Advances in Testing and Design-for-Test Solutions for M3D Integrated Circuits*

Sanmitra Banerjee, Arjun Chaudhuri, Shao-Chun Hung, and Krishnendu Chakrabarty
Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA

***Abstract*—Monolithic 3D (M3D) integration has the potential to achieve significantly higher device density compared to TSV-based 3D stacking. Sequential integration of transistor layers enables high-density vertical interconnects, known as inter-layer vias (ILVs). However, high integration density and aggressive scaling of the inter-layer dielectric make M3D integrated circuits especially prone to process variations and manufacturing defects. We explore the impact of these fabrication imperfections on chip-performance and present the associated test challenges. We introduce two M3D-specific design-for-test solutions – a low-cost built-in self-test architecture for the defect-prone ILVs and a tier-level fault localization method for yield learning. We describe the impact of defects on the efficiency of delay fault testing and highlight solutions for test generation under constraints imposed by the 3D power distribution network.**

I. Introduction

Conventional two-dimensional (2D) integrated circuits (ICs) face challenges associated with scalability, performance, and high power consumption at nanometer technology nodes. Novel architectures that enable 3D integration using high-density vertical interconnects are therefore being actively studied as promising alternatives to extend Moore's law. Several 3D integration technologies have been proposed in recent years, including though-silicon-via (TSV) based 3D [1], face-to-face bonded 3D [2], and monolithic 3D [3]. Among these, monolithic 3D designs utilize nanoscale inter-layer vias (ILVs) to enable massive vertical integration (up to 100 million/mm^2) with negligible RC delays and silicon area overhead [4].

Despite these benefits, a number of test challenges need to be addressed before M3D integration can become ready for commercial exploitation. As the thickness of the inter-layer dielectric scales down at advanced nodes, the nanoscale ILVs become prone to defects and electrostatic coupling between two layers is observed. Process variations and manufacturing defects originating in the immature fabrication flow results in timing variations on critical paths. This, in turn, degrades the efficiency of delay fault testing. In addition, power supply noise during the scan capture can lead to yield loss. In this paper, we explore these issues and highlight solutions to improve the testability of M3D ICs.

The remainder of the paper is organized as follows. In Section II, we present challenges in the immature M3D fabrication flow and review prior work on testing and DfT methods to address these challenges. Section III describes a built-in self-test (BIST) architecture that requires only two test patterns to detect stuck-at faults, hard shorts, and hard opens in ILVs. This solution is then extended to an improved architecture that guarantees minimal ILV fault masking. In Section IV, we analyze the impact of process variations and manufacturing defects on the testing of small delay defects. A power supply noise-aware delay testing method is presented in Section VI. We draw conclusions in Section VII.

II. Challenges in the M3D Fabrication Process

In the first step of the M3D fabrication flow, a standard high-temperature process is used to integrate the transistors and interconnects in the bottom tier. A thin inter-layer dielectric is then deposited followed by low-temperature molecular bonding of the silicon-on-insulator (SOI) substrate to obtain the top tier [5]. Finally, the ILVs are fabricated to connect the top and bottom tiers. These steps are repeated for the fabrication of additional layers.

Dense M3D integration can lead to routing congestion in the bottom tier, increased wire length, and thermal hotspots. Additionally, during the fabrication of the top tier, care must be taken not to damage the underlying interconnects and bottom-tier transistors. This is critical, especially for the dopant activation step, which is usually performed at temperatures higher than 800 °C. Novel process flows, like the ones proposed in [6] need to be used to achieve dopant activation at temperatures below 400 °C. Additionally, tungsten (W) or cobalt (Co) interconnects, which can endure higher temperatures, are often used for bottom-tier interconnects [7]. The low-temperature top-tier fabrication and the high-resistivity W/Co bottom-tier interconnects can lead to inter-tier performance variation. This needs to be taken into consideration during tier partitioning and ILV routing, especially for designs with tight timing requirements.

Aggressive scaling of the ILD thickness at the nanometer nodes can result in inter-tier coupling and timing variations. Under coupling, the threshold voltage (V_{th}) of top-tier transistors can vary significantly from their nominal value. Simulation results show that while this shift can be as high as 65 mV for ILD thickness less than 50 nm [7]. Manufacturing defects such as voids, delaminations, and foreign particles can occur at the bond interface during the wafer-bonding step. Such defects result in a degraded back-gate dielectric capacitance of the top-tier transistors. This, in turn, leads to variations in the on-current and the propagation delay.

Recent work based on both static and dynamic analysis has shown that compared to traditional 2D designs, M3D ICs are

*This research was supported in part by the National Science Foundation under grant CCF-1908045.

2021 Design, Automation & Test in Europe Conference & Exhibition (DATE) | 978-3-9819263-5-4/21/$31.00 ©2021110.23919/DATE51398.2021.9473921

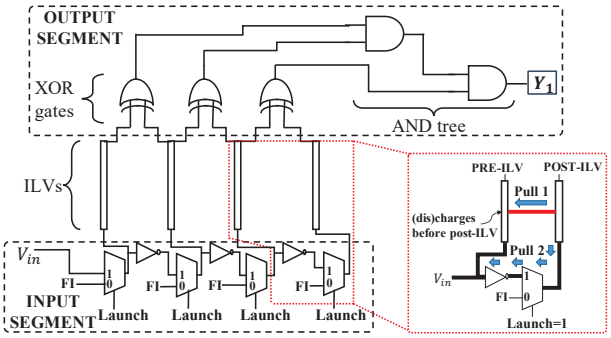

Fig. 1: Illustration of the XOR-BIST architecture.

more susceptible to power-supply noise (PSN) [8]. A major problem with the 3D power distribution network (PDN) in M3D ICs is that it can lead to high PSN during capture cycles in at-speed scan testing for transition delay faults. The PSN problem is more severe in test mode due to higher switching activities of circuit nodes compared to functional operation [9]. Therefore, the failure of good chips resulting from the PSN-induced voltage droop during scan testing is a major concern for M3D designs.

III. Built-In Self-Test for Inter-Layer Vias

Typical fault models for an ILV are shorts, opens, and stuck-at faults (SAFs) [10]. Particle contamination and metal diffusion lead to shorts. When an ILV fails to land on a contact pad, an open is created causing the ILV resistance to increase significantly.

To test ILVs, methods such as [11] deploy one scan flop per ILV, resulting in large area overhead and test time [12]. Interconnect test methods based on ATPG [13] are less effective for testing ILVs as I/O pins are present only on one tier in an M3D IC. As a result, the activated ILV faults have to be propagated through multiple ILVs and tiers, thereby increasing the risk of ILV-fault masking due to faults in the logic gates and hindering fault detection. In [14], a BIST framework has been proposed that can effectively detect single and multiple SAFs, shorts, and opens in ILVs. The proposed BIST methodology achieves nearly 100% fault coverage (both single and multiple faults) of the ILVs with only two test patterns.

A. XOR-BIST Architecture for Fault Detection

The BIST architecture to test for faults in ILVs is illustrated in Figure 1. On the output side of the ILVs, 2-input XOR gates are inserted between neighboring ILVs. For a set of N ILVs placed in a 1D array-like manner where every ILV has at most two nearest neighbors, $(N-1)$ XOR gates are inserted. The XOR outputs are fed as inputs to a space compactor which is an optimally balanced AND tree with $(N-1)$ inputs and a 1-bit output signature Y_1. By observing Y_1, it can be determined whether a fault is present in the ILVs under consideration. Test patterns are fed to the inputs of the ILVs from an input source V_{in}; V_{in} feeds an inverter chain that generates complementary signals to adjacent ILVs in the test mode. A 2:1 multiplexer is present at the input of every ILV to switch between test mode and mission mode (functional input—*FI*) based on the *Launch* signal.

The ILVs are tested in two clock cycles by switching V_{in}. The test patterns to the ILVs are "010..." ($V_{in} = 0$) and "101..." ($V_{in} = 1$) in the first and second cycles, respectively. It can be proven that a group of ILVs does not contain a hard fault if and only if Y_1 is 1 in both clock cycles. Aliasing occurs only when all ILVs are alternately stuck at 0 and 1, leading to masking of the ILV faults. However, the probability of occurrence of such a scenario is $\frac{2}{3^n}$, where n is the number of ILVs under test.

The inverter chain-based method of driving the ILVs in the test mode leads to a deterministic hard-short behavior. If a short is present between two ILVs, the ILV appearing first (pre-ILV) in the path of the incoming test signal from V_{in}, via the inverter chain, will drive the other ILV (post-ILV); this is illustrated in the inset of Figure 1. It is because the short provides a path of lower resistance ("pull 1") compared to the path through the multiplexer and inverter ("pull 2").

B. Dual-BIST Architecture

The BIST design described in Section III.A may be affected by SAFs, which in turn can potentially mask ILV fault(s). To reduce the likelihood of masking, a second propagation path is added from the ILV outputs to a 1-bit signature Y_2. The topology of this path to Y_2 (BIST-B) is identical to that of the path from the ILV outputs to Y_1 (BIST-A). The XOR and AND gates in BIST-A are substituted with the corresponding logical dual gates (XNOR and OR, respectively) in BIST-B. The ILVs under test, along with the "dual-BIST" engine, are considered to be fault-free if and only if $Y_1 = 1$ and $Y_2 = 0$ for both test patterns. With the "dual-BIST" architecture, it can be proven that a single fault in the dual-BIST engine cannot mask ILV fault(s). Furthermore, the probability of masking due to multiple faults in the dual-BIST engine is negligible.

IV. Tier-level Fault Localization in M3D ICs

Existing observation-point insertion (OPI) techniques do not address the problem of fault localization in M3D ICs. Motivated by the fact emerging M3D technology is susceptible to manufacturing defects and process variations, a topology-driven algorithm is proposed in [15] for OPI on the outgoing ILVs of an M3D tier with the objective of tier-level fault localization. The candidate ILVs are selected for OPI such that the number of fault-effects propagating through those ILVs to the next tier is maximized, thereby enabling tier-level fault localization.

A. Error Propagation due to Faults in Netlist

The strategy for selecting candidate ILVs for OPI is guided by the following factors: (i) a selected ILV should have a large number of gates in its fan-in cone; (ii) the fan-in cone of a selected ILV should have little overlap with the fan-in cones of other ILVs. The above factors enable a large number of gates on a given tier of an M3D IC to be observed through the candidate ILVs, i.e., the "gate coverage" is high. They also provide the

benefit that, with every candidate ILV chosen, a large number of previously uncovered gates are made observable through that ILV. If the selected set of ILVs provides a high gate coverage, the likelihood of a large number of fault-effects propagating through those ILVs increases. As all the tiers already have scan chains inserted in them, the scan chains also act as test points and contribute to increased fault coverage and tier localization. The set of OPs on ILVs with high gate coverage enables fault localization to a particular tier, especially for fault effects which cannot be localized by scan chains.

Given a graph model G for a scan-inserted circuit under test, the following attributes always hold when ATPG is run without any backtrack limit: (1) The error due to an irredundant fault on the output stem of a node (gate) V_i in G will always be propagated to one of the potential observation points—scan cells and POs. (2) Consider a node V_i having a fan-out of m to nodes V_j $(1 \leq j \leq m)$. The effect of an irredundant fault on the output stem of V_i always propagates downstream along one of the m fan-out branches.

In a baseline OPI method, we run constrained ATPG for single stuck-line (SSL) faults with a single ILV at a time as the OP and record its fault coverage. The ILVs with higher fault coverage are selected for OPI. If a circuit contains a million SSL faults and 100K ILVs, the total ATPG run-time will be $\sim$ 22083 days. The proposed NodeRank-based OPI method significantly reduces the run-time by executing the NodeRank algorithm only once for all SSL faults in the circuit.

When errors due to multiple faults propagate simultaneously along the graph's edges, they may reach a node common to the faults' fan-out cones. Let such an overlapping node V_j have a fan-in of f_j from nodes $\{V_i; 1 \leq i \leq f_j\}$. If a fault-effect is present on the output of V_i with likelihood $p(V_i)$, the probability of the fault-effect being propagated to the output of V_j is $p(V_i) \cdot p_{i,j}$, where $p_{i,j}$ is the error-propagation likelihood along the edge connecting V_i and V_j. As errors from f_j different nodes propagate to V_j with different probabilities, they accumulate on the output of V_j. In the next iteration, the accumulated errors in V_j propagate to its fan-out nodes, accumulating in them thereafter. The accumulation and forward propagation of errors, starting from the fault sites, continue until the errors reach and accumulate in the sink nodes, i.e., scan cells or nodes connected to POs and ILVs. We next introduce the definition of *NodeRank*.

Definition 1. *For the irredundant faults on the output stems of the nodes in G, the total fault-effect accumulation in a node V_x, after t iterations of fault-effect propagation, is given by its* ***NodeRank****, denoted by $NR^t(V_x)$.*

We define $p(V_i)$ as the error accumulated in V_i before the current iteration of error propagation: $p(V_i) = NR^{t-1}(V_i)$. Initially, faults are considered to be present on the output of every node in G. As the accumulated errors in non-sink nodes always propagate downstream, non-sink nodes do not retain any errors accumulated during the previous iterations. Therefore, for a non-sink node V_j: $NR^t(V_j) = \sum_{i=1}^{f_j} NR^{t-1}(V_i) \cdot p_{i,j}$. On the other hand, the accumulated errors in sink nodes have nowhere to propagate to; hence, sink nodes retain their old error accumulation which get added to the newly accumulated errors. Therefore, for a sink node V_j: $NR^t(V_j) = NR^{t-1}(V_j) + \sum_{i=1}^{f_j} NR^{t-1}(V_i) \cdot p_{i,j}$. Eventually, when all errors have propagated to the sink nodes, the error accumulation will be zero in all the non-sink nodes; only the sink nodes will have non-zero error accumulation. The total number of iterations of error propagation is equal to the maximum logic depth, d_G, of a root node from the sink nodes, i.e., d_G is the number of levels in G.

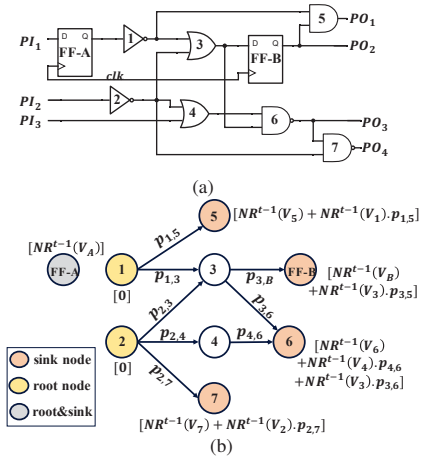

Fig. 2: (a) Example circuit and (b) corresponding circuit-graph to illustrate the NodeRank of nodes after t iterations.

During the NodeRank computation of graph nodes, the current error accumulation in a node is split among its fan-out branches, weighted by the error-propagation likelihoods on the corresponding branches. Hence, if all N_g nodes in G are initialized with NodeRank 1 (one fault per node), the total error accumulation in all the nodes after any number of iterations is always N_g. An ILV-connected sink node with higher NodeRank is a more favorable candidate for OPI.

Fig. 2 shows an example circuit and its corresponding circuit-graph. A node V_i in the graph is represented by the circle with the number i inscribed in it. The nodes V_B and V_A correspond to the flops FF-B and FF-A, respectively. Nodes V_5, V_6, V_7, and V_B are the ILV or sink nodes, as they are connected to ILVs, i.e., POs. The NodeRank $NR^t(V_x)$ of a root or sink node V_x $(x \in \{A, B, 1, 2, 5, 6, 7\})$, after t iterations of error propagation and accumulation $(t \geq 1)$, is shown inside '[]' beside V_x in Fig. 2(b). The NodeRank $NR^t(V_3)$ of the non-sink node V_3 is $NR^{t-1}(V_1) \cdot p_{1,3} + NR^{t-1}(V_2) \cdot p_{2,3}$. The NodeRank $NR^t(V_4)$ of the non-sink node V_4 is $NR^{t-1}(V_2) \cdot p_{2,4}$. The NodeRank of a node at $t = 0$ is 1. □

B. Matrix Formulation of NodeRank Computation

For achieving speed-up in computation, the NodeRank computation of the nodes is formulated as topological sorting (TPS)-based vector-vector dot-product operations. Let V be the set of all nodes in the graph G. Let NR be a $1 \times N_g$ array that stores the NodeRanks of all N_g nodes in the graph $(|V| = N_g)$. Let P_{dict} be a dictionary storing the fan-in

segment-probabilities $p_{i,j}$ for node V_j: the dictionary key is the node V_j and the corresponding value $P_{dict}(V_j)$ is a $1 \times f_j$ vector of fan-in probabilities $\{p_{i,j}; 1 \leq i \leq f_j\}$, where f_j is the fan-in of V_j. The function $TPS(G)$ returns a topologically sorted list, L_{order}, of nodes in G. The list L_{order} is then traversed to update NR. If $FI(V_j)$ is a $1 \times f_j$ list of adjacent fan-in nodes of V_j, $NR(FI(V_j))$ is a $1 \times f_j$ array storing the NodeRanks of the fan-in nodes of V_j. The NR array is updated in-place in the given manner: $NR(V_j) \leftarrow NR(V_j) + NR(FI(V_j)) \cdot P_{dict}(V_j)^T$.

After the termination of the algorithm, the NodeRanks of all non-sink nodes in the graph become 0. The worst-case time complexity of the NodeRank algorithm is $\mathcal{O}(N_g + E_g)$, where E_g is the number of edges in G. After NodeRank execution, the ILV-connected sink nodes are sorted in descending order of their NodeRanks and the top K candidate ILVs are selected for OPI from the ranked list, where K is the OP budget determined by the user based on area-overhead constraints.

C. Circuit Topology-based Splitting

We split the NodeRank of V_i equally among its m_i fan-out branches, thereby utilizing information on circuit-topology only and eliminating any gate type-related bias. The resultant error-propagation likelihoods $p_{i,j}$ are given by: $p_{i,j} = 1/m_i$. The NodeRanks of the sink nodes resulting from this equal-splitting heuristic are insightful — they indicate the effective fraction of nodes (or gates) uniquely observed by the corresponding sink nodes. Hence, we refer to a sink node's final NodeRank obtained via the equal-splitting heuristic as *uniqueness* of the sink node. Additionally, this heuristic enforces lower penalties on ILVs having high fan-out nodes in their fan-in cones. The higher the fan-out of a node, more paths are available for propagation to an error on the node's output; this implies that the node is potentially observed by multiple ILVs. If we end up not selecting a certain ILV for OPI, at least one of the other ILVs is very likely to compensate by getting selected for OPI.

V. Delay Testing in the Presence of Process Variations and Defects

Process variations and manufacturing defects result in parametric faults, increased propagation delay, and slower signal transitions. Although these small delay defects (SDDs) affect the timing slack of multiple paths through the affected cell(s), they can be detected only on paths with a small enough timing slack such that, in the presence of the fault, a signal transition is not captured within the rated clock period. Conventional SDD ATPG test patterns sensitize paths with the minimum nominal slack through each fault site. However, the timing slack on a path can vary under random process variations, coupling, and manufacturing defects. Thus, efficient SDD testing, especially in emerging technologies like M3D ICs, necessitates a variation-aware pattern generation flow.

A. Sources of SDDs in M3D ICs

Low-temperature wafer bonding is a key step in the M3D integration process. Due to increasing process variations at the nano-scale technology nodes, nanometer-sized voids are often formed in the inter-layer dielectric (ILD) during wafer bonding [10]. Such voids at the bond interface reduce the effective back-gate dielectric capacitance, which in turn affects the threshold voltage of the top-tier transistors in the M3D IC. The shift in the threshold voltage is especially significant for a low ILD thickness and leads to deviation in the ON-current and propagation delay.

Electrostatic coupling can occur between top- and bottom-tier transistors when the ILD is very thin. In the transistor-level design partitioning, the front gate of the transistor in the bottom tier can act as the bottom gate of the transistor in the top tier [10]. Therefore, the top-tier transistors behave as double-gate SOI transistors with asymmetric front and back gates (see Fig. 3(a)). Similarly, in gate-level partitioning, the uppermost metal line from the BEOL of the bottom layer acts as the bottom gate for the top-layer transistor and the ILD acts as the back-gate dielectric [10]. The voltage on the metal line impacts the electrical state of the channel in a top-layer transistor, thereby coupling the top and bottom device layers (see Fig. 3(b)).

B. Impact of Process Variations and Manufacturing Defects on SDD Testing

Coupling and void defects are typically manifested as SDDs at nano-scale technology nodes. Commercial SDD ATPG tools use variation- and defect-unaware static timing analysis (STA) to generate test patterns to sensitize long paths. However, the slack on a path can vary in the presence of fabrication imperfections; therefore, it is necessary to consider these effects while generating test patterns. This is highlighted using simulation results in [10]. The effectiveness of the commercial SDD test patterns is calculated using the SDQL metric [16] for two cases: (i) fault-free circuit: the nominal slack data and the test pattern set used as inputs during fault simulation, and (ii) faulty circuit: the slack data for a defective M3D instance and the test pattern set used as inputs during fault simulation. Over multiple defective instances, the SDQL values obtained in the faulty cases are greater than those obtained using the nominal (fault-free) timing library. Representative results for

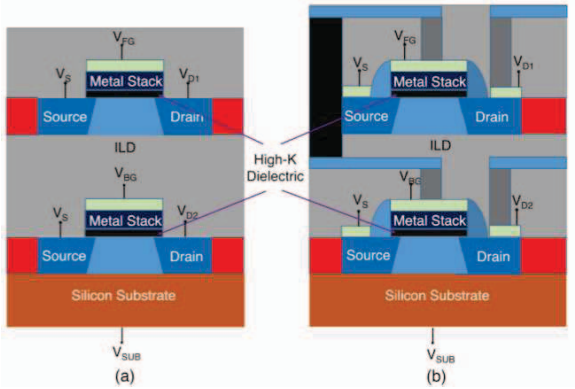

Fig. 3: Electrostatic coupling between top- and bottom-tier transistors in (a) transistor-level partitioned and (b) gate-level partitioned M3D ICs [10].

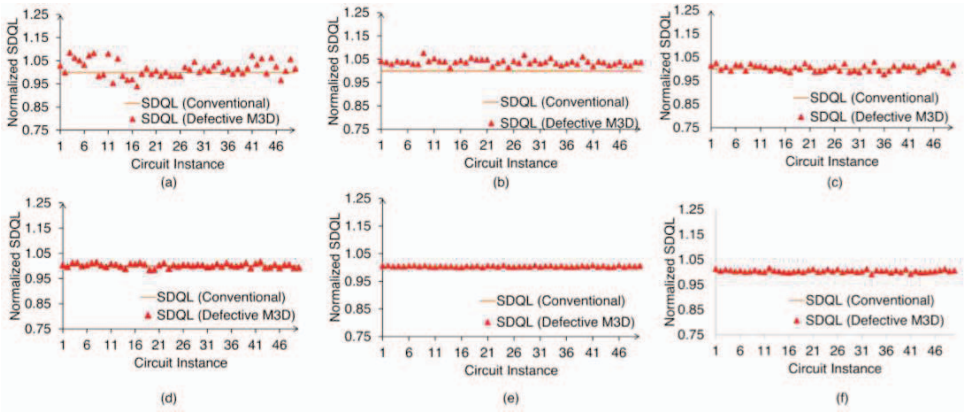

Fig. 4: Normalized SDQL values for IWLS 2005 benchmarks: (a) vga_lcd (23 nm thick ILD), (b) leon3 (23 nm thick ILD), (c) ethernet (23 nm thick ILD), (d) vga_lcd (100 nm thick ILD), (e) leon3 (100 nm thick ILD), (f) ethernet (100 nm thick ILD) [10].

three IWLS benchmarks are shown in Fig. 4.

Clearly, SDD ATPG tools need to consider the timing slack of critical paths under random process uncertainties to select appropriate critical paths for sensitization. Statistical STA-based ATPG methods take variability-aware delay data into consideration to generate efficient test patterns [17]. However, the extremely high run time associated with the multiple dynamic timing analysis runs render such methods inefficient for modern designs. Variation-aware delay fault ATPG methods such as [18], [19] have limited applicability as they only consider variations that have a linear impact on the propagation delay.

VI. Power Supply Noise-Aware Delay Testing

Power supply noise (PSN) in a power delivery network (PDN) is the difference in voltage between power supplies and local receivers. PSN-induced voltage droop can be categorized into two components: IR-drop and Ldi/dt drop. IR-drop results from instantaneous current flowing through the equivalent resistance along conduction paths when switching activities occur; while rapid changes in current with the parasitic inductance causes Ldi/dt drop. In M3D ICs, irregular placement of power MIVs, long resistive paths to bottom-tier receivers, and reduction in the number of C4 bumps lead to higher voltage droop compared with their 2D counterparts [20]. The testing mode suffers more from PSN than functional-mode operations due to high switching activity during scan shift and capture. Excessive PSN-induced voltage droop may slow down signal propagation through sensitized paths, resulting in the failure of good chips, i.e., yield loss. However, most previous work has optimized M3D PDN designs only for the functional mode [20]–[22]. A detailed analysis of PSN during scan testing and its impact on yield loss needs to be carried out.

To obtain the voltage droop value of each test pattern, a new analysis framework specific for M3D ICs has been developed in [23], as shown in Figure 5. Transition delay fault patterns are generated after place and route. For each pattern, a post-routed gate-level simulation is conducted to record the switching activities in a value change dump (VCD) file. Such a VCD file is imported into a commercial tool for vector-based dynamic power and rail analysis. Because commercial tools do not consider M3D designs, two tiers in an M3D IC have been analyzed separately with the 2D analysis flow. One major difference between the top and the bottom tier is that supply current for bottom-tier transistors flows through the top-tier PDN. Therefore, the top-tier PDN suffers from additional power consumption and current demand, while the bottom tier has a lower supply voltage than the nominal value due to voltage droop in the top tier. To simulate this scenario, top-tier power and rail analysis is first performed with the scaled current. A new voltage value is calculated by subtracting the maximum voltage droop in the top tier from the nominal supply voltage to analyze the worst-case scenario. This value is utilized as the power source of the bottom tier to complete the bottom-tier power and rail analysis.

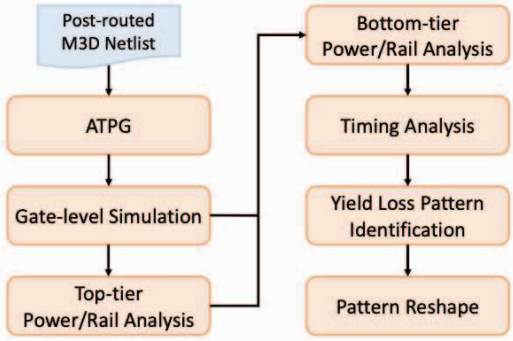

Fig. 5: Dynamic analysis flow for M3D ICs.

Due to tool limitations, the simulation window could not be extended to the whole testing process. The weighted switching activity (WSA) metric [24] is used to estimate power consumption during scan capture. Experimental results demonstrate that there is a high correlation between the WSA of the top tier only and the voltage droop in test mode. Therefore,

patterns with large WSA values are extracted to conduct power and rail analysis, in which the worst-case voltage droop during test application can be obtained. Next, PSN-aware timing analysis for each pattern is carried out by scaling delay with a factor based on the obtained voltage droop value. Increased delay of sensitized paths for a pattern leads to the reduction of slacks. Once the slack becomes negative, the corresponding pattern is identified to be susceptible to yield loss and is extracted to be reshaped.

In the pattern reshaping process, an algorithm based on integer linear programming (ILP) is developed. First, patterns that lead to yield loss are removed from the original set and the fault list is updated. Next, the ATPG process is performed to generate new patterns for undetected faults with don't-care bits unfilled. The ILP-based algorithm is to fill don't-care bits for each pattern such that the WSA of the top tier is minimized. In ILP modeling, the functionality of every Boolean logic gate is realized by a set of linear constraints. Note that there are two vectors V_1 and V_2 in a delay-fault pattern to represent the initial state and the launch state, respectively. Therefore, the forward implication is executed twice by applying such vectors continuously to formulate the ILP model constraints for the complete scan capture procedure. The objective function for the ILP model is shown as below:

$$\min \sum_i (i^1 \oplus i^2)(1 + N_{fo,i}), \text{ for all top-tier nets } i,$$

where i^1 and i^2 is the signal of net i with vector V_1 and V_2, respectively, and $N_{fo,i}$ is the fanout of net i. The solution to this ILP problem produces a fully specified pattern with a minimum WSA value of the top tier. This algorithm is applied to each pattern that requires to be reshaped to get the complete pattern set. Experimental results show that the pattern sets after reshaping eliminate the yield loss issue for all benchmark designs. Furthermore, compared to a 2D baseline don't-care bits filling algorithm in [25], the proposed method can achieve a lower number of paths with a marginal slack. This improvement helps prevent good chips from failing due to small process variations.

VII. Conclusion

While we have witnessed significant developments in monolithic integration in recent years, the exceptionally high 3D-interconnect density and the unique low-temperature fabrication process propose several challenges for testing and DfT. In this paper, we described a low-cost BIST architecture that uses only two test patterns to detect ILV faults with minimal probability of fault masking. To detect inter-tier variation and improve yield learning, we described a method for tier-level fault localization using observation points. The M3D fabrication flow is immature and prone to process variations and manufacturing defects – we explored the impact of these uncertainties on the delay fault testing. In addition, we have presented a framework to identify test patterns that are likely to cause yield loss due to droop-induced added delay on sensitized paths. Pattern reshaping using an ILP-based X-filling algorithm eliminates this yield loss during scan capture.

References

[1] J. H. Lau, "Evolution, Challenge, and Outlook of TSV, 3D IC Integration and 3D Silicon Integration," in *Int. Symp. on Advanced Packaging Materials*. IEEE, 2011, pp. 462–488.

[2] Z. Li et al., "Design and Package Technology Development of Face-to-Face Die Stacking as a Low Cost Alternative for 3D IC Integration," in *IEEE Electronic Components and Technology Conf.* IEEE, 2014, pp. 338–341.

[3] P. Batude et al., "Advances in 3D CMOS Sequential Integration," in *Proc. IEEE Int. Electron Devices Meeting*. IEEE, 2009, pp. 1–4.

[4] M. Vinet et al., "Monolithic 3D Integration: A Powerful Alternative to Classical 2D Scaling," in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conf.* IEEE, 2014.

[5] T. Yonehara, "Epitaxial layer transfer technology and application," in *SOI-3D*, 2015.

[6] J.-H. Park et al., "Low temperature ($\lesssim$380 °C) and high performance ge CMOS technology with novel source/drain by metal-induced dopants activation and high-k/metal gate stack for monolithic 3D integration," in *IEEE IEDM*, 2008.

[7] K. Chang et al., "Design automation and testing of monolithic 3D ICs: Opportunities, challenges, and solutions," in *2017 IEEE/ACM ICCAD*. IEEE, 2017.

[8] K. Chang et al., "System-level power delivery network analysis and optimization for monolithic 3D ICs," *IEEE Trans. on Very Large Scale Integr. (VLSI) Syst.*, vol. 27, pp. 888–898, 2019.

[9] J. Saxena et al., "A case study of IR-drop in structured at-speed testing," in *IEEE ITC*, 2003.

[10] A. Koneru et al., "Impact of electrostatic coupling and wafer-bonding defects on delay testing of monolithic 3D integrated circuits," in *JETC*, 2017.

[11] R. Pendurkar et al., "Switching activity generation with automated BIST synthesis for performance testing of interconnects," in *TCAD*, 2001.

[12] A. Jutman, "Shift register based TPG for at-speed interconnect BIST," in *24th Int. Conf. on Microelectronics*, 2004.

[13] D. Erb et al., "Multi-cycle circuit parameter independent ATPG for interconnect open defects," in *VTS*, 2015.

[14] A. Chaudhuri et al., "Built-in self-test for inter-layer vias in monolithic 3D ICs," in *IEEE European Test Symposium*, 2019.

[15] A. Chaudhuri et al., "NodeRank: observation-point insertion for fault localization in monolithic 3D ICs," in *IEEE Asian Test Symposium*, 2020 (to appear).

[16] Y. Sato et al., "Invisible delay quality-SDQM model lights up what could not be seen," in *IEEE Int. Test Conf.*, 2005.

[17] C.-T. Chao et al., "Pattern selection for testing of deep sub-micron timing defects," in *Proc. Design, Automation and Test in Europe*, vol. 2. IEEE, 2004, pp. 1060–1065.

[18] X. Lu et al., "Longest-path selection for delay test under process variation," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, pp. 1924–1929, 2005.

[19] R. Brawhear et al., "Predicting circuit performance using circuit-level statistical timing analysis," in *Proc. European Design and Test Conf.*, 1994.

[20] K. Chang et al., "System-level power delivery network analysis and optimization for monolithic 3D ICs," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 4, pp. 888–898, April 2019.

[21] S. Samal et al., "Full chip impact study of power delivery network designs in monolithic 3D ICs," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2014, pp. 565–572.

[22] A. Koneru et al., "Reliable power delivery and analysis of power-supply noise during testing in monolithic 3D ICs," in *IEEE 37th VLSI Test Symposium (VTS)*, April 2019.

[23] S.-C. Hung et al., "Power supply noise-aware scan test pattern reshaping for at-speed delay fault testing of monolithic 3D ICs," in *IEEE Asian Test Symposium*, 2020 (to appear).

[24] P. Girard, "Survey of low-power testing of VLSI circuits," *IEEE Design Test of Computers*, vol. 19, no. 3, pp. 82–92, 2002.

[25] S. Remersaro et al., "Preferred fill: A scalable method to reduce capture power for scan based designs," in *IEEE International Test Conference*, 2006.