Power Supply Noise-Aware At-Speed Delay Fault Testing of Monolithic 3-D ICs

Shao-Chun Hung, *Graduate Student Member, IEEE*, Yi-Chen Lu, Sung Kyu Lim, *Senior Member, IEEE*, and Krishnendu Chakrabarty, *Fellow, IEEE*

***Abstract*— Monolithic 3-D (M3-D) integration is an emerging technology that offers significant power, performance, and area benefits for an integrated circuit (IC) design. However, a problem with the 3-D power distribution network in such ICs is that it can lead to high power supply noise (PSN) during the capture cycles in at-speed scan testing for transition delay faults. Therefore, the failure of good chips (i.e., yield loss) resulting from the PSN-induced voltage droop is a major concern for M3-D designs. In this article, we first assess the PSN and voltage droop problems and their impact on path delays for at-speed testing of benchmark M3-D designs. Next, we present an analysis framework to identify test patterns that are most likely to lead to yield loss. We describe a test-pattern reshaping solution based on integer linear programming to make appropriate changes to the test patterns that cause yield loss. Simulation results for four M3-D benchmarks highlight the effectiveness of the proposed solution.**

***Index Terms*— Delay fault testing, monolithic 3-D integration, power distribution network (PDN), voltage droop.**

I. Introduction

THE 3-D integration provides a path to go beyond Moore's law, achieves higher circuit performance and package density, as well as reduces power consumption. Monolithic 3-D (M3-D) is a promising technology enabled by fine-grained vertical interconnects, known as monolithic interlayer vias (MIVs). MIVs are one to two orders smaller in size than the through-silicon-vias (TSVs) used in today's 3-D integration technology [1]. Such small MIV dimensions enable high precision alignment and area reduction in M3-D integration. Despite these benefits, a number of test challenges need to be addressed before M3-D integration can become ready for commercial exploitation. One of these challenges is related to power supply noise (PSN) during scan testing. Recent work based on both static and dynamic analyses has shown that compared with 2-D designs, M3-D suffers more from PSN problems [2]. A major problem with the 3-D power distribution network (PDN) in M3-D integrated circuits (ICs) is that it can lead to high PSN during the capture cycles in at-speed scan testing for transition delay faults. The PSN problem is more severe in the test mode due to higher switching activities of the circuit nodes compared to functional operation [3]. Therefore, the failure of good chips (i.e., yield loss) resulting from the PSN-induced voltage droop is a major concern for M3-D designs. Thus, the PSN should be carefully considered during test-pattern generation for at-speed testing of M3-D ICs.

At-speed scan testing is necessary for effective delay testing in today's scan-based designs [4]. The key idea underlying at-speed testing is to launch transitions at the start points of sensitized paths and capture responses at the end points within a specific timeframe that depends on the system clock period. However, a problem with at-speed scan testing is that the power consumption in the test mode is several times higher than in the functional mode and the current drawn from the PDN is also much higher than what is included in functional-mode specifications for designing the PDN [5]. This problem is especially severe for capture cycles because the rated functional clock frequency is used to simultaneously capture test responses in all scan flip-flops (FFs) in the design. The transitions at the outputs of FFs propagate through the combinational logic and lead to high toggle activity in the design. Excessive power consumption and high current drawn from the PDN lead to voltage droop, resulting in slower signal propagation through sensitized paths, the failure of good chips, and yield loss. Note that stuck-at faults are used for static testing and they do not affect PSN because static testing is not carried out at-speed. Therefore, in this work, we focus on the impact of PSN-induced voltage droop during delay fault testing.

Various strategies have been proposed in the literature to mitigate the problem of high power consumption during testing; these methods include test scheduling [6], circuit modification [7], test-pattern modification [8], and scan-chain ordering [9]. Algorithms based on the filling of don't-care bits (X-filling) have been proposed to manipulate test patterns to reduce power consumption. In [10], a justification-based algorithm was proposed to ensure low switching activities at FFs during scan capture. A probability-based X-filling algorithm is described in [11] to reduce the computation effort

Manuscript received March 4, 2021; revised June 17, 2021 and July 14, 2021; accepted August 11, 2021. Date of publication September 15, 2021; date of current version November 3, 2021. This work was supported in part by the DARPA Electronics Resurgence Initiative 3DSOC Program under Award HR001118C0096. The work of Shao-Chun Hung and Krishnendu Chakrabarty was supported in part by the National Science Foundation under Grant CCF-1908045. A preliminary version of this article was published in *Proc. IEEE Asian Test Symposium*, 2020. *(Corresponding author: Shao-Chun Hung.)*

Shao-Chun Hung and Krishnendu Chakrabarty are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27705 USA (e-mail: shaochun.hung@duke.edu).

Yi-Chen Lu and Sung Kyu Lim are with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TVLSI.2021.3108787.

Digital Object Identifier 10.1109/TVLSI.2021.3108787

associated with backtracing from output responses to input signals. In [12], an efficient solution is presented to identify don't-care bits in a test pattern without degrading test quality; it combines justification with probabilistic analysis to improve the effectiveness and scalability of X-filling techniques. However, previously proposed X-filling algorithms are of limited effectiveness for M3-D designs due to the differences in the layout and the PDN.

In this article, we first present an analysis framework to conduct dynamic power and rail analysis for each scan test pattern for an M3-D design. Based on this analysis, we determine the PDN voltage droop and compute the increase in delay for logic paths sensitized by each pattern. This information is used to determine the test patterns for which the slack on long paths becomes negative under the rated functional clock period. These test patterns, which contribute to yield loss, are then appropriately reshaped through X-filling to prevent excessive voltage droop. We present two X-filling techniques, based on integer linear programming (ILP) and simulated annealing (SA), respectively, to ensure that the test patterns are reshaped without any adverse impact on fault coverage. These reshaped patterns and the resulting test outcomes are not affected by voltage droop during scan capture.

The rest of this article is organized as follows. Section II provides an overview of M3-D integration and scan testing. Section III describes the design flow for M3-D ICs, especially the PDN and current delivery. Section IV presents the proposed framework for dynamic power and rail analysis and describes how we compute the delay of logic gates under voltage droop. A PSN-aware pattern reshaping algorithm is proposed in Section V. Section VI presents the simulation results for benchmark M3-D designs and a comparison with a baseline case that uses test vectors generated by an automatic test pattern generation (ATPG) tool. Section VII discusses the difference between our proposed methods and the impacts of process variations and test compression environments. Finally, Section VIII concludes this article.

II. Background

A. Monolithic 3-D Integration

M3-D integration has been made possible by significant breakthroughs in low-temperature manufacturing processes [13]. Manufacturing the upper tiers of an IC with low-temperature processing avoids damage to transistors and interconnects in the bottom tiers. M3-D design styles depend on the type of design-partitioning method employed; partitioning at the transistor level, gate level, and block level have been described in the literature [14]. In transistor-level M3-D ICs, P-channel and N-channel transistors are divided into different tiers; in gate-level M3-D, each tier is composed of standard cells. Functional blocks are partitioned into multiple tiers in a block-level M3-D IC. Fig. 1 demonstrates three design styles of M3-D. The gate-level design appears to be the most promising because a cumbersome redesign of standard cells is required for the transistor-level design, while the block-level design does not fully exploit the benefits of high-density MIVs [15]. A complete design flow for gate-level M3-D ICs

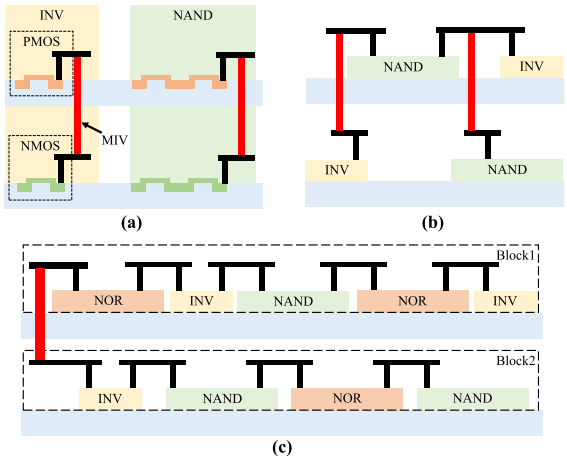

Fig. 1. M3-D design styles. (a) Transistor level. (b) Gate level. (c) Block level.

is proposed in [14]. The reduction in chip footprint is realized by redesigning larger standard cells by partitioning them into different tiers.

Despite advances in design techniques, much less effort has been devoted to the testing of M3-D ICs. In [16], a dedicated layer was introduced as a design-for-test solution for M3-D integration. In [17], a built-in self-test solution was presented for detecting MIV faults. A PDN design technique to alleviate reliability and PSN problems was presented in [18]. However, the problem of test generation for M3-D ICs, especially under PSN constraints, has not been addressed in prior work.

B. Power Supply Noise

PSN is defined as the difference between the nominal supply voltage value and the voltage level at local receivers [19]. PSN-induced voltage droop is composed of two components: IR-drop and Ldi/dt. When switching activities occur, instantaneous current flows through the PDN to cause transitions at the inputs of logic gates. The equivalent resistance along this conduction path causes IR-drop. In a high-speed circuit, rapid changes in the current drawn from the PDN and the parasitic inductance result in large Ldi/dt. For the M3-D IC design, considerable research efforts have been devoted to PDN optimization. In [2], system-level modeling and simulations were carried out in both the time and frequency domains. A comprehensive full-chip study, including the consideration of wire length, power consumption, MIV count, and thermal impact, was carried out in [20]. The design of a reliable PDN based on generic programming is described in [18]. However, the PDN in [18] and other related work is optimized only for functional-mode operations. Such a PDN design overlooks the PSN in the test mode and the impact of voltage droop on scan testing.

C. Delay Testing

Delay testing is used to detect timing-related faults. The transition delay fault is commonly used by ATPG tools [21]. A test pattern for a transition delay fault requires a sequence of two vectors (V_1, V_2). V_1 is first shifted into a full scan circuit for initialization. V_2 is applied to the circuit to launch

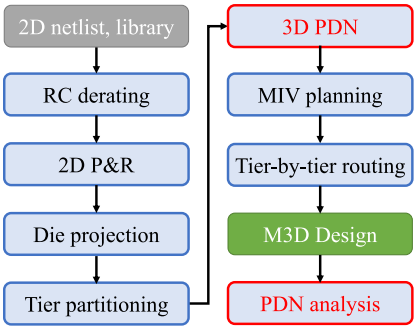

Fig. 2. Our M3-D design flow, including the integration of the PDN.

a transition for target faults, and a subsequent capture pulse is used to record the output responses. There are two methods typically used to implement the launch of a transition, namely *launch-off-shift* (LOS) and *launch-off-capture* (LOC). The LOS technique makes it easier for ATPG to generate transition delay fault patterns because of the controllable launch path. However, at-speed testing requires the scan-enable signal to change at-speed within the functional clock period, which increases design cost and effort. The LOC solution increases ATPG runtime and leads to lower fault coverage, but it is more practical because the scan enable signal does not have to switch at-speed. Therefore, we consider the LOC test-application method. In this article, our goal is to evaluate the impact of PSN and voltage droop on LOC-based transition delay-fault testing and reshape the LOC test patterns to minimize the yield loss with no adverse impact on fault coverage.

III. Design Flow

A. Overview

An overview of our design flow is shown in Fig. 2. The M3-D design flow we target is Compact2D [22], which is the state-of-the-art register-transistor level (RTL)-to-GDSII 3-D implementation flow that leverages 2-D commercial tools to build 3-D ICs [23]. Note that the original Compact2D flow does not consider PDNs. In this article, we have enhanced the original flow to incorporate a PDN in the final M3-D design.

To mimic the final 3-D design during 2-D stages, Compact2D first scales the RC parasitics by $1/(2)^{1/2}$ for placement and routing and then projects the entire design onto a tier with half of the original footprint. After the projection, a bin-based min-cut tier partitioning algorithm is utilized to transform the 2-D design into 3-D by assigning the z-location for every instance. This partitioning algorithm minimizes the overall connection between the two partitioned tiers (tiers), while balancing the cell area in both tiers.

After tier partitioning, we build the 3-D PDN before the original MIV planning stage in the Compact2D flow. The main reason is to avoid signal MIVs being placed at the rails of PDN or overlapped with power MIVs, thereby preventing PDN degradation caused by the conventional MIV planning. To build the 3-D PDN, we stacked the metal layers from both bottom tiers and top tiers. Note that the pins of the cells are annotated with respect to the original cell locations, so that the pins in the bottom tier will leverage the original M1 layer,

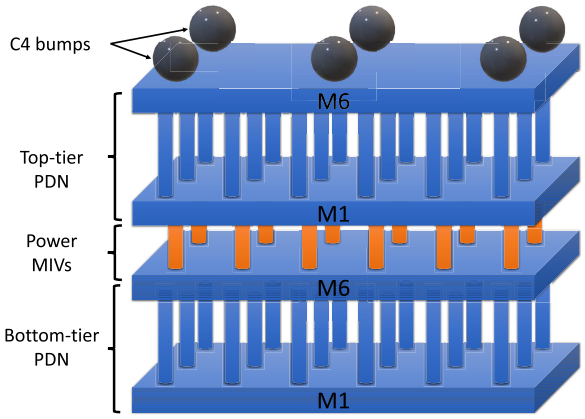

Fig. 3. Illustration of the cross-sectional view of an M3-D PDN.

and the pins that are original in the top tiers would utilize the M7 layer (assume a tier has six metal layers). After developing the power network, the power vias that connect the top metal layer of the bottom tier and the bottom metal layer of the top tier will be the power MIVs. An illustration of an M3-D PDN is shown in Fig. 3.

Following from the 3-D PDN stage, we perform the original MIV planning in the Compact2D flow. To determine the locations of signal MIVs, a 3-D global routing is performed on the stacked metal layers of both tiers, where signal MIVs are the vias that connect the bottom tier and top tiers as power MIVs.

After the MIV planning, legalization, and a timing-driven tier-by-tier routing are performed, it results in a fully placed and routed subdesigns in both tiers. Finally, the subnetlists in both tiers are merged into a single final M3-D design, where timing/power analysis as well as the PDN analysis are performed.

B. Tier Partitioning Strategies

Tier partitioning is one of the most critical stages of an M3-D design flow; it assigns standard cells to different tiers and directly determines the quality of the final full-chip designs. Recent work [24] describes a tier partitioning method that utilizes graph neural networks (GNNs) instead of the conventional bin-based min-cut method originally adopted by Shrunk2D and Compact2D. This new approach shows significant power, performance, and area (PPA) improvements. In this work, to thoroughly investigate the impact of different M3-D design flows, we perform our experiments using both tier partitioning strategies (min-cut-based, GNN-based) and present a detailed comparison.

IV. Dynamic Analysis

In this section, we describe a framework for the dynamic simulation and yield-loss analysis. The overall flow for the M3-D dynamic analysis with test patterns is shown in Fig. 4. Simulations were performed on four benchmark two-tier M3-D designs, namely low-density parity check (LDPC) and Tate Bilinear Pairing (Tate) from OpenCores, and netcard and leon3mp from the International Symposium on Physical Design (ISPD) 2012 benchmark suite [25]. Tables I and II

TABLE I
DESIGN MATRIX OF BENCHMARK MIN-CUT-BASED M3-D DESIGNS

Design	Frequency (MHz)	Footprint (μm^2)	# Cells	# FFs	# Power MIVs	# Patterns	Fault Coverage
LDPC	650	263 × 262	92010	2048	347	194	99.73%
Tate	714	631 × 630	209425	31409	12338	558	98.37%
netcard	500	816 × 815	250842	67446	19649	43103	96.71%
leon3mp	300	1035 × 1035	397420	108720	15328	17087	98.87%

TABLE II
DESIGN MATRIX OF BENCHMARK GNN-BASED M3-D DESIGNS

Design	Frequency (MHz)	Footprint (μm^2)	# Cells	# FFs	# Power MIVs	# Patterns	Fault Coverage
LDPC	833	210 × 209	49091	2048	190	203	99.74%
Tate	403	506 × 505	194067	31409	1772	592	98.26%
netcard	513	654 × 653	252483	67446	2394	43423	96.81%
leon3mp	313	803 × 829	343859	108720	4810	17158	98.82%

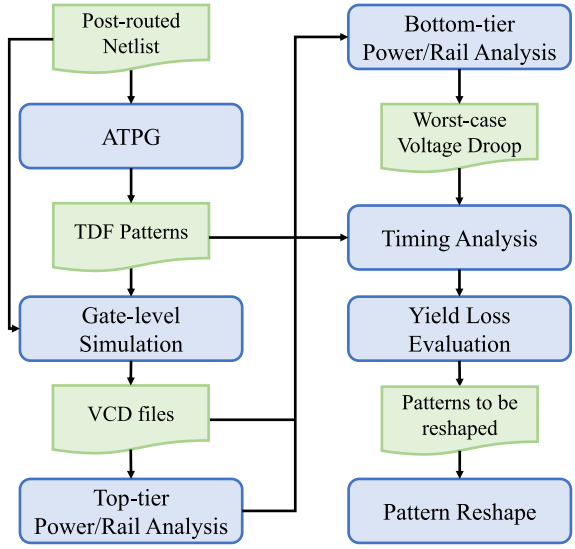

Fig. 4. Dynamic analysis flow for M3-D ICs.

provide the design matrix and ATPG results of min-cut-based M3-D designs and GNN-based M3-D designs, respectively. Experimental results highlight the problem of voltage droop due to PSN from the PDN. These results also motivate the need for an optimization method for pattern reshaping.

A. M3-D Power and Rail Simulation

We developed a framework to conduct dynamic power and rail analysis for M3-D ICs with Cadence Voltus. We generated transition-delay patterns after place and route. The patterns were written out in the STIL format and converted into a Verilog testbench by Synopsys Tetramax. Next, we used Mentor Graphic ModelSim to conduct post-routed gate-level simulation and dump the value change dump (VCD) files to record switching activities of each pattern. VCD files were imported into Cadence Voltus to perform vector-based dynamic power and rail analysis.

Because commercial tools do not consider M3-D designs, we created a method to analyze two tiers in an M3-D IC separately with the 2-D power and rail analysis flow. For the top tier, the PDN design can be extended to a system-level model considering printed circuit board (PCB), package, and C4 bumps [2]. The distance between two C4 bumps was set to 120 μm. One major difference between the traditional 2-D ICs and M3-D ICs is that the supply current of the bottom tier in M3-D flows through the top tier. Therefore, additional power consumption and current demand are superimposed on the top tier. To simulate this scenario, we scaled the current in the PDN of the top tier during power and rail analysis. For the bottom tier, the locations and the parasitics of power MIVs, that is, MIVs belonging to the PDN, were extracted during place and route. However, the reference voltage for the power MIVs was no longer the nominal value due to the voltage droop in the top tier. To analyze the worst case scenario, we subtracted the worst case voltage droop obtained in the rail analysis of the top tier from the nominal supply voltage and utilized this new value as the power source of the PDN in the bottom tier.

B. M3-D Dynamic Rail Analysis

Due to the limitations inherent in commercial tools with respect to M3-D, the simulation window could not be extended to the complete test procedure. Total power consumption is proportional to the occurrence of a switching activity of each net multiplied by its fan-out.

The weighted switching activity (WSA) [26] has been used in the literature to estimate power consumption during scan capture. Let V_1 and V_2 be a pair of test patterns and the state of each net n_i be a pair of $n_i(V_1)$ and $n_i(V_2)$ when V_1 and V_2 are applied, respectively. The calculation of WSA is carried out as follows:

$$\mathrm{WSA}(V_1, V_2) = \sum_{i}^{\mathcal{N}} \Big((F_i + 1) \cdot (n_i(V_1) \oplus n_i(V_2)) \Big) \quad (1)$$

where F_i is the number of fan-out of net n_i and $\mathcal{N}$ is the number of nets in the design. We first calculate the WSA of every pattern and extract patterns with large WSA values. Next, we simulate the extracted patterns to obtain the worst case voltage droop during test application.

Fig. 5 shows the voltage droop distributions of the PDN in the worst case scenario for scan capture, where the upper part of each figure refers to the top tier and the bottom part refers to the bottom tier. Note that the power source is different for each tier. For the top tier, the voltage is supplied from a dc power source with the nominal voltage; the voltage in the bottom tier is supplied from power MIVs with the reference voltage lower than the nominal value due to the voltage droop in the top tier, as discussed in Section IV-A. As the vertical connections

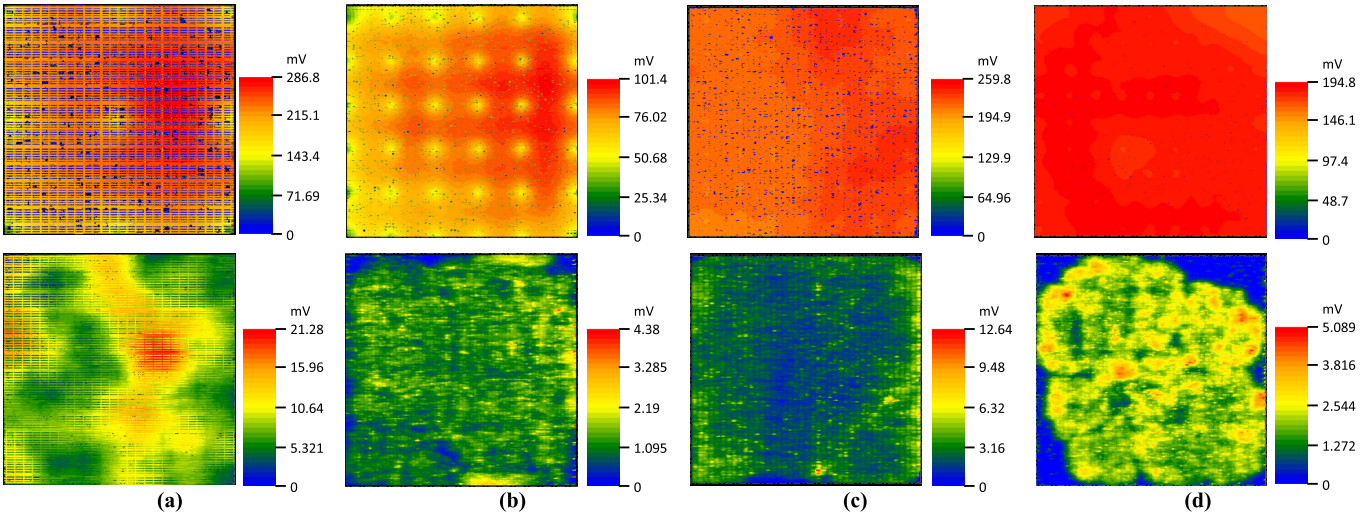

Fig. 5. Instantaneous voltage droop distributions in the worst case scenario during testing in min-cut-based M3-D designs. (a) LDPC. (b) Tate. (c) Netcard. (d) Leon3mp.

TABLE III
RELATIONSHIP BETWEEN WSA AND VOLTAGE DROOP FOR THE LDPC BENCHMARK

Pattern ID	WSA (total)	WSA (top tier)	Voltage Droop (mV)
41	130941	64149	264
6	130980	62253	247
14	131871	62695	247
Functional mode			238

and the 3-D design are important features different from the traditional 2-D ICs, our discussion focuses on the switching activities of the two tiers and their impacts on the voltage droop during testing. Note that the voltage droop problem in the bottom tier is less severe than in the top tier. As the size of the designs increases, this difference becomes more obvious. This phenomenon is observed in every benchmark design that we have considered, and it has been explained in [2]. In M3-D designs, high-density power MIVs provide a large number of current sources for the bottom tier, which prevents a large-magnitude current from flowing through power rails near power MIVs and therefore mitigates the IR-drop problem in the PDN. On the other hand, for the top tier, the number of C4 bumps is limited by the bump size. The reduction of footprint in an M3-D IC compared with its 2-D counterpart exacerbates this problem. Therefore, it is only expected that the voltage droop problem for the bottom tier is less severe than for the top tier. A large design requires an increase in the chip footprint, enabling the M3-D PDN to add more power MIVs between the two tiers to deliver current from the top to the bottom. Hence, this scenario can be observed more clearly in large designs.

To further examine this behavior in the test mode, we calculate the WSA for nets in the top tier only and make a comparison with voltage droop. The relation between the WSA of the top tier and the voltage droop for transition delay fault patterns is shown in Fig. 6. From these results, we conclude that there is a high positive correlation between the voltage droop in the test mode and the switching activities in the top tier.

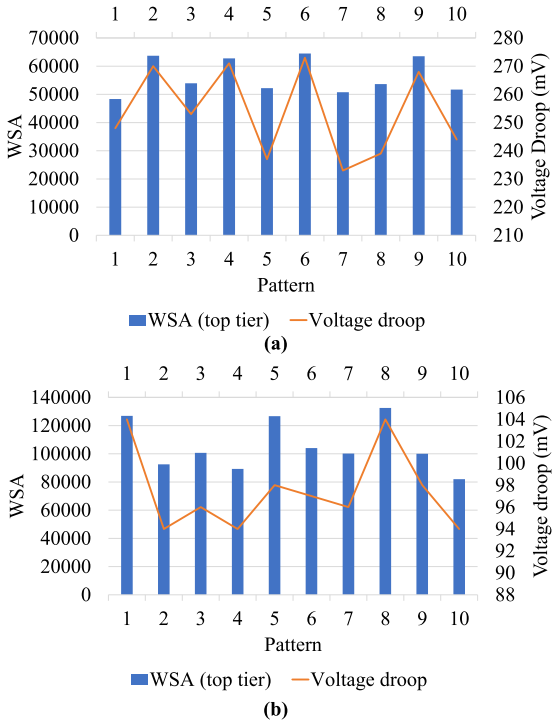

Fig. 6. WSA and voltage droop for the top tier for various test patterns. (a) LDPC. (b) Tate.

Table III provides a comparison between the three test patterns for LDPC that have similar total WSA but different WSA for the top tier. Pattern 41 and Pattern 6 have almost the same WSA for the whole design but have a 17 mV difference in voltage droop. The total switching activities are even larger in Pattern 14. However, the voltage droop is worse in Pattern 41 than in Pattern 14. Therefore, a key contributor to the voltage droop is the switching activities in the top tier, instead of total switching activities for the full M3-D design. ATPG tools for 2-D designs typically choose the easiest way to sensitize target faults and run dynamic compaction to reduce test set size and test power without

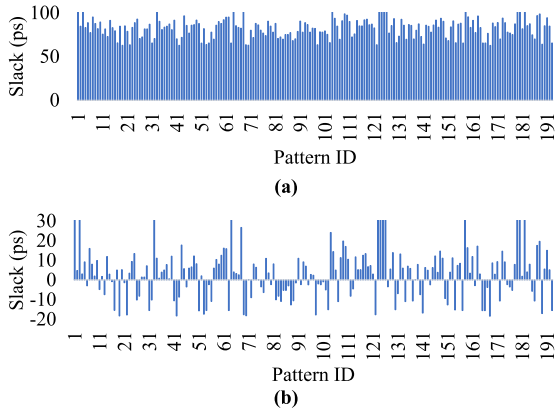

Fig. 7. Minimum slack for pattern for the min-cut-based LDPC benchmark. (a) Without voltage droop. (b) With voltage droop.

considering layout information. Prior work on capture-power reduction for scan testing use WSA for the whole design as the quality metric and optimization goal [10]–[12]. These methods do not provide an optimal solution for M3-D designs since they do not guarantee the minimum WSA for the top tier. Therefore, an M3-D-specific pattern generation algorithm is needed to mitigate the voltage droop problem during scan capture for M3-D designs; this method must consider the M3-D layout information.

C. Identification of Patterns That Cause Yield Loss

We next extract test patterns that cause excessive voltage droop, resulting in negative slack on sensitized paths. These patterns are likely to result in yield loss. During scan shifting, the only requirement to prevent yield loss is to guarantee the timing requirement between two scan FFs. In our experiments, the minimum slack in the scan shift is sufficient to endure the voltage droop problem. Therefore, this article focuses on mitigating the PSN-induced yield loss during scan capture. To calculate the additional delay due to voltage droop, we utilize a scale factor under the assumption that gate delay is computed using a first-order model based on which varies with the supply voltage and the velocity saturation $\alpha \approx 1$ in the nanometer regime [27]. Let V_{droop} be the voltage droop obtained during the dynamic rail analysis. The scaled delay T_{droop} is calculated as follows:

$$T_{\text{droop}} = T_{\text{nom}} \times \frac{1 - \frac{V_{\text{th}}}{V_{\text{nom}}}}{1 - \frac{V_{\text{th}}}{(V_{\text{nom}} - V_{\text{droop}})}} \tag{2}$$

where T_{nom} is the delay without the voltage droop, V_{nom} is the nominal supply voltage, and V_{th} is the threshold voltage. In our experiments, V_{nom} is 1.1 V and V_{th} is 0.15 V according to the Nangate 45-nm Open Process Design Kit. Increased delay of sensitized paths for a pattern leads to a reduction of the minimum slack for these paths. Once the slack becomes negative, the corresponding path violates the setup time violation, which may cause an erroneous response at the output (or scan FF) and hence result in the failure of a good chip and yield loss.

Note that we conduct timing analysis on the whole design, instead of each tier separately. A negative slack implies timing violations, while a positive slack is desirable to minimize yield loss. Fig. 7 compares the minimum slack of patterns with and without the impact of voltage droop during scan capture for the min-cut-based LDPC design. The voltage droop applied to each gate is obtained from the worst case scenario. Since we have established that the top-tier WSA is highly correlated with the PSN-induced voltage droop in Section IV-B, we first conduct dynamic power and rail analysis on the pattern with the largest top-tier WSA to obtain the worst case voltage droop. Next, this voltage droop is utilized to scale the delay of every gate in the design using (2). As shown in Fig. 7, a large proportion of patterns have negative slack after delay scaling. Therefore, pattern reshaping is necessary to mitigate the yield-loss problem.

TABLE IV
ANALYSIS OF ATPG-GENERATED PATTERNS

Design	# Patterns (Total)	# Patterns to be reshaped	Percentage of patterns to be reshaped
Min-cut based benchmarks			
LDPC	194	74	38.14%
Tate	558	117	20.93%
netcard	43103	451	1.05%
leon3mp	17087	258	1.51%
GNN-based benchmarks			
LDPC	203	50	24.63%
Tate	592	89	15.03%
netcard	43423	0	0.00%
leon3mp	17158	0	0.00%

Since it is time-consuming to obtain the voltage droop for each pattern, we apply the largest voltage droop during testing to every pattern assuming a worst case scenario. This is a conservative strategy that minimizes voltage droop during scan capture. The slack reports are obtained by conducting timing analysis using Synopsys PrimeTime. As shown in Fig. 7(b), 38% of the 194 patterns for LDPC generated by a commercial ATPG tool have a negative slack after the delay is scaled. Such patterns are identified to be susceptible to yield loss and imported into our reshaping algorithm for mitigating the voltage droop during scan capture.

Table IV shows the number and percentage of patterns that are likely to lead to yield loss for each design. At least 15% of original patterns for Tate cause yield loss, which is clearly unacceptable. Note that for GNN-based netcard and leon3mp, there is no pattern that needs to be reshaped. This is because patterns generated using the flow for transition-delay faults tend to sensitize paths that are much shorter than the critical path in static timing analysis. The slack of such paths remains positive after scaling with the worst case voltage droop using (2). In such cases, no pattern is susceptible to yield loss after yield-loss assessment. However, it is likely that long paths have negative slack due to voltage droop but they are not sensitized, which may result in test escapes. To compensate this situation, we customize our test generation flow to aggressively sensitize paths with marginal slack. We first conduct static timing analysis to capture long paths in the target circuit. For a tradeoff between runtime and fault coverage, top 2000 long paths are selected for the pattern generation. Next, we use a commercial ATPG tool to generate

TABLE V
ANALYSIS OF CUSTOMIZED PATTERNS WITH GNN-BASED BENCHMARKS

Design	# Patterns with the selected paths	# Patterns to be reshaped	# Top-off patterns	Fault coverage
netcard	84	17 (20.24%)	42567	96.80%
leon3mp	384	55 (14.32%)	17053	98.81%

patterns to detect delay faults through the selected paths. Yield-loss assessment is conducted on such patterns to identify which patterns are needed to be reshaped. Finally, we run a top-off ATPG process to detect remaining faults that cannot be propagated through the selected paths. Table V shows the number of patterns with our customized flow and the number of patterns that need to be reshaped. Around 20% and 14% of patterns generated with the selected paths cause yield loss for the GNN-based netcard and leon3mp, respectively. Therefore, a pattern reshaping procedure is necessary to obtain a new set of patterns with low dynamic voltage droop.

V. PATTERN RESHAPING

In this section, we describe pattern reshaping based on two approaches, ILP and simulated annealing. We first remove the extracted patterns from the original set and update the fault list. Next, ATPG is carried out to generate new patterns for undetected faults with don't-care bits unfilled. During pattern reshaping, our goal is to fill don't-care bits in each test pattern such that the voltage droop is minimized during scan capture.

A. ILP-Based Solution

The first step in ILP modeling is to declare all nets in the circuit to be binary variables. Since there are two vectors V^1 and V^2 for the initial state and the launch state in a transition delay fault pattern, respectively, we declare two variables n^1 and n^2 to represent signals of net n with V^1 and V^2, respectively. Let G be all the standard cells and N be all the nets in the circuit. We define two sets $S^1 = \{n^1 \mid \forall n \in N\}$ and $S^2 = \{n^2 \mid \forall n \in N\}$. Next, the functionality of each Boolean logic gate is realized by a set of linear constraints. Table VI shows the linear inequalities for four basic logic gates, where the inputs are denoted as x_i and the output is denoted as y. Every standard cell g in benchmark designs can be realized by a combination of the listed gate types. Note that the M3-D benchmarks are all full-scan circuits. The FF states are determined by the corresponding values in vectors V^1 and V^2. Therefore, we do not need to model sequential logic using linear constraints in our ILP model.

The constraints associated with g can be expressed in a canonical form as follows:

$$A_g^i \mathbf{x}_g^i \leq \mathbf{b}_g^i \tag{3}$$

where $i \in \{1, 2\}$ indicates that this constraint corresponds to vector V^i, $A_g^i \in \mathcal{R}^{m \times n}$, and $\mathbf{b}_g^i \in \mathcal{R}^m$ are the matrix and the vector of real numbers used to represent the functionality of cell g, and $\mathbf{x}_g^i \in (S^i)^n$ is a vector containing the variables associated with the fan-in net and the fan-out net. Both A_g^i and b_g^i can be easily derived from Table VI based on the gate type, while $\mathbf{x}_g^i$ depends on the topology of the input netlist.

TABLE VI
CONSTRAINTS THAT INCORPORATE THE FUNCTIONALITY OF LOGIC GATES

Gate Type	Linear Inequalities
AND	$-y + x_1 + x_2 \leq 1$ $y - x_1 \leq 0$ $y - x_2 \leq 0$
OR	$y - x_1 - x_2 \leq 0$ $-y + x_1 \leq 0$ $-y + x_2 \leq 0$
XOR	$y - x_1 - x_2 \leq 0$ $-y + x_1 - x_2 \leq 0$ $-y - x_1 - x_2 \leq 0$ $y + x_1 + x_2 \leq 2$
INV	$y + x_1 \leq 1$ $-y - x_1 \leq -1$

After the circuit is modeled using the constraints described above, the objective function is formulated. To evaluate the switching activity of net n, we define a binary variable $n_{\text{toggle}} = n^1 \oplus n^2$. The XOR gate is used because it can demonstrate the transition states at net n. If net n is switching, n^1 and n^2 must have opposite values, that is, n^1 equals to 0(1) and n^2 equals to 1(0), making the output of the XOR gate become 1. Therefore, whenever n_{toggle} equals 1, a transition occurs at net n. The constraints for n_{toggle} can be expressed as shown below

$$A_{\text{XOR}} \mathbf{x}_n \leq \mathbf{b}_{\text{XOR}} \tag{4}$$

where $A_{\text{XOR}} \in \mathcal{R}^{4\times 3}$ and $\mathbf{b}_{\text{XOR}} \in \mathcal{R}^4$ are the matrix and vector to realize an XOR gate, and $\mathbf{x}_n = \{n_{\text{toggle}}, n^1, n^2\}$. As discussed in Section IV-B, the dynamic voltage droop is greatly influenced by switching activities in the top tier. Such an influence is not considered when patterns are reshaped for conventional 2-D ICs. Two-dimensional X-filling algorithms [10]–[12] aim at minimizing transition states at scan FFs without considering the toggling of combinational logic gates. These algorithms may occasionally trigger large switching activities in the top tier, increasing the probability of yield loss induced by high voltage droop. Therefore, a new algorithm is needed for M3-D designs that takes M3-D layout information and combinational gates in the top tier into consideration. In our ILP model, the goal is to minimize the weighted switching activities in the top tier. Let a subset $N_{\text{top}} = \{n \in N \mid n \text{ belongs to the top tier}\}$. The objective function is shown as

$$\min \sum_n w_n n_{\text{toggle}} \quad \forall n \in N_{\text{top}} \tag{5}$$

where w_n is the weight of net n and equal to 1 plus the number of its fan-out gates. The overall ILP model is represented as follows:

$$\begin{aligned} \min \quad & \sum_n w_n n_{\text{toggle}} \quad \forall n \in N_{\text{top}} \\ \text{s.t.} \quad & A_g^1 \mathbf{x}_g^1 \leq \mathbf{b}_g^1 \quad \forall g \in G \\ & A_g^2 \mathbf{x}_g^2 \leq \mathbf{b}_g^2 \quad \forall g \in G \\ & A_{\text{XOR}} \mathbf{x}_n \leq \mathbf{b}_{\text{XOR}} \quad \forall n \in N_{\text{top}}. \end{aligned} \tag{6}$$

However, modeling large designs completely with variables in this manner requires an enormous number of constraints and consumes considerable runtime during optimization. To relax

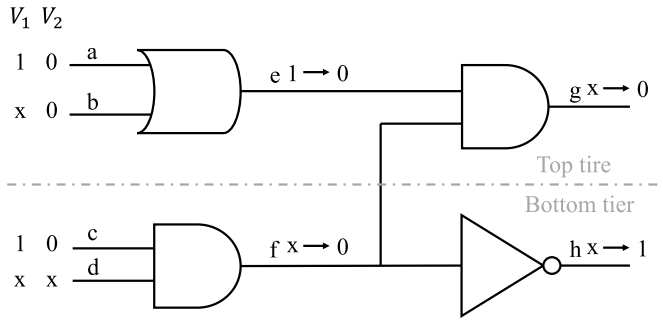

Fig. 8. Example circuit used to illustrate the ILP model.

the constraints, we first carry out a forward implication with test patterns. Only inputs and gate outputs with an unknown signal are included in the ILP model. Note that for a delay-fault test pattern, both the initial state and the launch state need to be taken into consideration. An example circuit is shown in Fig. 8. With test vectors V_1 and V_2, we perform forward implication twice by applying two vectors contiguously. In this case, only $(b^1, d^1\ f^1, g^1, h^1, d^2)$ will be defined as binary variables in our ILP model, and the constraints corresponding to them are formulated using the linear inequalities mentioned above. Next, nets a, b, e, g are located in the top tier. After constraints relaxation, it is obvious that a_{toggle} and e_{toggle} have already been determined. The inclusion of these two variables in the objective function is unnecessary. Therefore, the objective function is formulated as: Minimize $\{2b_{\text{toggle}} + 2g_{\text{toggle}}\}$. The complete ILP model for this example is described below

$$
\begin{aligned}
&\min \quad 2b_{\text{toggle}} + 2g_{\text{toggle}} \\
&\text{s.t.} \\
&\quad -f^1 + d^1 \le 0 \\
&\quad f^1 - d^1 \le 0 \\
&\quad -g^1 + f^1 \le 0 \\
&\quad g^1 - f^1 \le 0 \\
&\quad h^1 + f^1 \le 1 \\
&\quad -h^1 - f^1 \le -1 \\
&\quad b_{\text{toggle}} - b^1 \le 0 \\
&\quad -b_{\text{toggle}} + b^1 \le 0 \\
&\quad b_{\text{toggle}} + b^1 \le 2 \\
&\quad g_{\text{toggle}} - g^1 \le 0 \\
&\quad -g_{\text{toggle}} + g^1 \le 0 \\
&\quad g_{\text{toggle}} + g^1 \le 2 \\
&\quad \text{Binary variables}: \ b^1, d^1 f^1, g^1, h^1, d^2, b_{\text{toggle}}, g_{\text{toggle}}. \quad (7)
\end{aligned}
$$

The solution to this ILP problem provides a fully specified test pattern with minimum WSA for the top tier. This ILP model is invoked for every test pattern that needs to be reshaped.

B. Simulated Annealing

The ILP-based algorithm is guaranteed to reshape each pattern with the minimum WSA value for the top tier. However, the high runtime is a problem for large designs or when a large number of patterns need to be reshaped. To ensure scalability for large designs, we present another reshaping algorithm based on simulated annealing [28]. The ILP-based algorithm can reshape patterns with optimal solutions, but it needs to find a tradeoff between performance and runtime by dropping patterns exceeding the runtime limit, which results in a loss of fault coverage. The simulated annealing-based algorithm can finish execution in a relatively short amount of time, but it may reshape patterns with suboptimal solutions, leading to a greater reduction of slack than patterns reshaped by the ILP-based algorithm. Therefore, the ILP-based solution is suitable for circuits that have small slack margins but can accept a slightly loss of fault coverage, while the simulated annealing-based solution can be applied when the design is large or the slack margin is not tight. In addition, optimal results from the ILP approach can be used to assess the quality of the heuristic solution for smaller designs.

Fig. 9 sketches the steps involved in the reshaping process. We utilize an inhomogeneous annealing schedule in our algorithm, that is, the temperature decreases after a change of state. The initial temperature t is a user-defined constant that is independent of the design benchmark. The number of steps n_s controls the decreasing rate of temperature, which can be fine-tuned for each design to find a tradeoff between runtime and performance. Lines 2–18 iterate through all patterns that need to be reshaped. In Line 3, the initial state is created by filling don't-care bits in the target pattern p with Algorithm [11], which has been shown to be scalable for large designs. Lines 6–16 iterate through the simulated annealing process for n_s times. Line 7 finds a neighbor of the current state by randomly choosing a bit among don't-care bits in p and convert its value from 0(1) to 1(0). Line 8 calculates the difference between the top-tier WSA values of S_i and S_{current}. In Lines 9–11, if the acceptance probability, calculated by $\exp(-\Delta E/t_{\text{current}})$, is greater than a random number between 0 and 1, the current state is updated; else, the current state remains unchanged for the next iteration.

In the early stages of optimization when t is high, the acceptance probability is close to 1. Thus, a nonimproving solution is highly likely to be accepted, helping the searching process to escape from a local minima. In the later stages of optimization, the acceptance probability becomes close to 0 due to low temperature, so the process gradually converges to an equilibrium state. Note that if ΔE is negative, the current state is almost certain to be updated. Lines 12–14 changes the best state to S_i if the state S_i has the lowest top-tier WSA value up to this point. Line 15 updates the current temperature. In our algorithm, we use the linear reduction rule and set the termination condition to be $t = 0$. Therefore, the temperature is changed by decreasing t/n_s each time. In Line 17, the best state S_{best} is appended to the final pattern set P'.

The time complexity of the above simulated annealing-based algorithm can be analyzed as follows. Given a pattern set P with n_P patterns, Lines 6–16 will be executed n_s times to find the best state. For each iteration, the bottleneck occurs at Line 8, where the fault simulation is conducted to calculate the top-tier WSA of S_i. Let the number of gates in the design D be n_G. During fault simulation, each

TABLE VII
RESULTS FOR MIN-CUT-BASED BENCHMARKS WITH AND WITHOUT RESHAPING

Design	# Patterns	Mean WSA (top tier)	Standard deviation of WSA (top tier)	Fault coverage	# Paths with negative slack
Without Pattern Reshaping					
LDPC	194	58830	5818.91	99.73%	74
Tate	558	99629.05	11763.54	98.37%	117
netcard	43103	46710.27	19333.75	96.71%	451
leon3mp	17087	35241.07	9390.20	98.87%	258
With Simulated Annealing-based Pattern Reshaping					
LDPC	204	56674.65	5928.68	99.73%	0
Tate	669	81546.31	35459.48	98.37%	0
netcard	43288	46241.82	19486.18	98.71%	0
leon3mp	17624	26119.89	9441.09	98.88%	0
With ILP-based Pattern Reshaping					
LDPC	200	58150.98	5355.39	99.64%	0
Tate	443	99132.92	11503.97	97.98%	0
netcard	43288	45626.45	19892.42	96.71%	0
leon3mp	16932	35061.30	9569.05	98.81%	0

Input: A pattern set P with don't care bits unfilled, the corresponding design D, the number of steps n_s, and the temperature t
Output: A pattern set P' after reshaping
$P' := \emptyset$
for each $p \in P$
  $S_{current} \leftarrow$ Initialize(p)
  $S_{best} \leftarrow S_{current}$
  $t_{current} \leftarrow t$
  for i = 1 to n_s
    $S_i \leftarrow$ Neighbor($S_{current}$)
    $\Delta E \leftarrow$ Top-tierWSA(D, S_i) - Top-tierWSA(D, $S_{current}$)
    if exp($-\Delta E/t_{current}$) > rand()
      $S_{current} \leftarrow S_i$
    end if
    if Top-tierWSA(D, S_i) < Top-tierWSA(D, S_{best})
      $S_{best} \leftarrow S_i$
    end if
    $t_{current} \leftarrow$ UpdateTemperature($t_{current}$)
  end for
  $P' := P' \cup S_{best}$
end for
return P'

Fig. 9. Pseudo-code for the simulated annealing-based reshaping algorithm.

gate is evaluated twice for a transition-delay test pattern, hence the time complexity is $\mathcal{O}(n_G)$. The other steps can be completed in constant time. Therefore, the total time complexity is $\mathcal{O}(n_P n_s n_G)$.

VI. EXPERIMENTAL RESULTS

We implemented a program in Python to formulate the ILP models for patterns with don't-care bits. We utilized the Python application programming interface and the ILP solver of the Gurobi optimizer [29]. Our code was run on a 64-bit Linux Server with a 10-core Intel Xeon 2.40 GHz CPU and 12 GB memory.

Even though the number of variables and constraints of the ILP model are linear in the size of circuit, runtime overhead is a major concern for large designs. To find a tradeoff between performance and efficiency, we define a runtime threshold to compare the results of each reshaping algorithm within the same timing constraint. In our experiments, the runtime limit is 0.5 h for each pattern and 48 h for the whole reshaping process. With a negligible loss of fault coverage, patterns that exceed the runtime limit are removed during optimization. The CPU time required to generate the ILP model is negligible.

A. Min-Cut-Based Benchmarks

Table VII shows the test pattern results obtained after reshaping for the min-cut-based benchmarks. It is expected that the number of patterns increases slightly due to the lack of pattern compaction when we generate a new set with don't-care bits before the optimization. For the Tate and leon3mp benchmarks with the ILP-based reshaping algorithm, our solution leads to a reduction in the number of test patterns, with a negligibly small decrease in fault coverage. For all our benchmarks, the average WSA of patterns decreases with pattern reshaping without any adverse impact on fault coverage.

Note that for the original ATPG-generated pattern set, many sensitized paths have negative slack due to voltage droop. These paths are likely to lead to yield loss. In Table VII, we list one path for each test pattern; this is the path with the minimum slack for the corresponding pattern. We consider the worst case scenario and extract patterns with negative minimum slack and then use our approaches to reshape such patterns. If such a path has positive slack, we can ensure that no sensitized path has negative slack; hence, yield loss is eliminated. We leave the patterns that only sensitize paths with positive slack unchanged and extract the other patterns from the original pattern set. Next, we reshape the extracted patterns with our proposed ILP-based and simulated annealing-based optimization methods. A timing-analysis verification step is carried out to compute the slack after pattern reshaping. As shown in Table VII, all sensitized paths have positive slack for the reshaped patterns.

Note, however, that the reduction in average WSA value is not so noticeable. This is because the average WSA is dominated by patterns in the original test set that have a large WSA but only sensitize paths without a small slack. Those patterns are unlikely to fail a good chip due to the voltage droop problem and thus these patterns do not have to be reshaped.

We next compare the reshaped patterns with a 2-D baseline X-filling algorithm [11]; the results are shown in Table VIII.

TABLE VIII
COMPARISONS BETWEEN PROPOSED RESHAPING METHODS WITH A 2-D BASELINE X-FILLING ALGORITHM [11] FOR MIN-CUT-BASED BENCHMARKS

Design	# Patterns reshaped	Fault coverage	# Patterns having paths with negative slack	# Patterns with min. slack below 3% of the clock period	CPU time
[11]					
LDPC	84	99.73%	0	7	5m
Tate	228	98.37%	1	18	20m
netcard	636	96.71%	47	98	17m 22s
leon3mp	795	98.88%	7	10	31m 7s
Simulated Annealing-based Optimization					
LDPC	84	99.73%	0	0	3h 20m
Tate	228	98.37%	0	0	2h 26m
netcard	636	96.71%	0	31	5h 28m
leon3mp	795	98.88%	0	3	9h 57m
ILP-based Optimization					
LDPC	80	99.64%	0	0	9h 10m
Tate	2	97.98%	0	0	48h
netcard	636	96.71%	0	6	48h
leon3mp	103	98.81%	0	0	48h

We first carry out timing analysis to evaluate the reduction in yield loss after pattern reshaping. We also consider sensitized paths with the minimum slack for each pattern and record the number of paths whose slack is no more than 3% of the functional clock period.

In Table VIII, the number of reshaped patterns is obtained based on the ATPG generation process. We update the undetected fault list after extracting patterns from the original set and regenerate a new set of patterns with don't-care bits unfilled. Therefore, this number is different from the number of patterns-to-be-reshaped listed in Table IV due to the lack of pattern compaction. Note that for the results with the ILP-based optimization method, the number of patterns is lower in Table VIII due to optimizations carried out to reduce runtime. For Tate and leon3mp, the drop in the number of patterns is large because the reshaping procedure is terminated when it reaches the runtime limit, which is 48 h in our experiments. During ILP optimization, we remove patterns after timeout and patterns that require high runtime for analysis and evaluate the remaining patterns. Since the loss of fault coverage is negligible, further optimization does not have to be conducted. For the simulated annealing-based optimization method, all benchmarks can finish execution within the runtime limit without any loss of fault coverage.

In [11], a probabilistic method is presented to ensure low capture power without any loss of fault coverage. However, the randomness in this method tends to occasionally sensitize paths with a small slack. Moreover, because [11] does not take the M3-D layout and PDN information into account, it can lead to considerable voltage droop for some test patterns during scan capture. As a result, the patterns obtained from [11] can lead to either negative slack or considerably reduced slack margin. Negative slack will always lead to yield loss, while reduced slack will magnify the detrimental impact of small-delay defects and also likely to lead to the failure of a good chip with small process variations. In the proposed methods, the reshaped patterns do not cause negative slack under voltage-droop conditions on sensitized paths and the problem of slack-margin reduction is also mitigated. Also, Table VIII shows that the proposed methods lead to fewer sensitized paths with reduced slack margins compared with the baseline.

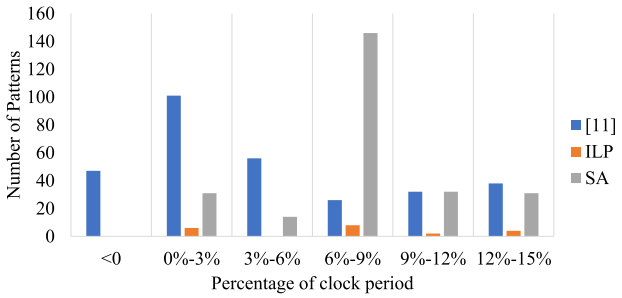

Fig. 10. Minimum slack distribution with voltage droop for the reshaped patterns for the min-cut-based netcard benchmark, where SA is the simulated annealing-based algorithm.

We next evaluate the impact on slack of different optimization methods. Figs. 10 and 11 show the slack distributions for the reshaped patterns with respect to the percentage of the clock period for the min-cut-based netcard and leon3mp benchmark, respectively. Note that patterns with minimum slack larger than 15% of the clock period are not of concern because such patterns are unlikely to result in yield loss due to small-delay defects or process variations. Among the three optimization methods, the ILP-based solutions provide the greatest improvement in the slack margin. The slack distributions of the baseline solutions [11] are more similar to the proposed simulated annealing-based solutions than to the ILP-based solutions. This is reasonable because during simulated annealing, we utilize the baseline results as initial states. It is likely that the best solutions the heuristic can find are local minima near initial states. However, for all the benchmarks in Table VIII, the simulated annealing-based method can get better solutions compared to the baseline.

B. GNN-Based Benchmarks

Table IX shows the test pattern results obtained after reshaping for the GNN-based benchmarks. Note that for netcard and leon3mp, the pattern sets are generated by our customized flow, as described in Section IV-C. The proposed

TABLE IX
RESULTS FOR GNN-BASED BENCHMARKS WITH AND WITHOUT RESHAPING

Design	# Patterns	Mean WSA (top tier)	Standard deviation of WSA (top tier)	Fault coverage	# Paths with negative slack
Without Pattern Reshaping					
LDPC	203	39142.29	3282.08	99.74%	50
Tate	592	107900.45	12544.87	98.26%	89
netcard	42651	53880.73	25361.12	96.80%	17
leon3mp	17437	38673.78	9903.05	98.81%	55
With Simulated Annealing-based Pattern Reshaping					
LDPC	218	37084.16	4053.50	99.74%	0
Tate	659	90981.13	34217.86	98.26%	0
netcard	42651	53848.34	25371.23	96.80%	0
leon3mp	17705	38519.45	10149.40	98.81%	0
With ILP-based Pattern Reshaping					
LDPC	218	37919.70	3323.10	99.74%	0
Tate	509	107432.23	14146.12	98.18%	0
netcard	42651	53584.62	25373.49	96.80%	0
leon3mp	17389	38833.83	9900.18	98.80%	0

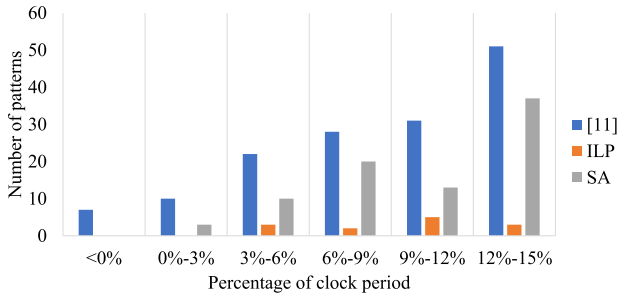

Fig. 11. Minimum slack distribution with voltage droop for the reshaped patterns for the min-cut-based leon3mp benchmark, where SA is the simulated annealing-based algorithm.

reshaping methods can mitigate the yield-loss problem for all the benchmarks. Note that with ILP-based pattern reshaping, the drop in fault coverage for Tate is larger than for other benchmarks. This is because during reshaping, few patterns can finish within the runtime limit, that is, 0.5 h for each pattern. However, the loss on fault coverage is only 0.08%, which is acceptable.

Table X shows the comparisons for the reshaped patterns between the baseline algorithm [11] and the proposed methods for the GNN-based benchmarks. For netcard and leon3mp, the number of patterns with minimum slack below 3% of the clock period is higher than that for LDPC and Tate. This is expected because patterns for netcard and leon3mp are generated by our customized flow. The sensitized paths tend to have marginal slack even without considering voltage droop. Therefore, the slack-margin reduction is not as significant as the alleviation of the yield-loss problem.

Furthermore, patterns with marginal slack are highly sensitive to PSN-induced voltage droop. A small increase in voltage droop can make positive slack become negative, leading to yield loss. As a result, reshaping algorithms require to be extremely effective to prevent yield loss. For all the benchmarks, the proposed methods can eliminate the number of patterns with negative slack, while two and one patterns have negative slack in the baseline solutions for netcard and leon3mp, respectively. Although the proposed methods require higher CPU runtime than the baseline, such full elimination is very important. This is because even only one pattern in the pattern set has negative slack, and it always leads to yield loss. For the slack-margin reduction, the proposed solutions outperform the baseline solutions in reducing the number of patterns with slack below 3% of the clock period for the largest two benchmarks.

VII. DISCUSSION

A. Comparisons Between the Proposed Methods

The ILP-based reshaping method guarantees the minimum top-tier WSA value for each reshaped pattern. Among the three reshaping algorithms evaluated in Tables VIII and X, it always produces the best results. Furthermore, it can improve the slack margin considerably to help prevent yield loss due to small-delay defects. However, a major drawback is that a few patterns may be dropped due to runtime limits, leading to a small loss of fault coverage. We have shown that the fault coverage loss for all benchmarks is within 0.08%, which is sometimes acceptable in practice. Moreover, our method can be combined with stuck-at fault patterns to prevent test escape of real stuck-at faults due to fault coverage loss. Stuck-at faults do not affect PSN because they are used in static testing, which is not at-speed. A second concern is the scalability to large designs. The number of variables in our ILP model is proportional to the size of benchmarks. Therefore, it may take a long time when patterns are reshaped for large benchmarks and some patterns may be dropped. Nonetheless, the results obtained by ILP-based optimization are useful because they can be used to evaluate the performance of other heuristics for smaller designs.

The simulated annealing-based method guarantees that there is no drop in fault coverage, and therefore it is an alternative to ILP when even negligible drop in fault coverage is unacceptable. Furthermore, the simulated annealing-based method is scalable for large designs. Although it needs more computational effort compared to [11], the reshaped pattern set has fewer patterns with negative slack. This is very important because even a single pattern with negative slack can lead to yield loss. The disadvantage of this method is that it can only obtain suboptimal solutions. Therefore, our ILP-based method is suitable for designs that can endure a small loss of fault

TABLE X
COMPARISONS BETWEEN PROPOSED RESHAPING METHODS WITH A 2-D BASELINE X-FILLING ALGORITHM [11] FOR GNN-BASED BENCHMARKS

Design	# Patterns reshaped	Fault coverage	# Patterns having paths with negative slack	# Patterns with min. slack below 3% of the clock period	CPU time
[11]					
LDPC	65	99.74%	0	0	23s
Tate	156	98.26%	0	0	3m 20s
netcard	17	96.80%	2	3	28s
leon3mp	323	96.81%	1	91	12m 47s
Simulated Annealing-based Optimization					
LDPC	65	99.74%	0	0	1h 40m
Tate	156	98.26%	0	0	1h 53m
netcard	17	96.80%	0	2	58m 52s
leon3mp	323	96.81%	0	65	3h 51m
ILP-based Optimization					
LDPC	65	99.74%	0	0	2h 6m
Tate	6	98.18%	0	0	48h
netcard	17	96.80%	0	3	5h 52m
leon3mp	7	98.80%	0	1	48h

TABLE XI
ANALYSIS OF ATPG-GENERATED PATTERNS OF BENCHMARK MIN-CUT-BASED M3-D DESIGNS IN THE TEST-COMPRESSION ENVIRONMENT

Design	# Test inputs	# Scan chains	# Patterns	Fault coverage	# Patterns to be reshaped	Percentage of patterns to be reshaped
LDPC	5	100	187	99.62%	14	7.49%
Tate	5	100	538	99.36%	74	13.75%
netcard	10	200	44231	96.66%	486	1.10%
leon3mp	10	200	17406	98.79%	214	1.23%

coverage but have a tight slack margin, while the simulated annealing-based algorithm can be applied to the designs when absolutely no loss in fault coverage is permitted.

B. M3-D Process Variations

Process variations have a significant impact on delay testing because they lead to slack-distribution variations from chip to chip. M3-D process variations can be attributed to two key reasons: (i) low-temperature manufacturing process for top-tier transistors and (ii) stability of intermediate back-end-of-line (iBOEL). In an M3-D design, device tiers are sequentially fabricated on the same wafer. Conventional manufacturing steps (e.g., epitaxy, annealing, and dopant activation) often have high thermal budgets, which inevitably causes damages to cells and wires in the bottom tier during top-tier device fabrication. Therefore, low-temperature process is mandatory for M3-D integration. Solid-phase epitaxy regrowth (SPER) and laser annealing have been developed to successfully realize top-tier transistors without damaging the bottom-tier components. However, such processes lead to high source/drain resistance and low on-current, degrading the performance of top-tier transistors. The metal usage for iBOEL is another concern since standard copper/low-k materials have a high risk of contamination when fabricating the top tier. A high-temperature anneal may increase the sheet resistance and leakage current [30].

The above M3-D process variation issues have been largely resolved in recent years due to breakthroughs in 3-D integration technology. For example, it has been demonstrated in [31] that high-performance fully depleted silicon-on-insulator (FDSOI) transistors can be fabricated at a temperature below 500 °C. In [32], copper/low-k iBOEL has been shown to be stable and reliable under the standard 28-nm design rules. These breakthroughs make the impacts of process variations in M3-D ICs similar to that for conventional 2-D designs. Solutions have been proposed in the literature or adopted in practice to handle such variations [33], [34], and these solutions can be easily used for M3-D designs. Therefore, we do not consider process variations in our analysis and pattern reshaping approaches.

C. Pattern Reshaping With Test Compression

Test compression is widely used today to achieve a significant reduction in test time and data volume. To discuss the PSN-induced yield loss in a test-compression environment, we insert compression designs into our min-cut-based benchmarks using Synopsys Testmax. Table XI shows the number and percentage of patterns that are required to be reshaped in the compression environment. Up to 13% of original patterns for Tate cause yield loss, which is unacceptable. Therefore, pattern reshaping is required to mitigate PSN-induced voltage droop in M3-D designs with test compression. However, X-filling reshaping algorithms are not compatible with test compression due to the difficulty of finding don't-care bits in a compressed pattern.

To conduct pattern reshaping in the test compression environment, we extend our proposed simulated-annealing-based method. During the simulated annealing process, we change the way in which we find a neighbor of the current state in a compressed pattern. Instead of searching for a don't-care bit, we randomly choose a bit among test inputs and convert its value from 0 (1) to 1 (0). This modification may occasionally influence the sensitized paths after reshaping, leading to a loss of fault coverage. To compensate for the loss, we run a top-off ATPG process following simulated annealing. Finally, yield-loss assessment is conducted on the reshaped patterns

TABLE XII
RESULTS FOR MIN-CUT-BASED M3-D DESIGNS IN THE TEST-COMPRESSION ENVIRONMENT AFTER RESHAPING

Design	# Total patterns	Loss of fault coverage due to reshaping	# Top-off patterns	Fault coverage with top-off and reshaped patterns	# Patterns with negative slack	Test time increase
LDPC	221	0.12%	34	99.63%	0	18.18%
Tate	641	0.07%	101	99.36%	0	18.70%
netcard	44362	0.01%	131	96.66%	0	0.29%
leon3mp	17638	0.02%	214	98.82%	0	1.33%

and top-off patterns to evaluate the effectiveness of yield-loss mitigation. Pattern reshaping results with test compression are shown in Table XII. PSN-induced yield loss can be fully eliminated for all the benchmarks. Note that the loss of fault coverage due to the simulated-annealing-based reshaping process is extremely low; therefore, few paths need to be sensitized by each top-off pattern, leading to a low value of WSA. We also fine-tuned the merging step during pattern generation to ensure that the WSA value of every top-off pattern is low enough to prevent yield loss. Hence, while the proposed solution does not guarantee that the top-off patterns will not lead to any increase in the WSA, we have incorporated optimization steps to ensure that the yield-loss problem is avoided. An increase in test time due to the top-off patterns is another concern, but this increase is negligible for the two largest benchmarks. Moreover, in view of the high compression ratio, this increase might be acceptable. As part of ongoing work, we are assessing new reshaping algorithms that are specifically aimed at M3-D designs with test compression.

VIII. CONCLUSION

We have proposed a framework to conduct dynamic power and rail analysis for M3-D ICs. We have demonstrated that the magnitude of the voltage-droop problem in scan test mode depends on the switching activities in the top tier of a two-tier design. We have also shown how we can identify test patterns that are likely to fail a fault-free chip, that is, cause yield loss, due to the droop-induced added delay on sensitized paths. We have presented an ILP-based X-filling algorithm and a simulated annealing-based algorithm for M3-D pattern reshaping. Experimental results for OpenCore and the ISPD 2012 benchmarks show that the average WSA of the top tier is reduced after pattern reshaping and there is no decrease in the slack of sensitized paths. The proposed methods significantly mitigate the PSN-induced yield-loss problem during scan capture for M3-D designs. As part of ongoing work, we are assessing our solutions for M3-D designs with more than two tiers.

REFERENCES

[1] P. Batude, T. Ernst, J. Arcamone, G. Arndt, P. Coudrain, and P.-E. Gaillardon, "3-D sequential integration: A key enabling technology for heterogeneous co-integration of new function with CMOS," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 2, no. 4, pp. 714–722, Dec. 2012.

[2] K. Chang, S. Das, S. Sinha, B. Cline, G. Yeric, and S. K. Lim, "System-level power delivery network analysis and optimization for monolithic 3-D ICs," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 4, pp. 888–898, Apr. 2019.

[3] J. Saxena *et al.*, "A case study of IR-drop in structured at-speed testing," in *Proc. Int. Test Conf. (ITC)*, 2003, p. 1098.

[4] X. Lin *et al.*, "High-frequency, at-speed scan testing," *IEEE Des. Test Comput.*, vol. 20, no. 5, pp. 17–25, Sep./Oct. 2003.

[5] P. Pant and J. Zelman, "Understanding power supply droop during at-speed scan testing," in *Proc. 27th IEEE VLSI Test Symp.*, May 2009, pp. 227–232.

[6] R. M. Chou, K. K. Saluja, and V. D. Agrawal, "Scheduling tests for VLSI systems under power constraints," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 5, no. 2, pp. 175–185, Jun. 1997.

[7] B. Yang, A. Sanghani, S. Sarangi, and C. Liu, "A clock-gating based capture power droop reduction methodology for at-speed scan testing," in *Proc. Design, Autom. Test Eur.*, Mar. 2011, pp. 1–7.

[8] S. Kajihara, K. Ishida, and K. Miyase, "Test vector modification for power reduction during scan testing," in *Proc. 20th IEEE VLSI Test Symp. (VTS)*, 2002, pp. 160–165.

[9] Y. Bonhomme, P. Girard, C. Landrault, and S. Pravossoudovitch, "Power driven chaining of flip-flops in scan architectures," in *Proc. Int. Test Conf.*, 2002, pp. 796–803.

[10] X. Wen, Y. Yamashita, S. Kajihara, L.-T. Wang, K. K. Saluja, and K. Kinoshita, "On low-capture-power test generation for scan testing," in *Proc. 23rd IEEE VLSI Test Symp. (VTS)*, May 2005, pp. 265–270.

[11] S. Remersaro, X. Lin, Z. Zhang, S. Reddy, I. Pomeranz, and J. Rajski, "Preferred fill: A scalable method to reduce capture power for scan based designs," in *Proc. IEEE Int. Test Conf.*, Oct. 2006, pp. 1–10.

[12] X. Wen *et al.*, "A novel scheme to reduce power supply noise for high-quality at-speed scan testing," in *Proc. IEEE Int. Test Conf.*, Oct. 2007, pp. 1–10.

[13] S. Wong, A. El-Gamal, P. Griffin, Y. Nishi, F. Pease, and J. Plummer, "Monolithic 3D integrated circuits," in *Proc. Int. Symp. VLSI Technol., Syst. Appl. (VLSI-TSA)*, 2007, pp. 1–4.

[14] S. A. Panth, K. Samadi, Y. Du, and S. K. Lim, "Design and CAD methodologies for low power gate-level monolithic 3D ICs," in *Proc. Int. Symp. Low Power Electron. Design*, Aug. 2014, pp. 171–176.

[15] S. Panth, K. Samadi, Y. Du, and S. K. Lim, "Placement-driven partitioning for congestion mitigation in monolithic 3D IC designs," *IEEE Trans. Comput.-Aided Design Integr.*, vol. 34, no. 4, pp. 540–553, Apr. 2015.

[16] A. Koneru and K. Chakrabarty, "An interlayer interconnect BIST and diagnosis solution for monolithic 3-D ICs," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 10, pp. 3056–3066, Oct. 2020.

[17] A. Koneru and K. Chakrabarty, "An inter-layer interconnect BIST solution for monolithic 3D ICs," in *Proc. IEEE 36th VLSI Test Symp. (VTS)*, Apr. 2018, pp. 1–6.

[18] A. Koneru, A. Todri-Sanial, and K. Chakrabarty, "Reliable power delivery and analysis of power-supply noise during testing in monolithic 3D ICs," in *Proc. IEEE 37th VLSI Test Symp. (VTS)*, Apr. 2019, pp. 1–6.

[19] W Chen, *The Electrical Engineering Handbook*. Boston, MA, USA: Academic, 2005.

[20] S. K. Samal, K. Samadi, P. Kamal, Y. Du, and S. K. Lim, "Full chip impact study of power delivery network designs in monolithic 3D ICs," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2014, pp. 565–572.

[21] J. A. Waicukauski, E. Lindbloom, B. K. Rosen, and V. S. Iyengar, "Transition fault simulation," *IEEE Design Test Comput.*, vol. DTC-4, no. 2, pp. 32–38, Apr. 1987.

[22] 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 *Proc. Int. Symp. Phys. Design*, Mar. 2018, pp. 90–97.

[23] H. Park, B. W. Ku, K. Chang, D. E. Shim, and S. K. Lim, "Pseudo-3D approaches for commercial-grade RTL-to-GDS tool flow targeting monolithic 3D ICs," in *Proc. Int. Symp. Phys. Design*, Mar. 2020, pp. 47–54.

[24] Y.-C. Lu, S. S. Kiran Pentapati, L. Zhu, K. Samadi, and S. K. Lim, "TP-GNN: A graph neural network framework for tier partitioning in monolithic 3D ICs," in *Proc. 57th ACM/IEEE Design Autom. Conf. (DAC)*, Jul. 2020, pp. 1–6.

[25] M. Ozdal, C. Amin, A. Ayupov, S. Burns, G. Wilke, and C. Zhuo. (2012). *The ISPD-2012 Discrete Cell Sizing Contest and Benchmark Suite*. [Online]. Available: http://www.ispd.cc/contests/12/ispd2012_contest.html
[26] P. Girard, "Survey of low-power testing of VLSI circuits," *IEEE Design Test Comput.*, vol. 19, no. 3, pp. 82–92, May 2002.
[27] A. Ramalingam, S. V. Kodakara, A. Devgan, and D. Z. Pan, "Robust analytical gate delay modeling for low voltage circuits," in *Proc. Asia South Pacific Conf. Design Autom.*, 2006, pp. 61–66.
[28] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.
[29] LLC Gurobi Optimization. (2020). *Gurobi Optimizer Reference Manual*. [Online]. Available: http:///www.gurobi.com
[30] A. Vandooren *et al.*, "Sequential 3D: Key integration challenges and opportunities for advanced semiconductor scaling," in *Proc. Int. Conf. IC Design Technol. (ICICDT)*, Jun. 2018, pp. 145–148.
[31] C. Fenouillet-Beranger *et al.*, "First demonstration of low temperature ($\leq 500°C$) CMOS devices featuring functional RO and SRAM bitcells toward 3D VLSI integration," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2020, pp. 1–2.
[32] L. Brunet *et al.*, "Breakthroughs in 3D sequential technology," in *IEDM Tech. Dig.*, Dec. 2018, pp. 7.2.1–7.2.4.
[33] M. Wagner and H.-J. Wunderlich, "Probabilistic sensitization analysis for variation-aware path delay fault test evaluation," in *Proc. 22nd IEEE Eur. Test Symp. (ETS)*, May 2017, pp. 1–6.
[34] S. Banerjee, A. Chaudhuri, A. Ning, and K. Chakrabarty, "Variation-aware delay fault testing for carbon-nanotube FET circuits," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 29, no. 2, pp. 409–422, Feb. 2021.

Shao-Chun Hung (Graduate Student Member, IEEE) received the B.S. degree from the National Taiwan University, Taipei, Taiwan, in 2019. He is currently working toward the Ph.D. degree in electrical and computer engineering at Duke University, Durham, NC, USA.

He was an Intern with Texas Instruments, Taipei, and NVIDIA Corporation, Santa Clara, CA, USA. His current research interests include reliability, testing, and diagnosis of monolithic 3-D integrated circuits.

Yi-Chen Lu received the B.S. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 2017, and the M.S. degree in electrical and computer engineering from Georgia Institute of Technology, Atlanta, GA, USA, in 2019, where he is currently working toward the Ph.D. degree under Professor Sung Kyu Lim's guidance.

His research focuses on devising machine learning, reinforcement learning, and graph algorithms to enhance the electronic design automation (EDA) flow for 2-D and 3-D integrated circuits (ICs).

Sung Kyu Lim (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Department of Computer Science, University of California, Los Angeles (UCLA), Los Angeles, CA, USA, in 1994, 1997, and 2000, respectively.

He joined the School of Electrical and Computer Engineering, Georgia Institute of Technology, in 2001, where he is currently a Professor. He has authored *Practical Problems in VLSI Physical Design Automation* (Springer, 2008) and *Design for High Performance, Low Power, and Reliable 3-D Integrated Circuits* (Springer, 2013). He has published more than 400 articles on 2.5-D and 3-D ICs. His research focus is on the architecture, design, test, and electronic design automation (EDA) solutions for 2.5-D and 3-D ICs. His research is featured as Research Highlight in the Communication of the Association for Computing Machinery (ACM) in January, 2014.

Dr. Lim received the National Science Foundation Faculty Early Career Development (CAREER) Award in 2006. He received the ACM SIGDA Distinguished Service Award in 2008. He was an Associate Editor of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS from 2007 to 2009 and the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS from 2013 to 2018. He received the best paper award from several conferences in EDA. His works have been nominated for the best paper award at several top venues in EDA and circuit/packaging design.

Krishnendu Chakrabarty (Fellow, IEEE) received the B.Tech. degree from Indian Institute of Technology, Kharagpur, India, in 1990, and the M.S.E. and Ph.D. degrees from the University of Michigan, Ann Arbor, MI, USA, in 1992 and 1995, respectively.

He is currently the John Cocke Distinguished Professor and the Department Chair of Electrical and Computer Engineering (ECE), and a Professor of Computer Science with Duke University, Durham, NC, USA. His current research projects include design-for-testability of integrated circuits and systems, microfluidic biochips, hardware security, neuromorphic computing systems, and artificial intelligence (AI) for healthcare.

Dr. Chakrabarty was a recipient of the National Science Foundation CAREER Award, the Office of Naval Research Young Investigator Award, the Humboldt Research Award from the Alexander von Humboldt Foundation, Germany, the IEEE Transactions on CAD Donald O. Pederson Best Paper Award (2015), the IEEE Transactions on VLSI Systems Prize Paper Award (2021), the Association for Computing Machinery (ACM) Transactions on Design Automation of Electronic Systems Best Paper Award (2017), multiple IBM Faculty Awards and HP Labs Open Innovation Research Awards, and over a dozen best paper awards at major conferences. He is also a recipient of the IEEE Computer Society Technical Achievement Award in 2015, the IEEE Circuits and Systems Society Charles A. Desoer Technical Achievement Award in 2017, the IEEE Circuits and Systems Society Vitold Belevitch Award in 2021, the IEEE-KHN Asad M. Madni Outstanding Technical Achievement and Excellence Award in 2021, the Semiconductor Research Corporation Technical Excellence Award in 2018, and the IEEE Test Technology Technical Council Bob Madge Innovation Award in 2018. He was a 2018 recipient of the Japan Society for the Promotion of Science (JSPS) Invitational Fellowship in the "Short Term S: Nobel Prize Level" category. He is a Research Ambassador of the University of Bremen, Bremen, Germany, and he was a Hans Fischer Senior Fellow at the Institute for Advanced Study, Technical University of Munich, Munich, Germany, from 2016 to 2019. He is a fellow of ACM, a fellow of the American Association for the Advancement of Science (AAAS), and a Golden Core Member of the IEEE Computer Society. He was a Distinguished Visitor of the IEEE Computer Society from 2005 to 2007 and from 2010 to 2012, a Distinguished Lecturer of the IEEE Circuits and Systems Society from 2006 to 2007 and from 2012 to 2013, and an ACM Distinguished Speaker from 2008 to 2016. He served as the Editor-in-Chief of *IEEE Design & Test of Computers* from 2010 to 2012, *ACM Journal on Emerging Technologies in Computing Systems* from 2010 to 2015, and IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS from 2015 to 2018.