyay
Georgia Institute of Technology

Editor's note:
Monolithic 3-D (M3D) technology enables unprecedented degrees of integration on a single chip. The miniscule monolithic intertier vias (MIVs) in M3D are the key behind higher transistor density and more flexibility in designing circuits compared to conventional through silicon via (TSV)-based architectures. This article presents a comprehensive design and test techniques for emerging M3D-enabled circuits and systems.

—Partha Pratim Pande, Washington State University

The emergence of monolithic 3-D (M3D) integrated chips (ICs) [1] has enabled the high-density vertical integration of heterogeneous dies via nanoscale monolithic intertier vias (MIVs). A key prerequisite for fully benefitting from M3D is the provisioning of electronic design automation (EDA) tools. However, such tools are not yet commercially available. In the process of aggressive scaling of MIVs for high-density integration, MIVs have become increasingly susceptible to process variations and defects [2], [3]. As MIVs form the electrical links between adjacent dies, monitoring MIV health is crucial for guaranteeing error-free operation, quality assurance, and yield ramp-up. Resistive random access memory (ReRAM) has emerged as a promising candidate for low-power nonvolatile memory (NVM) technology, enabling high-bandwidth logic-memory integration. Despite the development of novel ReRAM devices and circuits [4], [5], there is a need for EDA tools for automated ReRAM macrogeneration.

In this article, we present advances in the integration of M3D design and test flows covered in [6].

Digital Object Identifier 10.1109/MDAT.2020.2988657
Date of publication: 20 April 2020; date of current version: 1 September 2020.

M3D physical design

Design flow with pseudo-3-D approach

The Shrunk-2-D-based design flow introduced in 2014 [7] is the first commercial-quality RTL-to-graphic design system (GDS) flow for M3D ICs that has inspired few follow-up work [8], [9]. It introduced an initial pseudo-3-D design named the *Shrunk-2-D* design, which has all the geometric elements (i.e., cell width and height, wire width, and spacing) scaled down to $1/\sqrt{n}$ of the original dimensions for an n-tier 3-D design. When all standard cells in the Shrunk-2-D design are popped up to their original sizes after Shrunk-2-D place and route (P&R), we consider this design as a commercial-tool-based 3-D P&R result with all cells projected to a single tier.

Figure 1a shows an overview of the full M3D design flow combining M3D synthesis, P&R, insertion of test architecture, and integration of ReRAM. It starts from a 2-D synthesized netlist, and macroblock technology files if the design contains predefined macroblocks (e.g., memory modules). We need to

2168-2356/20©2020 IEEE Copublished by the IEEE CEDA, IEEE CASS, IEEE SSCS, and TTTC

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on October 24,2020 at 10:23:51 UTC from IEEE Xplore. Restrictions apply.

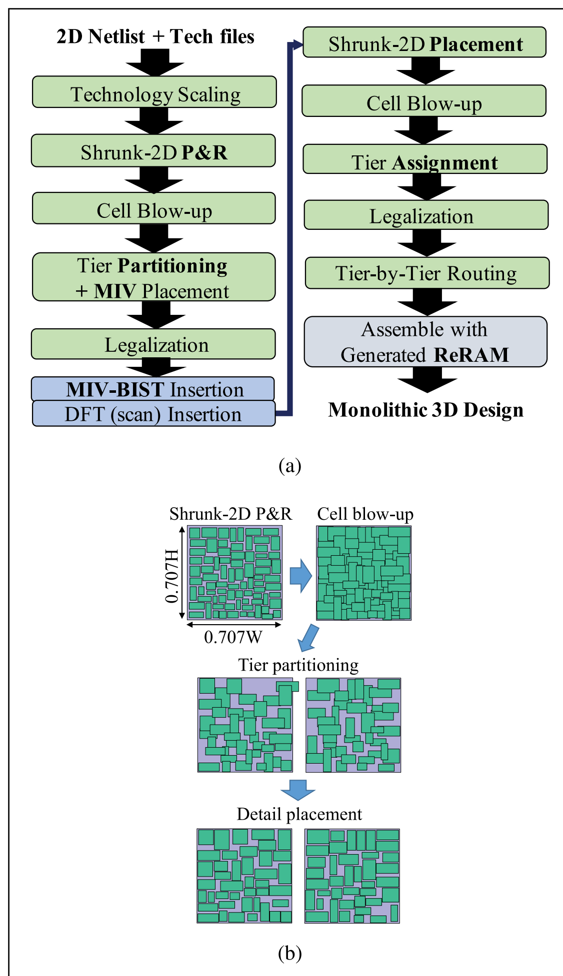

Figure 1. (a) RTL-to-GDS design flow of test-inserted M3D IC. (b) Illustration of 3-D placement in Shrunk-2-D flow.

prepare for an additional scaled library exchange format (LEF) file, so that the standard cell and metal width/height are scaled down to $0.707(1/\sqrt{2})$ for a two-tier 3-D design. If there exists any macroblocks, their LEF files are also modified to have the smallest block size (one "site") with block I/O pins remaining at the same positions, which are outside the minimized block boundary.

In the first block preplacement stage, each macroblock is preassigned and placed in its corresponding tier, and we lay a partial placement blockage on the area of its original size, to restrict the standard cell placement to 50% density in this area. As all standard cells within this area will be assigned to the same tier (different from the macroblock), at least half of this area should be left empty to guarantee the accommodation of these standard cells when they are returned to their original sizes. Using the generated floorplan and shrunk technology files, we use a commercial 2-D P&R tool to perform Shrunk-2-D P&R.

With the intermediate Shrunk-2-D results, standard cells are blown up to their original sizes and 3-D placement is performed with in-house tools as shown in Figure 1b. First, standard cells in the Shrunk-2-D design are separated into two different tiers with bin-based Fiduccia–Mattheyses (FM) min-cut partitioning [10], which divides the footprint into square bins and performs the FM min-cut algorithm on each bin. They are legalized in the next tier-by-tier detail placement stage, which uses the refinement function of the commercial 2-D P&R tool. Next, we stack up both tiers vertically with a modified LEF file and perform routing of cross-tier nets with a commercial router to find efficient spots for MIVs.

On completion of MIV planning and placement, BIST is inserted tier-wise on the inputs and outputs of MIVs, followed by the insertion of scan chains in both tiers. After test insertion, Shrunk-2-D placement is rerun to optimize the cell placement in the test-inserted tiers. Then, tier assignment is performed, which is a minor version of tier partitioning that every cells already have their own assigned tiers. It uses the netlist generation function of the tier-partitioning tool. Finally, tier-by-tier routing with the commercial 2-D P&R tool is performed. P&R of ReRAM modules is then executed in the respective tiers to finalize the M3D design.

For an n ($n > 3$)-tier M3D IC, the same concepts as presented in this article can be utilized, with some changes with respect to the details. Technology should be down-scaled by $1/\sqrt{n}$ instead of $0.707(1/\sqrt{2})$, and there must be a suitable tier-partitioning algorithm to partition the intermediate Shrunk-2-D design (it is still possible to apply the bin-based FM min-cut algorithm multiple times to have a partitioning result). Other stages will operate as is, with the only change of the tier count from 2 to n.

Design comparison

We compare M3D design with 2-D IC counterparts in terms of power–performance–area (PPA) metrics using three design benchmarks: RISC-V Rocketcore design with two cores (Rocket2), AES-128, and AVC-Nova. The commercial 28-nm physical design kit (PDK) is used for the entire experiment. We used the Synopsys Design Compiler for RTL synthesis, the Cadence Innovus Implementation System for P&R

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on October 24,2020 at 10:23:51 UTC from IEEE Xplore. Restrictions apply.

along with our in-house binaries for 3-D placement, and Cadence Tempus Timing Signoff for PPA data generation. In M3D design, all external I/O ports are only connected to the top tier.

Table 1 summarizes comparison between 2-D and M3D designs. The M3D design has 50% smaller footprint and remarkable wirelength reduction than the 2-D design, with the aid of vertical connections through MIVs. Shorter wirelength also leads to smaller net switching power followed by total power reduction. The M3D design flow performs tier-by-tier final routing separately, which means that it is not feasible to perform accurate timing optimization throughout the full design. This makes the timing closure of the M3D design incomplete, resulting in the existence of a certain amount of worst-case negative slack (WNS) despite its design efficiency. There is a solution to mitigate the timing closure problem provided in [9] with post-tier-partitioning timing closure, but it requires the modification of commercial-grade PDKs into an intolerable shape for the CAD tool, and thus is not available with our commercial PDK. In our benchmarks, WNS is in a similar range for both M3D and 2-D designs, which is less than 10% of the target clock period. Figure 2 shows the GDS layouts for the 2-D and M3D designs in Table 1.

During sequential M3D fabrication with the flow shown in Figure 1a, the lower tier may suffer from distortion during the fabrication of the upper tier, thereby incurring transistor degradation and resistance–capacitance (RC) parasitic increments in the lower tier. There are some techniques available to cope with such a situation by forcing the tier-partitioning tool to allocate the M3D design's critical path cells and nets to the faster upper tier [11].

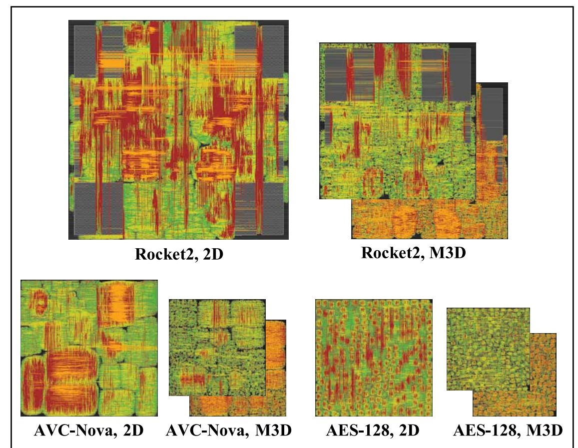

Figure 2. GDS design layouts of the benchmarks in Table 1.

Design-for-test for MIVs

Well-known fault models for MIVs are shorts, opens, and stuck-at faults (SAFs), similar to those for 2-D interconnects. Hard shorts typically occur when a particle contaminates the spacing between adjacent MIVs; an SAF is a hard short between an MIV and the power/ground rail. Resistive shorts occur due to the diffusion of the MIV metal, causing partial contact with an adjacent MIV. Hard opens occur when an MIV does not make contact with its landing pad. Resistive opens in MIVs are of relatively smaller size and can occur due to interface defects, hairline cracks, etc.

Methods like [12] use dedicated control points for launching patterns to test MIVs and dedicated observation points for capturing the test responses. The resultant test time and area overheads can become prohibitively large for large MIV counts [13]. ATPG-based interconnect test methods [14] are not

Table 1. Comparison between 2-D and gate-level partitioned M3D designs on the basis of P&R results with a commercial 28-nm PDK.

	Rocket2, 800MHz			AVC-Nova, 730MHz			AES-128, 3.44GHz		
	2D	M3D	$\Delta(\%)$	2D	M3D	$\Delta(\%)$	2D	M3D	$\Delta(\%)$
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**
MIV 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
WNS (ps)	-54.27	-49.47		-60.48	-135.57		-14.88	-26.87	

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on October 24,2020 at 10:23:51 UTC from IEEE Xplore. Restrictions apply.

suitable for MIV testing as I/O pins are available only on one tier; error due to activated faults must propagate through multiple tiers and MIVs to get detected. This adds to the ATPG complexity and impedes fault localization and diagnosis. Moreover, faults in multiple MIVs on a test path can potentially mask each other leading to test escape and field failures.

MIVs XOR-BIST architecture

An MIV XOR-BIST architecture for testing hard faults in MIVs was introduced in [15] and is illustrated in part (I) of Figure 3a. Two-input XOR gates are inserted between the outputs of adjacent MIVs in the output segment of the BIST engine.

Hence, $(N - 1)$ XOR gates are inserted for a cluster of N MIVs. The XOR outputs are fed to a balanced AND tree-based space compactor with $(N - 1)$ inputs and a one-bit pass/fail response, Y_1; one observation point (scan flop) is sufficient for monitoring the health of multiple MIVs. A fault in the given set of MIVs can be detected by observing Y_1. In the input segment of the BIST engine, an input source V_{in} provides complementary inputs to adjacent MIVs in the test mode via an inverter chain. A 2:1 multiplexer, present at the input of every MIV, enables switching between functional (*FI*) and test modes based on the *Launch* signal.

Two clock cycles are used to test the MIVs for hard faults by setting $V_{in} = 0$ and 1 in consecutive cycles. The resulting test patterns are "101..." in the first cycle and "010..." in the second cycle. It can be shown that a set of MIVs does not contain a hard fault if and only if Y_1 is observed to be 1 in both clock cycles. Aliasing occurs when all MIVs are alternately stuck at 0 and $1 - Y_1$ will be 1 in both cycles. However, such a scenario is highly unlikely with a likelihood of $2/3^m$, where m is the number of MIVs in the set.

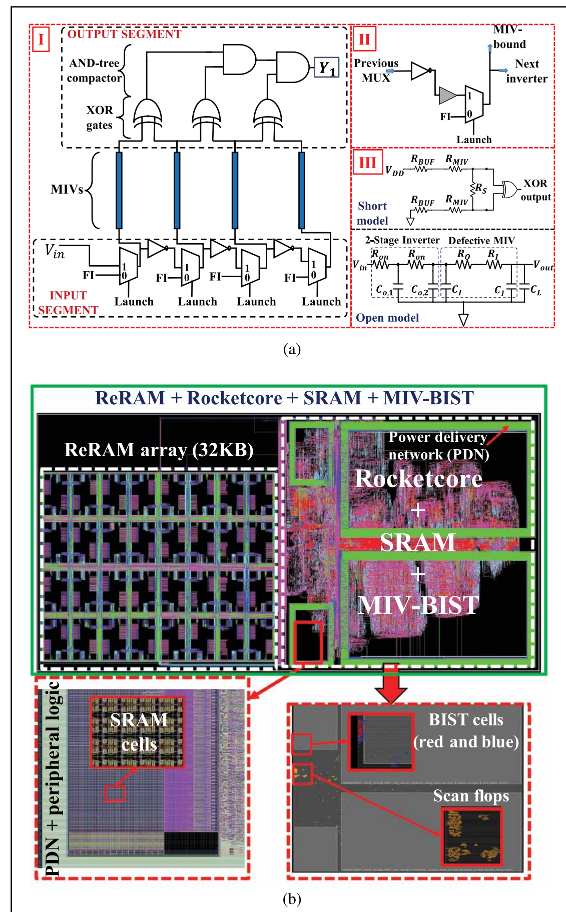

Figure 3. (a) (I) MIV XOR-BIST architecture for hard fault detection, (II) enhancement in input segment of BIST engine for resistive fault detection, and (III) lumped circuit models for resistive short and open tests. (b) Top view of the superimposed layout of a two-tier M3D gate-level partitioned Rocketcore design integrated with ReRAM and MIV-BIST using a commercial 90-nm PDK.

MIVs dual-BIST architecture

The XOR-BIST engine mentioned in the previous section is itself prone to SAFs which, in turn, can mask MIV faults. To minimize the masking probability, a parallel propagation path is added from the MIV outputs to a second one-bit signature Y_2 via a second BIST engine, BIST-B. The topology and connectivity of gates inside BIST-B are identical to that of BIST-A, except that the XOR and AND gates in BIST-A are substituted with their logic duals (XNOR and OR, respectively) in BIST-B. With this "dual-BIST" architecture, it can be shown that MIV fault(s) cannot be masked by a single fault in the dual-BIST engine [15]. The MIV fault-masking probability due to multiple concurrent faults in the dual-BIST engine is also negligible.

Enhanced dual-BIST for resistive fault detection

The size of short or open defect determines the magnitude of its resistance—larger-sized short (open) has lower (higher) resistance and vice versa. For example, a hard short (open) is characterized by very low (high) resistance. As demonstrated in [15],

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on October 24,2020 at 10:23:51 UTC from IEEE Xplore. Restrictions apply.

the range of detectable defect size can be expanded by adding a (programmable) delay element—an N-stage CMOS inverter (N is the number of cascaded single-stage inverters)—to the input of every MIV in the test mode. Part (II) of Figure 3a describes the enhanced dual-BIST with the delay element being a buffer (N = 2); the added buffer is shaded. The addition of the delay stage enables the detection of resistive faults caused by the short and open defects of intermediate sizes.

Resistive shorts

Part (III) of Figure 3a illustrates the lumped circuit models for the testing of resistive short and resistive open. In the resistive short model, the added buffer's equivalent resistance is denoted by R_{BUF}, the MIV self-resistance is denoted by R_{MIV}, and R_S denotes the resistance of the short defect. Complimentary test inputs are applied at the inputs of the buffers—1 (V_{DD}) and 0—to ensure a fault-free XOR output of the differential input voltage of the two-input XOR is given by $\Delta V_{\mathrm{out}} = V_{\mathrm{DD}} R_S/(2R_{\mathrm{MIV}} + 2R_{\mathrm{BUF}} + R_S)$. A short is detected when the XOR output is 0 instead of 1 (fault-free scenario). The XOR output will flip to 0 when its differential input voltage becomes lower than a certain threshold (input offset voltage). As ΔV_{out} for $R_{\mathrm{BUF}} > 0$ is less than ΔV_{out} for $R_{\mathrm{BUF}} = 0$, the addition of the buffer stage decreases the magnitude of the differential input voltage, causing the XOR output to flip $1 \rightarrow 0$. Therefore, the maximum detectable short resistance can be increased by increasing R_{BUF}. Smaller the size of the short defect, higher the resistance. Therefore, enhanced dual-BIST facilitates highly granular detection of a wide range of defect sizes. With the buffer-based enhanced dual-BIST (designed using Nangate 45-nm open-cell library), a resistive short as large as 12 KΩ is detected.

Table 2. Impact of BIST on circuit PPA with a commercial 28-nm PDK.

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

Resistive opens

For a given functional clock frequency f_{clk}, let the minimum open resistance detected by dual-BIST be $R_{o,\min}$. The enhanced dual-BIST can detect open resistances lower than $R_{o,\min}$ by adding a programmable delay to the small delay of a resistive open defect. Part (III) of Figure 3a illustrates the lumped circuit model of the two-stage inverter, enabling the test of a resistive open. Here, R_{on} is the ON-resistance of a transistor, $C_{o,x}$ (x = 1, 2) are the output capacitances of the two inverter stages, C_I is the parasitic MIV-to-substrate capacitance, C_L is the MIV load (fan-out) capacitance, and R_O indicates the open resistance. Using the Elmore–Delay model, the total propagation delay, D, from the buffer's input V_{in} to the MIV output V_{out} is given by

$$D = (\ln 2) \times (R_{\mathrm{on}}(C_{o,1} + 2C_{o,2} + 4C_I + 2C_L)) + (\ln 2) \times (R_O + R_I)(C_I + C_L).$$

A sufficient condition for the detection of a resistive open is: $D + \Delta_{\mathrm{XOR}} > 1/f_{\mathrm{clk}}$, where Δ_{XOR} is the XOR gate delay. Therefore, without increasing f_{clk}, the minimum detectable open resistance can be lowered below $R_{o,\min}$ by increasing R_{on}—by decreasing transistor widths or increasing the number of inverter stages, N. With the buffer-based enhanced dual-BIST (designed using Nangate 45-nm open-cell library), a resistive open as small as 25 KΩ is detected with f_{clk} = 2 GHz. To summarize, increasing the propagation delay of the N-stage inverter increases the range of resistive shorts and opens detected by the enhanced dual-BIST engine.

Overhead for dual-BIST

The impact of MIV dual-BIST on circuit PPA is evaluated for four block-level partitioned M3D benchmarks: Rocket2 (301K cells, 1,200 MIVs, and 350 MHz), AVC-Nova [17] (134K cells, 317 MIVs, and 366 MHz), AES-128 (I) [17] (102K cells, 425 MIVs, and 2,273 MHz), and AES-128 (II) [17] (90K cells, 426 MIVs, and 1,250 MHz). All four M3D benchmarks are obtained using the design flow described in the "M3D physical design" section with a commercial 28-nm PDK.

Table 2 presents the comparison of power consumption (P_{tot}) and the total cell area (A_{tot}) between the BISTed designs (BI) and the non-BISTed designs (N-BI). It can be seen that the area overhead of dual-BIST is negligible. The ~9% power overhead of dual-BIST is manageable by shutting off the BIST engines when MIVs are not being

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on October 24,2020 at 10:23:51 UTC from IEEE Xplore. Restrictions apply.

tested. The dual-BIST engines will be switched off in functional (mission) mode and switched on only in the test mode; such switching of modes is enabled with multiplexers inserted on the BIST engine inputs. Switching off the dual-BIST engine will significantly reduce the switching activity, and hence dynamic power consumption, of the BIST cells in the functional mode.

For all four M3D benchmarks, the dual-BIST scheme achieves greater than 1.5× area reduction than the trivial DfT scheme where test points are added at both ends of an MIV for direct control and observation.

Figure 3b demonstrates the top view of the BIST-inserted M3D layout of the gate-level partitioned design of the Rocketcore integrated with on-chip 32-kB ReRAM array using commercial 90-nm PDK. In this top view, the two M3D tiers are superimposed. MIVs are implemented on the right-hand side (Rocketcore + SRAM + MIV-BIST) as a via layer between the two tiers. SRAMs work as L1 caches for the Rocketcore, and thus they are within the Rocketcore logic floorplan (right side). The ReRAM array functions as an external DRAM, which stays outside the Rocketcore logic.

ReRAM compiler

ReRAMs have emerged as attractive candidates for NVM technology. There have been significant efforts in developing ReRAM device technologies as well as analysis tools for ReRAM architecture and their impact on processor design [18]. When used as a memory device, ReRAMs encode bits with the resistance of the device. However, the stochastic behavior of ReRAM devices during read and write has resulted in many challenges during physical implementation.

Many custom design of ReRAM macros [4], [5], [19] have been proposed and developed with various techniques to address these issues. One of the more distinguishable developments is the implementation of write termination (WT) circuits. This specific type of circuit is designed to address the large variation in the switching time of ReRAM devices. By stopping the process of writing when the resistance of cells reach a predesigned threshold, overstress and performance loss in fast-switching cells can be mitigated even when using a long write time to guarantee successful writing of slow-switching cells [5].

Design of complex peripheral circuits to adapt to device variations on-the-fly is crucial for reliable control for read/write processes. However, as they can greatly impact the performance of subarrays in practice, it is equally crucial to include their effects during modeling.

With WT, the overall write energy of the ReRAM subarrays becomes not only dependent on the switching mechanism of bit cells and the switching of word-lines (WLs)/bit-lines (BLs), but also the latency of the WT to respond to changes in BL after a device switches. Although conventional memories such as SRAMs and DRAMs can model read/write energy and latency in a decoupled fashion with parasitic passive models and peripheral circuit specifications accurately, circuit-level simulations with device models and peripheral circuits attached to layout-extracted parasitics becomes indispensable in performing accurate design space exploration for ReRAM subarrays. However, there is a lack of EDA methods and tools for the physical design of ReRAM subarrays, and the generation of large ReRAM macros.

Compiler architecture

Fully automated and scalable physical implementation with performance evaluation are key foundations of the ReRAM compiler. A complete ReRAM module generator will also need to partition ReRAM macros into optimized ReRAM subarrays of different dimensions (i.e., $N_{\text{row}} \times N_{\text{column}}$), and accurately evaluate their performance/power when assembled into a full array module.

An EDA methodology for the automated generation of ReRAM memory arrays with the ability to perform the layout-precise evaluation of the performance is presented. The proposed approach combines the circuit-level design of cells and peripherals with the automated layout generation of custom circuits to enable the physical design of ReRAM subarrays. The autogenerated ReRAM subarrays are coupled with autogenerated (logic synthesis and P&R) connectivity between subarrays to design a large array. An evaluation framework for the fast circuit level and layout accurate ReRAM performance/power analysis is embedded to enable design space exploration and optimization. This enables design space exploration and optimization for full array designs. The tool structure is shown in Figure 4.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on October 24,2020 at 10:23:51 UTC from IEEE Xplore. Restrictions apply.

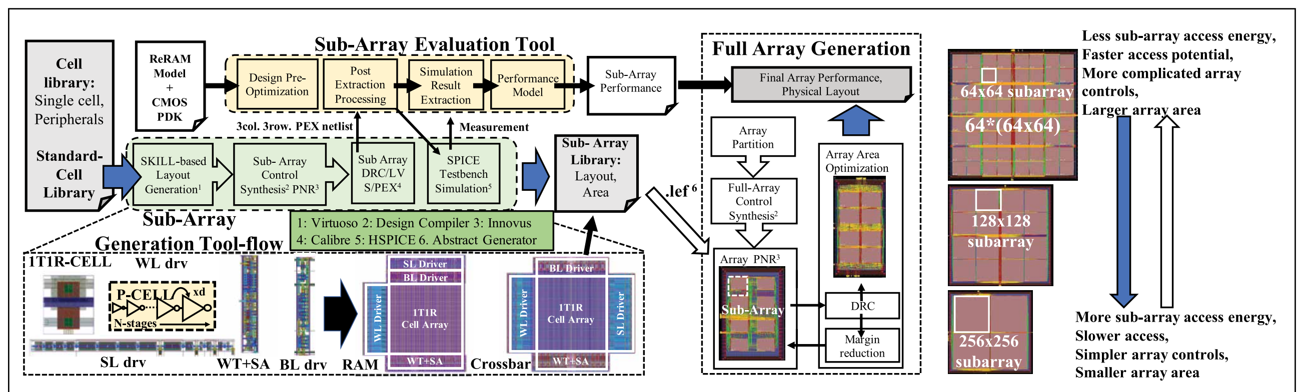

Figure 4. Top-level view of tool architecture and sample layouts.

Predesigned peripheral circuits and single-cell layouts are taken as inputs to the tool for layout generation along with PDK models and ReRAM device models for performance evaluation. Subarray layouts are generated with Cadence SKILL to be sourced in full-array generation, where it is synthesized with array controls and placed and routed (PnR) to final array layouts. Parasitic extracted 4 × 4 subarrays are simulated to extract latency and power for single-bit access, which is then combined with full-array PnR results to estimate final performance.

Array generation results

Results of the developed ReRAM compiler are demonstrated for 1T-1R ReRAM cells considering 65-nm CMOS technology and open-source ReRAM models [20]. The experimental results show that our tool can generate the physical design of ReRAM subarrays for different cell topology and subarray dimensions including read/write circuits. The proposed tool allows us to optimize the subarray configurations (i.e., number or rows versus columns) for different target subarray sizes, as well as select optimal partitions for large ReRAM arrays. Table 3 summarizes performance and area for generated arrays of different (8 and 32 kB) capacities.

Table 3. Design results for ReRAM compiler with a 65-nm PDK.

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

* R.E./W.E. = Read/Write Energy
** Top. E. = Top module energy for controlling sub-arrays.

Although smaller subarrays can realize faster access, it results in higher area and power overhead at the array level for more subarrays. Top module energy is calculated using the P&R tool power estimations at designated frequency. Read/write energy includes both subarray internal energy consumption and top module energy. As seen for the 8 kB array capacity, although the top module energy for the subarray is smallest for the 256 × 256 subarray, the 128 × 128 subarray shows the minimum read operation energy due to the lower read energy consumption of the ReRAM subarray. Also, although the write energy is smallest for the 64 × 64 subarray for a 8-kB total capacity, it increases with top module energy for larger array capacities, and the 128 × 128 subarray in turn becomes the optimum design regarding write energy consumption for a design of 32-kB total capacity. For both sizes, using a 256 × 256 subarray results in the minimum total area due to the less area overhead to P&R the top module and less peripheral circuit overhead within subarrays. These tradeoffs highlight the need for an automated memory compiler for design space exploration.

The compiler structure can easily be adapted to other technologies/custom design libraries by replacing the input cell library; this has been facilitated in the design demonstrated in Figure 3b. The M3D ReRAM arrays have been designed with peripheral control circuits placed not at the boundary of the cell array, but directly on top, to reduce the 2-D footprint of subarrays that result from large drivers designed for ReRAM programming.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on October 24,2020 at 10:23:51 UTC from IEEE Xplore. Restrictions apply.

In this article, we presented recent advances in the design and test of M3D ICs comprising end-to-end automated design flow, low-cost test solution for MIVs, and ReRAM array compiler for high-density memories. Further research is needed for enhancing the capabilities of these EDA tools to boost high-volume production of M3D ICs for applications in low-power high-performance computing. ■

Acknowledgments

This work was supported by the DARPA ERI 3DSOC Program under Award HR001118C0096.

References

[1] S. Wong et al., "Monolithic 3-D integrated circuits," in *Proc. VLSI-TSA*, Apr. 2007, pp. 1–4.

[2] A. Koneru, S. Kannan, and K. Chakrabarty, "A design-for-test solution based on dedicated test layers and test scheduling for monolithic 3-D integrated circuits," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 38, no. 10, pp. 1942–1955, Oct. 2019.

[3] A. Koneru, S. Kannan, and K. Chakrabarty, "Impact of electrostatic coupling and wafer-bonding defects on delay testing of monolithic 3-D integrated circuits," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 13, no. 4, pp. 1–23, Aug. 2017.

[4] C.-C. Chou et al., "An N40 256K x 44 embedded RRAM macro with SL-precharge SA and low-voltage current limiter to improve read and write performance," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2018, pp. 478–480.

[5] M.-F. Chang et al., "19.4 embedded 1 Mb ReRAM in 28 nm CMOS with 0.27-to-1 V read using swing-sample-and-couple sense amplifier and self-boost-write-termination scheme," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2014, pp. 332–333.

[6] H. Park et al., "RTL-to-GDS tool flow and design-for-test solutions for monolithic 3-D ICs," in *Proc. DAC*, 2019, pp. 1–4.

[7] S. A. Panth et al., "Design and CAD methodologies for low power gate-level monolithic 3-D ICs," in *Proc. ISLPED*, 2014, pp. 171–176.

[8] K. Chang et al., "Cascade2-D: A design-aware partitioning approach to monolithic 3-D IC with 2-D commercial tools," in *Proc. ICCAD*, Nov. 2016, pp. 1–8.

[9] B. W. Ku, K. Chang, and S. K. Lim, "Compact-2-D: A physical design methodology to build commercial-quality face-to-face-bonded 3-D ICs," in *Proc. ISPD*, 2018, pp. 90–97.

[10] C. M. Fiduccia and R. M. Mattheyses, "A linear-time heuristic for improving network partitions," in *Proc. DAC*, 1982, pp. 175–181.

[11] S. K. Samal et al., "Tier partitioning strategy to mitigate BEOL degradation and cost issues in monolithic 3-D ICs," in *Proc. ICCAD*, Nov. 2016, pp. 1–7.

[12] R. Pendurkar, A. Chatterjee, and Y. Zorian, "Switching activity generation with automated BIST synthesis for performance testing of interconnects," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 20, no. 9, pp. 1143–1158, 2001.

[13] A. Jutman, "Shift register based TPG for at-speed interconnect BIST," in *Proc. ICM*, 2004, pp. 751–754.

[14] D. Erb et al., "Multi-cycle circuit parameter independent ATPG for interconnect open defects," in *Proc. VTS*, Apr. 2015, pp. 1–6.

[15] A. Chaudhuri et al., "Built-in self-test for inter-layer vias in monolithic 3-D ICs," in *Proc. ETS*, May 2019, pp. 1–6.

[16] *RISC-V Cores and SoC Overview*. Accessed: Jul. 1, 2018. [Online]. Available: https://riscv.org/risc-v-cores

[17] *OpenCore Benchmark Suite*. Accessed: Jul. 1, 2018. [Online]. Available: http://www.opencores.org/

[18] X. Dong et al., "NVSim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 31, no. 7, pp. 994–1007, Jun. 2012.

[19] S.-S. Sheu et al., "A 4 Mb embedded SLC resistive-RAM macro with 7.2ns read-write random-access time and 160ns MLC-access capability," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2011, pp. 200–202.

[20] H. Li et al., "Variation-aware, reliability-emphasized design and optimization of RRAM using SPICE model," in *Proc. DATE*, 2015, pp. 1425–1430.

Arjun Chaudhuri is currently pursuing a PhD in electrical and computer engineering with Duke University, Durham, NC. He was an intern at Global Foundries Inc., Malta, NY and Alibaba Computing Technology Lab., Sunnyvale, CA. His current research interests include design-for-testability of monolithic 3-D ICs and neuromorphic computing systems. Chaudhuri has a BTech from the Indian Institute of Technology Kharagpur, Kharagpur, India (2017).

Sanmitra Banerjee is currently pursuing a PhD in electrical and computer engineering with Duke University, Durham, NC. He was an intern at Texas Instruments Inc., Bengaluru, India, and Intel Corporation, Folsom, CA. His current research interests include fault modeling and design-for-testability

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on October 24,2020 at 10:23:51 UTC from IEEE Xplore. Restrictions apply.

solutions for emerging technologies like carbon nanotube field-effect transistors (FETs) and monolithic 3-D ICs. Banerjee has a BTech from the Indian Institute of Technology Kharagpur, Kharagpur, India (2018).

Heechun Park is currently working as a Post-Doctoral Fellow with the Georgia Institute of Technology, Atlanta, GA. His current research interests include electronic design automation of 3-D and 2.5-D IC. Park has a BS in electrical engineering from Seoul National University, Seoul, South Korea (2011), and integrated MS and PhD from the same department (2018).

Jinwoo Kim is currently pursuing a PhD in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA. His current research interests include 2.5-D/3-D IC design and analysis. Kim has a BS in electrical and computer engineering and an MS in electrical engineering and computer science from Seoul National University, Seoul, South Korea, in 2011 and 2013, respectively.

Gauthaman Murali is currently pursuing an MS and a PhD in electrical and computer engineering with the Georgia Institute of Technology, Atlanta, GA, where he worked as an SoC Design Engineer at Intel, Bengaluru, India, for two years. His research interests include 2.5-D/3-D IC physical design. Murali has a BE in electrical and communication engineering from the PSG College of Technology, Coimbatore, India (2016).

Edward Lee is currently pursuing a PhD with the Georgia Institute of Technology, Atlanta, GA. His current research interests include low-power circuits and power management for energy harvesting and volatile/nonvolatile memory designs with emerging technologies. Lee has a BS in electrical engineering (EE) from National Taiwan University, Taipei City, Taiwan (2015) and an MS in electrical and computer engineering (ECE) from the Georgia Institute of Technology (2017).

Daehyun Kim is currently pursuing an MS and a PhD in electrical and computer engineering with the Georgia Institute of Technology, Atlanta, GA, under the supervision of Prof. S. Mukhopadhyay. His research interests include hardware accelerators and memory compiler. Kim has a BS in semiconductor system engineering from Sungkyunkwan University, Seoul, South Korea (2016).

Sung Kyu Lim is currently a Full Professor with the Georgia Institute of Technology, Atlanta, GA, where he joined the School of Electrical and Computer Engineering (2001). His research focus is on the architecture, circuit design, and physical design automation for 3-D ICs. He has published more than 300 articles on 3-D ICs. Lim has a PhD from the University of California at Los Angeles (UCLA), Los Angeles, CA (2000).

Saibal Mukhopadhyay is currently a Joseph M. Pettit Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA. His research interests include design of energy-efficient, intelligent, and secure systems by exploring a cross-cutting approach spanning algorithm, architecture, circuits, and emerging technologies. Mukhopadhyay has a PhD in electrical and computer engineering from Purdue University, West Lafayette, IN (2006). He is a Fellow of IEEE.

Krishnendu Chakrabarty is currently the John Cocke Distinguished Professor and the Department Chair of electrical and computer engineering, and a Professor of computer science with Duke University, Durham, NC. His current research projects include design-for-testability of 3-D ICs, microfluidic biochips, hardware security, failure prediction, and neuromorphic computing. Chakrabarty has a BTech from the Indian Institute of Technology Kharagpur, Kharagpur, India (1990) and an MSE (1992) and a PhD (1995) from the University of Michigan, Ann Arbor, MI. He is a Fellow of AAAS, a Fellow of IEEE, and a Fellow of ACM.

■ Direct questions and comments about this article Arjun Chaudhuri, Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA; arjun.chaudhuri@duke.edu.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on October 24,2020 at 10:23:51 UTC from IEEE Xplore. Restrictions apply.