Built-in Self-Test and Fault Localization for Inter-Layer Vias in Monolithic 3D ICs

ARJUN CHAUDHURI and SANMITRA BANERJEE, Duke University
JINWOO KIM, HEECHUN PARK, and BON WOONG KU, Georgia Institute of Technology
SUKESHWAR KANNAN, Broadcom Inc.
KRISHNENDU CHAKRABARTY, Duke University
SUNG KYU LIM, Georgia Institute of Technology

Monolithic 3D (M3D) integration provides massive vertical integration through the use of nanoscale inter-layer vias (ILVs). However, high integration density and aggressive scaling of the inter-layer dielectric make ILVs especially prone to defects. We present a low-cost built-in self-test (BIST) method that requires only two test patterns to detect opens, stuck-at faults, and bridging faults (shorts) in ILVs. We also propose an extended BIST architecture for fault detection, called Dual-BIST, to guarantee zero ILV fault masking due to single BIST faults and negligible ILV fault masking due to multiple BIST faults. We analyze the impact of coupling between adjacent ILVs arranged in a 1D array in block-level partitioned designs. Based on this analysis, we present a novel test architecture called Shared-BIST with the added functionality of localizing single and multiple faults, including coupling-induced faults. We introduce a systematic clustering-based method for designing and integrating a delay bank with the Shared-BIST architecture for testing small-delay defects in ILVs with minimal yield loss. Simulation results for four two-tier M3D benchmark designs highlight the effectiveness of the proposed BIST framework.

CCS Concepts: • **Hardware → Hardware test**; **Design for testability**; **Built-in self-test;**

Additional Key Words and Phrases: Monolithic 3D IC, design-for-test

ACM Reference format:
Arjun Chaudhuri, Sanmitra Banerjee, Jinwoo Kim, Heechun Park, Bon Woong Ku, Sukeshwar Kannan, Krishnendu Chakrabarty, and Sung Kyu Lim. 2021. Built-in Self-Test and Fault Localization for Inter-Layer Vias in Monolithic 3D ICs. *J. Emerg. Technol. Comput. Syst.* 18, 1, Article 22 (November 2021), 37 pages.
https://doi.org/10.1145/3464430

This research was supported in part by the DARPA ERI 3DSOC Program under Award HR001118C0096. The work of A. Chaudhuri, S. Banerjee, and K. Chakrabarty was also supported in part by the National Science Foundation under grant CCF-1908045. A preliminary version of this paper was presented at the IEEE European Test Symposium, 2019 [8].
Authors' addresses: A. Chaudhuri, Apt 15B, 2748 Campus Walk Avenue, Durham, NC 27705; email: arjun.chaudhuri@duke.edu; S. Banerjee, Apt 19E, 2748 Campus Walk Avenue, Durham, NC 27705; email: sanmitra.banerjee@duke.edu; J. Kim, KACB 2360, GTCAD Laboratory, Klaus Advanced Computing Building, 266 Ferst Drive, Atlanta, GA 30332; email: jinwookim@gatech.edu; H. Park, KACB 2360, GTCAD Laboratory, Klaus Advanced Computing Building, 266 Ferst Drive, Atlanta, GA 30332; email: phcssl0112@gmail.com; B. W. Ku, KACB 2360, GTCAD Laboratory, Klaus Advanced Computing Building, 266 Ferst Drive, Atlanta, GA 30332; email: bwku@gatech.edu; S. Kannan, 1320 Ridder Park Dr, San Jose, CA 95131; email: sukeshwar.kannan@broadcom.com; K. Chakrabarty, Box 90291, 130 Hudson Hall, Science Drive, Durham, NC 27710; email: krish@duke.edu; S. K. Lim, KACB 2360, GTCAD Laboratory, Klaus Advanced Computing Building, 266 Ferst Drive, Atlanta, GA 30332; email: limsk@ece.gatech.edu.

1550-4832/2021/11-ART22 $15.00
https://doi.org/10.1145/3464430

1 INTRODUCTION

The advent of 3D integration has extended Moore's law by enabling novel architectures through high-density vertical interconnects [48]. 3D **integrated circuits** (**ICs**) are typically realized by one of three integration approaches, namely, die stacking, wafer stacking, and monolithic integration [33]. The diameter and pitch of the high-density vertical interconnects differ for these approaches, typically ranging from a few microns for die/wafer stacking to only a few nanometers in the case of monolithic integration [49].

The nanoscale **inter-layer vias** (**ILVs**) used in **monolithic 3D** (**M3D**) designs enable massive vertical integration, resulting in far denser vertical connections compared to conventional TSV-based 3D ICs [43]. However, high integration density and aggressive scaling of the inter-layer dielectric make ILVs especially prone to defects [26, 27]. ILV testing is therefore needed to ensure effective defect screening and quality assurance. While ILVs can conceivably be tested together with the M3D logic/memory tiers, defect isolation and yield learning require a test solution that can exclusively target the ILVs in an M3D IC. **Design-for-testability** (**DfT**) for ILVs is therefore a promising step toward this direction.

There are three different types of M3D–IC designs, namely, transistor-level, gate-level, and block-level [29, 36, 37]. In transistor-level design, all p-type and n-type MOSFETs are partitioned into different tiers using ILVs, whereas cells of both p-type and n-type MOSFETs are tier-partitioned in gate-level design. In block-level design, functional blocks are floorplanned into different tiers. We focus on the block-level design methodology for our experimental designs because it requires the smallest number of ILVs and promotes the reuse of pre-designed macros and IPs.

In this article, we present a low-cost **built-in self-test** (**BIST**) framework to detect opens, **stuck-at faults** (**SAFs**), bridging faults (shorts), and crosstalk-induced faults in the ILVs of block-level partitioned M3D–ICs. The key contributions of this article are as follows:

- We present a low-cost BIST architecture called *XOR-BIST* that requires only two test patterns for detecting SAFs, hard shorts, and hard opens in ILVs.
- We investigate the probability of ILV fault masking due to faults in the BIST hardware. We propose an extended BIST architecture called *Dual-BIST* that guarantees zero ILV fault masking due to single BIST faults and negligible ILV fault masking probabilities for multiple BIST faults.
- We explore the detectability of resistive faults and present an enhanced Dual-BIST design that can detect a wide range of resistive shorts and opens.
- We analyze the impact of crosstalk between adjacent ILVs arranged in a 1D array in block-level partitioned designs. We present a first-order approximation of the ILV-to-ILV coupling model, which is validated by S-parameter analysis. Based on the analysis, we propose a minimal test sequence to cover all possible crosstalk-induced faults in ILVs.
- We present a new test architecture called *Shared-BIST* that is capable of detecting and localizing faults (single and multiple), including crosstalk-induced faults.
- We present the **design-space exploration** (**DSE**) of the Shared-BIST architecture for selecting an optimal design configuration with the area and test-time overheads within acceptable limits.
- We describe a clustering-based methodology for designing a delay bank—as a subcircuit of the Shared-BIST architecture—to enable the testing of **small-delay defects** (**SDDs**) in ILVs.
- We evaluate the Dual-BIST and Shared-BIST overheads for M3D benchmarks and compare them to a baseline DfT method that uses flip-flops at the two ends of an ILV for controllability and observability.

The remainder of the article is organized as follows. Section 2 presents an overview of M3D technology and related prior work on interconnect BIST and ILV testing. Section 3 describes the XOR-BIST solution. Section 4 describes the Dual-BIST architecture that minimizes the probability of ILV fault masking. The enhanced BIST architecture for resistive faults is described in Section 5. Section 6 presents the analysis and test of crosstalk-induced faults in ILVs and evaluates their impact on the quality of SDD testing. Section 7 describes the Shared-BIST architecture and the fault localization solution, and presents the design methodology for the delay bank. Section 8 presents evaluation results and Section 9 concludes the article.

2 BACKGROUND

2.1 M3D Fabrication Process

The first step in M3D fabrication involves a standard high-temperature process to integrate the transistors and interconnects in the bottom layer. A thin inter-layer dielectric is then created over the bottom layer and low temperature molecular bonding of the **silicon-on-insulator** (**SOI**) substrate is used to obtain the top layer [4, 54]. The ILVs are finally fabricated to connect the top and bottom layers. The above steps are repeated for the fabrication of additional layers.

2.2 M3D Technology Qualification with Block-level Partitioned Designs

Block-level partitioned designs are preferred to gate-level partitioned designs when logic and memory modules cannot be partitioned across multiple tiers [24]. The unification of logical and physical hierarchies and hierarchical uniformity are achieved using block-level partitioning [34]. In [3], the 3D–IC technology roadmap envisioned by CEA-Leti shows that block-level partitioning appears prior to gate-level designs in the roadmap of sequential or M3D integration.

While conventional 3D-stacked integration involves only core-level and block-level partitioned designs, sequential integration will be applied to block-level designs (intermediate partitioning) for the qualification of the sequential fabrication process before transferring to more fine-grained gate-level designs. Hence, ILV–BIST for block-level designs is essential for qualifying the ILV manufacturing process during this transition from 3D–SiC (W2W bonding) to gate-level 3D–SoC (active layer bonding). Specifically, because block-level designs have fewer ILVs, ILV–BIST can be deployed with low overhead to enable M3D–PDK qualification for gate-level designs that are next in line. Therefore, in addition to facilitating trustworthy IP reuse and heterogeneous integration [24], block-level partitioned designs can be leveraged by design-houses as "test vehicles" for overall M3D process development.

Transistor-level partitioned designs have dense ILVs crossing tiers, causing global P-to-N skew and routing congestion. This results in variations in the threshold voltages as well as coupling-induced delay degradation in a large majority of the nets [32]. As a result, block-level integration, followed by gate-level integration, is likely to be the preferred choice, especially for qualifying the immature M3D technology.

2.3 ILV Fault Models

The target fault models for ILVs are the same as those for interconnects in an IC with one active (device) layer because the dimensions of an ILV are comparable to that of vias in today's ICs [32]. During fabrication, the ILVs are treated as back-end-of-line vias that are susceptible to open and short defects [18]. Therefore, typical fault models for an ILV are shorts, opens, and SAFs [25, 26]. ILV faults can be further classified into hard and resistive categories based on the type and size of the underlying defects. Hard shorts can occur due to imperfect design and circuit synthesis, or particle contamination during fabrication [6]. A resistive short has a resistance in the intermediate

range, typically ranging from a few KΩs to several hundred KΩs [39]. Resistive shorts can occur when the ILV metal diffuses through the ILD to make a partial contact with another nearby ILV [16], or due to defects at the interface between two tiers after vertical integration of the top tier's device layer [26].

A hard open occurs when an ILV fails to land on a contact pad. This gap in connection leads to a very high open resistance, typically in the order of MΩs [26, 39]. A resistive open is of relatively smaller size and has a resistance in the intermediate range, typically ranging from a few KΩs to several hundred KΩs [39]. Resistive opens can occur due to bonding defects [26], mechanical stress-induced striations on the underside of an ILV [30], air-voids inside an ILV due to imperfect electro-chemical deposition of metal [15], hairline cracks, and pinhole defects [46].

2.4 Related Prior Work

The testing of M3D–ICs in general, and ILVs, in particular, has remained largely unexplored thus far. Due to the high ILV integration density (30 million per mm^2 [32]), retrofitting of conventional interconnect BIST approaches can introduce significant overhead. Methods such as [40, 41] use dedicated scan elements (test points) for test access. However, these solutions require large test application time since the number of test patterns required for high fault coverage can become prohibitively large for high ILV density [21]. Moreover, the number of required test points is directly proportional to the ILV count. ATPG-based interconnect test methods, such as [13], are likely be less effective for ILV testing because I/O pins are available only on one layer in an M3D IC; either test data or test responses—or both in the case of ILVs that do not land on the bottom tier—must be propagated through multiple tiers and the associated ILVs. This requirement adds significantly to the propagation constraints for ATPG. Even if tests can be found by an ATPG tool, additional ILV faults on test paths, which is a likely scenario due to high ILV density, will impede testability. Commercial ATPG tools tend to target single faults for test-pattern generation. However, multiple faults are likely for dense ILV layouts; hence, test escapes might occur if tests are generated under the single-fault assumption. The proposed BIST alleviates these problems by using a compact set of test patterns that exhaustively test for single or multiple ILV fault scenarios with test-output compaction and negligible fault-masking probability.

The concept of **Known Good Die** (**KGD**) is applicable to TSV-based 3D ICs where the cost of die-to-wafer or die-to-die stacking is reduced by bonding dies (i.e., 3D tiers) to a (separate) KGD that has already passed all necessary pre-bond tests [14]. In contrast, an M3D IC employs *in-situ* sequential fabrication of tiers in an integrated process, instead of stacking independently processed tiers. Therefore, the notions of KGD and pre-bond testing do not exist in the context of M3D–ICs.

While post-bond TSV testing techniques can be extended to post-assembly M3D testing, recently proposed methods such as [35] need a die-wrapper register cell on both ends of the ILV for controllability and observability. The drawbacks of applying the IEEE 1838 3D test standard to M3D–ICs, which contain many more vertical connections than in 3D-stacked ICs, include: (i) the current test standard does not provide on-chip ILV-fault localization capabilities; (ii) dedicated die-wrapper registers on every ILV will have high area overhead; (iii) inland wrapping (to avoid adding dedicated wrapper cells) will test both ILV and tier logic as part of "EXTEST" leading to potential test escapes of ILV faults, which are more likely to occur than logic faults. In [45], a design flow is proposed to test the full M3D stack, including ILVs, but without on-chip fault localization. In [22, 27], an inter-layer ILV–BIST solution is presented using interface scan cells and a twisted ring counter. However, it mandates a dedicated test layer, which can have a significant impact in terms of the number of fabrication steps and area overhead. This technique also assumes that the number of upward-facing ("up") ILVs is equal to the number of downward-facing ("down") ILVs

between the two tiers. However, in real designs, this assumption is unlikely to hold and dummy ILVs must be added to equalize the ILV counts.

A novelty of the proposed ILV–BIST methods is that faults can be detected without an additional test layer and without an equal number of up and down ILVs between two tiers. During post-bond or pre-assembly testing of block-level partitioned ICs, BIST can screen faulty M3D dies with low test escape. Consequently, resources will not be spent on packaging a known bad die. This will reduce the **unit/per hour** (**UPH**) cost, assembly cycle time, as well as the assembly material cost. Targeted ILV–BIST will reduce test escapes and accelerate failure analysis once faults are localized during production testing (or post-assembly testing). The proposed BIST methods can be applied for in-field testing to enable early detection of aging-induced reliability failures and possible repair or recovery in the field.

3 PROPOSED BIST APPROACH AND ANALYSIS OF FAULT DETECTABILITY

3.1 BIST Architecture

The BIST architecture for testing shorts, opens, and SAFs in ILVs is shown in Figure 1. The BIST design is partitioned into two segments. The first segment corresponds to the tier (Tier 1) that includes the driving side of the ILVs. The second segment corresponds to the tier (Tier 2) that is driven by the ILVs under test. This segmentation enables testing of both bidirectional and unidirectional ILVs. On the output side of the ILVs, we insert 2-input XOR gates between adjacent ILVs. During ILV planning and placement in the design phase of a block-level partitioned M3D–IC, ILVs of the same bus tend to be placed closer to each other; every ILV can therefore be shorted to at most two adjacent ILVs. This is accounted for by our XOR placement to reduce gate count. There are $(N-1)$ XOR gates in Tier 2 for a set of N ILVs. The XOR outputs feed a space compactor. This compactor is an optimally balanced AND tree with $(N-1)$ inputs and a 1-bit output signature Y_1. We can determine if there is a fault in the given set of ILVs by observing Y_1.

In Tier 1, test data are fed to the ILVs from an input source V_{in}, which provides complementary signals to adjacent ILVs in the test mode via an inverter chain. Therefore, for N ILVs, there are $(N-1)$ inverters in Tier 1. A 2:1 **multiplexer** (**MUX**) is present at the input of every ILV to switch between functional and test modes based on the *Launch* signal. Note that the inverter chain is not on the functional path; therefore, it does not impact timing closure in functional mode. Nevertheless, a long inverter chain can impact the frequency at which V_{in} can be switched. In such a scenario, the inverter chain can be broken into sub-chains and the test inputs to the ILVs adjusted accordingly to ensure that adjacent ILVs receive complementary values.

3.2 Detection of Hard Faults

The ILVs are tested in two clocks cycles and V_{in} is switched between these cycles. In the first (second) clock cycle, V_{in} is set to 1 (0). The resulting test patterns to the ILVs are therefore "101..." in the first cycle and "010..." in the second cycle. It is shown in the appendix that these test patterns and the space compactor can together detect all single and multiple hard faults in the ILVs.

The manner in which the ILVs are driven in the test mode leads to a deterministic hard-short behavior; this is illustrated in the inset of Figure 1. If two ILVs are shorted, the ILV that appears first (pre-ILV) in the path of the incoming test signal from V_{in} will drive the other ILV (post-ILV). This is because the post-ILV will experience opposite pulls via two paths: one through the highly conductive short to the pre-ILV (Pull 1) and the other through the MUX and inverter to the test signal (Pull 2). HSPICE simulations show that the latter is of higher resistance and hence offers weaker pull.

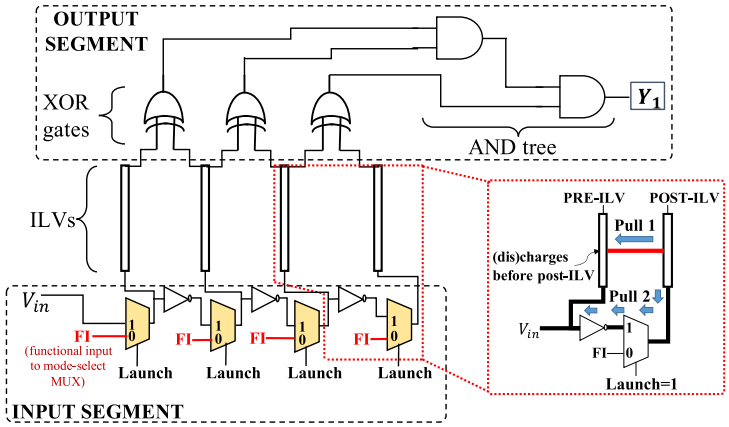


Fig. 1. Simplest form of the BIST architecture and illustration of the hard-short behavior.

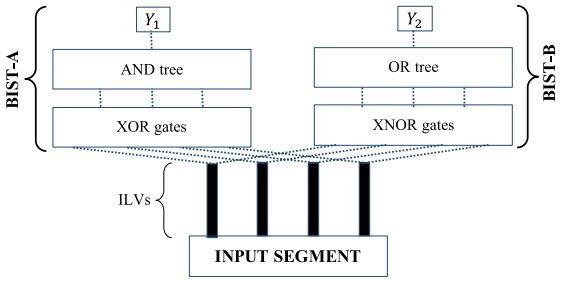


Fig. 2. Illustration of the "Dual" BIST architecture.

4 AVOIDANCE OF ILV FAULT MASKING

4.1 Proposed Architecture

The BIST design of Section 3.1 is also susceptible to SAFs. Some of these faults may mask fault(s) in the ILVs under test. If this happens, Y_1 will be 1 in both clock cycles even if the ILVs contain fault(s). To minimize the probability of ILV fault masking due to faults in the BIST design (referred to as BIST-A), we add a parallel propagation path from the ILV outputs to a second 1-bit signature Y_2 (we refer to this part of the BIST logic as BIST-B). The physical structure of this parallel path to Y_2 (PY_2) is identical to that of the path from the XOR inputs to Y_1 (PY_1). The XOR and AND gates in PY_1 are replaced with their Dual counterparts (XNOR and OR, respectively) in PY_2. This "Dual" BIST architecture is shown in Figure 2.

4.2 Masking of ILV Faults due to Single SAF in BIST

In the Dual-BIST architecture, let the 2-bit signature for the ILVs be $\{Y_1, Y_2\}$, where Y_1 and Y_2 represent the outputs from the XOR-AND and XNOR-OR compactors, respectively. The XOR-AND and XNOR-OR compactors are denoted as BIST-A and BIST-B, respectively. The proposed BIST method cannot detect the fault scenarios in which all the ILVs are stuck at 0's and 1's alternately (010101... or 101010...). This is because the ILVs are stuck at values identical to the test patterns and $Y_1 = 1$ in both the test clock cycles, resulting in fault masking. We show in the appendix that if this low-probability scenario does not occur, ILV faults cannot be masked by a single BIST fault.

4.3 Masking of ILV Faults due to Multiple SAFs in BIST

For analyzing ILV fault masking due to multiple BIST faults, we only need to address the cases when we observe $Y_1 = 1$ and $Y_2 = 0$ in both cycles. We first define a faulty ILV location as a pair of adjacent ILVs where either there is a short between them, or both of them are stuck-at or open, or a combination of these faults occurs. If the fault is activated, the corresponding error will propagate through the XOR (X_1) and XNOR (XN_1) gates that are driven by the two ILVs. Note that every adjacent ILV pair drives a unique XOR/XNOR pair, which propagates the error to Y_1 and Y_2.

Let V_1 and V_2 be a pair of faulty ILVs that are XOR-ed in the compactor. A fault, involving an ILV or a pair of ILVs, will be masked by the BIST logic only if any of the following conditions is true:

(1) Condition (1): Let the path from X_1's output to Y_1 be called P1. Let the path from XN_1's output to Y_2 be P2. Here, X_1 and XN_1 denote the XOR and XNOR gates driven by the ILVs V_1 and V_2, respectively. If there exists at least one s/1 fault in P1 and at least one s/0 fault in P2, any ILV fault, if present, will be masked as we will obtain $Y_1 = 1$, $Y_2 = 0$ in both cycles.
(2) Condition (2): The inputs of all XOR and XNOR gates are stuck at complementary values.

Let F be the event that a pair of ILVs contains fault(s) and M be the event that the fault(s) is (are) masked by the BIST logic. We next calculate the probability $P[F \cap M]$. For a set of $c + 1$ ILVs and c XOR/XNOR outputs, if the AND and OR trees are well-balanced, the total number of nodes N_e from any given XOR (XNOR) output to Y_1 (Y_2) will be either $\lceil \log_2 c \rceil$ or $\lceil \log_2 c \rceil + 1$. The paths from X_1 to Y_1 and XN_1 to Y_2 will have the same number of nodes because of the identical structures of the AND and OR trees.

Let p_b be the probability that a node in the BIST logic is faulty, with the s/0 and s/1 faults being equiprobable. Let p_s be the probability of a short between adjacent ILVs. Let p_o be the probability that an ILV is s/0 (open) or s/1. Let p_m be the probability that any given ILV pair is faulty. Therefore, p_m is defined as: $p_m = 1 - (1 - p_s)(1 - p_o)^2$. This expression can be interpreted as follows—an ILV location is not faulty if there is no short and none of the two adjacent ILVs are open or stuck-at. We know from basic probability theory that $P[F \cap M] = P[F].P[M|F]$; here, $P[F]$ is the probability that a given ILV pair is faulty and $P[M|F]$ is the probability that, given the ILV pair contains fault(s), the fault(s) is (are) masked by faults in the BIST. With this understanding, $P[F] = p_m$.

To derive $P[M|F]$ and consequently $P[F \cap M]$, we add the probabilities for the two mutually exclusive cases of multiple fault masking described earlier. If Condition (1) is true, there are N_e nodes in the paths P1 and P2, and a fault is masked if s/1 and s/0 faults are present in at least one of the N_e nodes in P1 and P2, respectively. The probability for (1) is therefore $p_m(1 - (1 - \frac{p_b}{2})^{N_e})^2$. As P1 and P2 enforce different masking probabilities for different ILVs due to different path lengths, we conservatively consider the worst-case masking scenario by associating the highest possible fault-masking probability with all the paths from ILVs to Y_1 and Y_2. A larger number of edges in a path implie a higher chance of masking. Therefore, we consider $N_e = \lceil \log_2 c \rceil + 1$. If Condition (1) does not hold, the associated probability will be $(1 - \frac{p_b}{2})^{2N_e}$; we multiply this probability with the probability for the case when Condition (2) is true. If (2) holds, the probability is $2^{2c} p_m \cdot (\frac{p_b}{2})^{4(c+1)}$. Therefore, if we let $P_{FM} = P[F \cap M]$, we have

$$P_{FM} = p_m \left(1 - \left(1 - \frac{p_b}{2}\right)^{N_e}\right)^2 + \left(1 - \frac{p_b}{2}\right)^{2N_e} .2^{2c} p_m \cdot \left(\frac{p_b}{2}\right)^{4(c+1)}.$$

Here, $P[M|F] = (1 - (1 - \frac{p_b}{2})^{N_e})^2 + (1 - \frac{p_b}{2})^{2N_e} .2^{2c} \cdot (\frac{p_b}{2})^{4(c+1)}$. Since $p_b < 1$ and $p_b N_e << 1$, we use the approximation $(1 - x)^\alpha \approx (1 - \alpha x)$ for $|x| < 1$ and $|\alpha x| << 1$ to simplify P_{FM} as follows:

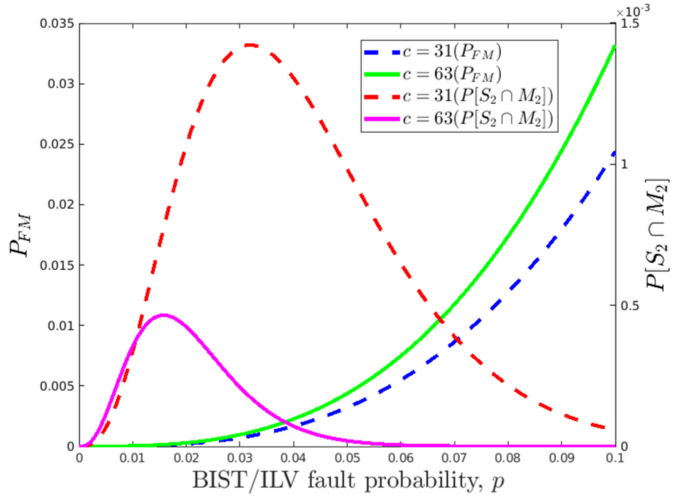


Fig. 3. Multiple-fault masking probability.

$$P_{FM} \approx p_m \left(1 - \left(1 - \frac{p_b N_e}{2}\right)\right)^2 + (1 - p_b N_e).2^{2c} p_m . \left(\frac{p_b}{2}\right)^{4(c+1)}$$
$$= p_m \left(\frac{p_b^2 N_e^2}{4} + \frac{(1 - p_b N_e) p_b^{4c+4}}{2^{2c+4}}\right) \approx \frac{p_m p_b^2 N_e^2}{4},$$

since $2^{2c+4} > 4$, $p_b^{4c+4} < p_b^2$, and $(1 - p_b N_e) < N_e^2$. We next derive the probabilities that specific ILV fault scenarios are masked by multiple BIST faults (assuming independence of events):

Case-I: If S_1 is the event that k known faulty ILV pairs are present in the set of $c + 1$ ILVs and M_1 is the event that this fault scenario is masked by the BIST logic, then the masking probability can be estimated as: $P[S_1 \cap M_1] = (P_{FM})^k$.

Case-II: Let us consider the case where any k out of the c ILV pairs contain fault(s) (event S_2) and the fault(s) is (are) masked by faults present in the BIST engine (event M_2). This is different from *Case-I* as *Case-I* considered the identities of the k faulty ILV pairs to be known *a priori*. The number of ways of choosing k pairs (to contain fault(s)) from c pairs is $\binom{c}{k}$. For each of these mutually exclusive ways, the probability that the k-fault scenario is masked is $P[S_1 \cap M_1]$. Furthermore, for every choice of k pairs that contain fault(s), we are considering the remaining $c - k$ pairs to be fault-free. Therefore, the probability $P[S_2 \cap M_2]$ can be expressed as:

$$P[S_2 \cap M_2] = \binom{c}{k} P[S_1 \cap M_1](1 - p_m)^{c-k} = \binom{c}{k} (P_{FM})^k (1 - p_m)^{c-k}.$$

Figure 3 shows the variation of P_{FM} and $P[S_2 \cap M_2]$ with p, where $p = p_s = p_o = p_b$, $k = 1$, and $c \in \{31, 63\}$. The *width* of a BIST engine is defined as the number of ILVs $(c+1)$ tested concurrently by the engine. For a wider BIST engine, P_{FM} is higher because more logic gates are present in the path between ILV and BIST output. In contrast, $P[S_2 \cap M_2]$ is generally lower for higher engine width due to the presence of a higher number of ILVs. The different approaches undertaken to calculate the two probabilities lead to opposite trends in their values as the engine width is varied. Table 1 gives the values of $P[S_2 \cap M_2]$ for $k > 1$, $c = 31$, and $p = 0.05$. We observe that the probability of ILV fault(s) being masked by multiple BIST faults is very low.

Table 1. ILV Fault-masking Probabilities Due to Multiple BIST Faults

k	$P[S_2 \cap M_2]$
2	5.5228×10^{-5}
3	1.9982×10^{-6}
4	5.2354×10^{-8}
5	1.0582×10^{-9}

Case-III: Let S_3 be the event that a set of $c+1$ ILVs (c adjacent ILV pairs) contains one or more faulty ILV pairs. Let M_3 be the event that faults present in the BIST mask any fault(s) that may be present in the ILVs. The probability $P[S_3 \cap M_3]$ can be expressed as:

$$P[S_3 \cap M_3] = \sum_{k=1}^{c} P[S_2 \cap M_2] = \sum_{k=1}^{c} (P_{FM})^k (1-p_m)^{c-k} = (P_{FM} + 1 - p_m)^c - (1-p_m)^c$$

Since $|p_m - P_{FM}| < p_m < 1$ and $|cp_m| << 1$, we can further simplify $P[S_3 \cap M_3]$ as follows:

$$P[S_3 \cap M_3] \approx (1 - c \cdot (p_m - P_{FM})) - (1 - c \cdot p_m) = (1 - c \cdot p_m + c \cdot P_{FM}) - 1 + c \cdot p_m = c \cdot P_{FM}.$$

The above analysis derives closed-form expressions for the masking probability of ILV faults due to multiple faults in the BIST logic. Such expressions can be evaluated quickly to determine the optimum number of ILVs (c) to be tested concurrently by the Dual-BIST engine such that the likelihood of ILV-fault masking is minimized.

5 ENHANCED DUAL-BIST FOR RESISTIVE FAULTS

Targeted BIST for ILVs potentially reduces test escapes of resistive faults and enables on-chip fault localization. Testing of ILVs for resistive faults is essential especially in the field when such faults arise due to aging and other reliability-related root causes such as time-dependent dielectric breakdown between ILVs and electromigration. Regular monitoring of ILV health by BIST can lead to early detection of such failures in the field which, in turn, can facilitate potential repair.

The range of detectable defect size (i.e., magnitude of open or short resistance) can extended by the addition of a delay element—an N_s-stage inverter (where N_s is the number of cascaded single-stage CMOS inverters)—to the input of every ILV in the test mode. Figure 4 describes this design for $N_s = 2$; the buffers are shaded. We next show how the addition of the delay stage facilitates the detection of resistive faults caused by defects of intermediate size.

5.1 Resistive Shorts

A resistive short is a short between ILVs that has a resistance in the intermediate range, typically ranging from few to several hundred KΩs [39]. Figure 5(a) illustrates the lumped circuit model for the testing of a resistive short. The buffer's equivalent resistance is denoted by R_{BUF}, the ILV self-resistance is denoted by R_{ILV}, and R_S denotes the resistance of the short. Complimentary test inputs are applied at the inputs of the buffers—$V_{DD}(1)$ and Ground (0). The differential input voltage of the 2-input XOR is given by $\Delta V_{out} = V_{out,1} - V_{out,2} = V_{DD}R_S/(2R_{ILV} + 2R_{BUF} + R_S)$. A short is detected if and only if it forces the XOR output to 0 instead of 1 (fault-free case). The XOR output is 0 when its differential input voltage is less than a preset threshold. The addition of the buffer stage decreases the magnitude of this differential input voltage since ΔV_{out} for $R_{BUF} > 0$ is

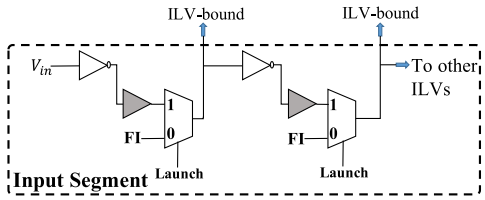


Fig. 4. Enhanced Dual-BIST for resistive faults.

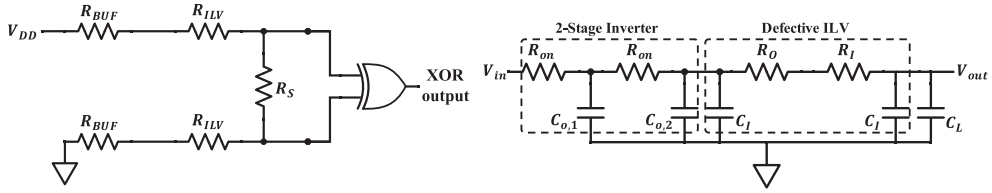


Fig. 5. Circuit models for resistive (a) short, and (b) open.

less than ΔV_{out} for $R_{BUF} = 0$. Therefore, the maximum detectable short resistance can be increased by increasing R_{BUF}.

5.2 Resistive Opens

For a given functional clock period T_{clk}, let the range of opens detected be $[R_{o,min}, \infty)$. The enhanced Dual-BIST can detect open resistances lower than $R_{o,min}$ by adding a buffer delay to the small delay of the open defect. Figure 5(b) illustrates the lumped circuit model of the buffer aiding the test of a resistive open. Here, R_{on} is the ON-resistance of a transistor, $C_{o,x}$ $(x = 1, 2)$ are the output capacitances of the two inverter stages in the buffer, C_I is the parasitic ILV-to-substrate capacitance, C_L is the ILV load (fan-out) capacitance, and R_O denotes the open resistance. Using the Elmore–Delay model for an RC ladder, the total delay, D, of a signal from the buffer input V_{in} to the ILV output V_{out} can be expressed as:

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

A sufficient condition for the detection of an open is given by: $D + \Delta_{XOR} > T_{clk}$, where Δ_{XOR} is the XOR gate delay. Therefore, without decreasing T_{clk}, the minimum detectable open resistance can be decreased beyond $R_{o,min}$ by increasing R_{on}, i.e., by decreasing transistor width or increasing N_s.

Resistive defects can lead to an increase in the propagation delay of paths through the defect sites. The additional delay contribution due to such defects is typically small (hence, SDDs). SDDs 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. Therefore, to detect SDDs, it is critical that the longest (minimum-slack) paths through a fault site are sensitized. Commercial EDA tools attempt to ensure this by using a "constrained" transition delay fault model—only the paths having a nominal timing slack lower than a predefined margin are considered for testing [2]. The insights presented in this section are utilized later in Section 7.4 to design a delay bank for testing SDDs in ILVs.

6 ANALYSIS AND TEST OF CROSSTALK-INDUCED FAULTS

6.1 Analysis of Crosstalk

6.1.1 Sources of Crosstalk. The issue of crosstalk noise arising from TSV coupling is well documented as a principal cause of failure in 3D ICs [50–52]. TSVs can be coupled to an adjacent TSV or proximal active devices. TSV-to-TSV coupling can be classified into two categories: capacitive and inductive coupling [51]. Inductive coupling between TSVs is typically observed in applications with higher data transmission frequencies where significant electromagnetic interference, crosstalk, attenuation, and other signal discontinuity issues are observed [50]. Capacitive coupling is typically dominant for operational frequencies lower than 5 GHz [52]. TSV-to-active device coupling is observed when the substrate contact are placed close to the TSVs. The SiO_2 insulation layer between the TSVs and the silicon substrate has a thickness of a few hundredths of a nanometer, and thus results in high capacitance between the TSV and the silicon substrate. Due to this, high frequency noise in the silicon substrate can be coupled to the TSV through the large capacitance [9]. The **keep-out zone** (**KOZ**) around a TSV is characterized to prevent TSV-induced thermo-mechanical stress from causing **hot-carrier-injection** (**HCI**) and **bias-temperature-instability** (**BTI**) related failures in nearby transistors [12, 17]. The KOZ is not a concern for ILVs because they are much smaller structures (several nm) compared to TSVs (several μm), resulting in lower stress and fewer stress-induced reliability failures.

Various methods have been proposed to mitigate the impact of TSV coupling. Grounded TSVs can be added surrounding the victim TSV [31] or into the empty spots in the design [44]. However, this leads to large area overhead. A coding scheme can be used to reduce the maximum crosstalk in a TSV matrix [28]; however, this is only applicable for capacitive coupling and for a limited mesh size. Clearly, the state-of-the-art mitigation approaches are insufficient to handle the coupling issue. This necessitates a dedicated testing approach for couple-induced fault detection.

The M3D fabrication flow is still immature, and therefore we expect coupling between neighboring ILVs (which are analogous to TSVs in conventional 3D design) and between ILVs and the active circuit to be a critical issue in high-volume manufacture. In addition to increasing path delay, coupling results in additional current flowing through the victim ILV during signal transition. In the presence of electromigration, even a small amount of additional current can significantly reduce the **mean-time-to-failure** (**MTTF**) of the chip. Therefore, the impact of coupling on delay and MTTF of the chip varies depending on: (i) the presence of other defects in the ILVs; (ii) defects in the physical medium between the ILVs; (iii) presence of electromigration; and (iv) physics of the material composition of an ILV and its interaction with the surrounding medium, and several other non-idealities. Therefore, it is not possible to extend the design rules of conventional vias to the ILVs without separate characterization of ILV-to-ILV coupling using specialized BIST.

During early phases of the product tapeout, customers put their designs on **Multi-Part-Wafers** (**MPWs**) to validate them using functional specifications (e.g., target frequency) before doing a product tapeout. Typically, the space is limited on MPWs and having BIST would help validate the designs and provide valuable feedback on the impact of coupling in ILVs. This is a more practical solution than just increasing the ILV spacing far enough in a product tapeout so that the impact of coupling is minimized. Having targeted BIST structures on MPWs for technology qualification is a standard business practice followed by customers [7]. Note that BIST for detecting coupling may no longer be needed in MPWs after several revisions of the design rules.

We present an ILV-specific lumped coupling model and explore the impact of coupling on several performance aspects. Coupling affects the slack on critical paths; therefore, the test efficiency of delay faults can also be impacted.

6.1.2 ILV-to-ILV Coupling. The uppermost metal layer in the bottom M3D tier (UB) and the lowermost metal layer in the top tier (LT) are connected using ILVs. The active layer (containing the transistors) in the top tier and the UB are separated by an ILD of thickness t_{ILD}. The ILVs are manufactured using a flow adapted from that for vertical BEOL plugs in 2D ICs [5]. The fabrication of an ILV is done using a via-etch process followed by diffusion-barrier and seed deposition. Then the ILV is electroplated with a "filler" metal. The filler metal is typically tungsten (W) as its use is compatible with the thermal budget of M3D processing. The diffusion barrier is present around and at the bottom of the ILV, i.e., where the ILV lands, to prevent the ILV metal from diffusing into the surrounding dielectric (silicon) medium and causing unwanted shorts with nearby interconnects (active devices). The diffusion barrier is typically made of **titanium nitride** (**TiN**). Figure 9(a) illustrates the side-view of a two-tier M3D IC with the ILV, ILD, and diffusion barrier annotated.

In a 1D array of ILVs, an ILV can be coupled to at most two adjacent ILVs, i.e., ones to the left and right. We model the electrostatic coupling between an ILV and its two adjacent ILVs using lumped resistances and capacitances; the approach is similar to what is commonly adopted for TSV-to-TSV coupling [31]. Figure 6 shows the proposed lumped ILV-to-ILV coupling model. In the three-ILV coupling model, the ILV in the middle is considered to be a *victim* and the adjacent ILVs are considered as *aggressors*. Considering 2–4 GHz as the upper-end operational frequency of our M3D designs, inductances are not included in the ILV-to-ILV coupling model. The fabrication-induced taper of an ILV does not significantly affect the model's parameter values and is not considered in the calculations. An ILV is considered to have a cuboidal shape with a square base of dimensions $w \times w$ and height $t_{ILD} + t_{Si}$, where t_{Si} is the thickness of the top tier's active silicon layer. Let the thickness of the diffusion barrier, surrounding the entire length of the ILV, be t^s_{db}. Let the thickness of the diffusion barrier at the bottom of the ILV, i.e., near the landing pad on UB, be t^b_{db}. Therefore, the length of the ILV t_W containing filler-metal tungsten is $t_{ILD} + t_{Si} - t^b_{db}$; the square base occupied by tungsten has a side-length of $w - 2t^s_{db}$.

Before connecting to LT in the top tier, the ILV passes through the silicon-based active layer of thickness t_{Si}. Therefore, we can divide the ILV into three segments based on the ILV composition and surrounding medium:

- **A:** This segment constitutes the lower portion of the ILV near the UB. The segment contains only the diffusion barrier, is surrounded by ILD, and extends for a length of t^b_{db}.
- **B:** This segment constitutes the bottom portion of the ILV near the UB. The segment contains tungsten surrounded by the diffusion barrier, is surrounded by ILD, and extends for a length of $t_W - t_{Si}$.
- **C:** This segment constitutes the upper portion of the ILV near the LT. The segment contains tungsten surrounded by the diffusion barrier, is surrounded by the top tier's active layer, and extends for a length of t_{Si}.

6.1.3 Derivation and Validation of Coupling Parasitics. We express the lumped resistances and capacitances of each segment of the coupling model as functions of the physical dimensions defined in Section 6.1.2. Let the separation between victim and aggressor ILVs be d. The resistance R^A_{db} of the diffusion barrier in Segment A is: $R^A_{db} = \frac{\rho_{db} \cdot t^b_{db}}{w^2}$, where ρ_{db} is the resistivity of the diffusion-barrier material. The coupling capacitance C^A_{cc} of Segment A between victim and aggressor ILVs is: $C^A_{cc} = \frac{\epsilon_{ILD} \cdot w \cdot t^b_{db}}{d}$, where ϵ_{ILD} is the permittivity of the ILD.

In Segment B, the resistance R^B_{db} of the diffusion barrier is: $R^B_{db} = \frac{\rho_{db} \cdot (t_W - t_{Si})}{w^2 - (w - 2t^s_{db})^2}$. The resistance R^B_W of the ILV filler-metal is: $R^B_W = \frac{\rho_W \cdot (t_W - t_{Si})}{(w - 2t^s_{db})^2}$, where ρ_W is the resistivity of the filler-metal. The coupling capacitance C^B_{cc} of Segment B between victim and aggressor ILVs is: $C^B_{cc} = \frac{\epsilon_{ILD} \cdot w \cdot (t_W - t_{Si})}{d}$.

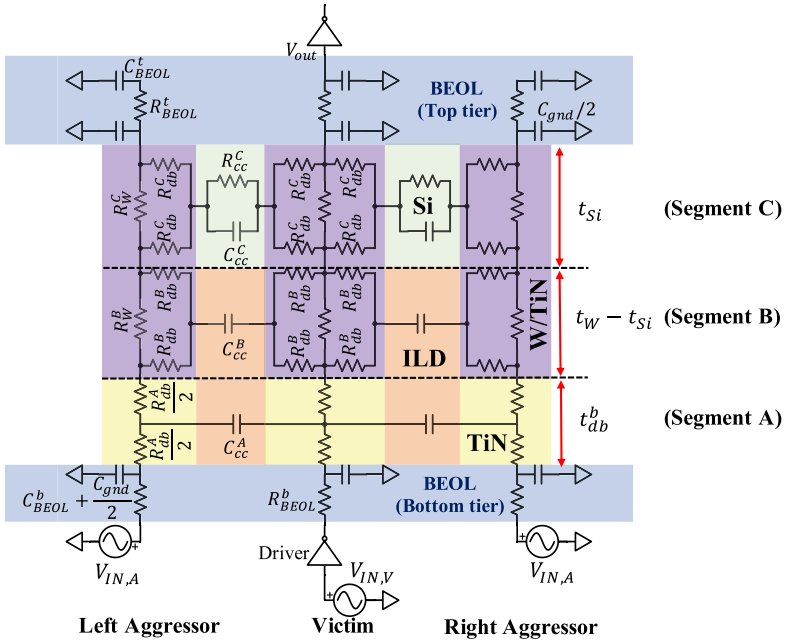


Fig. 6. ILV-to-ILV lumped coupling model for a victim ILV with two aggressor ILVs.

In Segment C, the resistance R^C_{db} of the diffusion barrier is: $R^C_{db} = \frac{\rho_{db} \cdot t_{Si}}{w^2 - (w - 2t^s_{db})^2}$. The resistance R^C_W of the ILV filler-metal is: $R^C_W = \frac{\rho_W \cdot t_{Si}}{(w - 2t^s_{db})^2}$. The coupling capacitance C^C_{cc} of Segment C between victim and aggressor ILVs is: $C^C_{cc} = \frac{\epsilon_{ILD} \cdot w \cdot t_{Si}}{d}$. The coupling resistance R^C_{cc} of Segment C between victim and aggressor ILVs is: $R^C_{cc} = \frac{\rho_{Si} \cdot d}{w \cdot t_{Si}}$, where ρ_{Si} is silicon resistivity. The coupling resistance is in parallel with the coupling capacitance. As the ILD resistance is very high, coupling resistances in Segments A and B are neglected.

The ILV-to-ground lumped capacitance C_{gnd} is approximated by $c_{BEOL} \cdot t_{ILD}$, where c_{BEOL} is the average BEOL wire capacitance (wire-to-ground and wire-to-wire combined) per unit length. To model a full flop-to-flop functional path, P, on which an ILV lies, the total BEOL wire capacitance C^b_{BEOL} of the portion of functional path in the bottom tier is given as: $C^b_{BEOL} = FO_b \cdot c_{BEOL} \cdot w_{BEOL}$. Here, FO_b is the number of fan-out branches of P in the bottom tier and w_{BEOL} is the average length of a fan-out branch. The total BEOL wire capacitance C^t_{BEOL} of the portion of functional path in the top tier is given as: $C^t_{BEOL} = FO_t \cdot c_{BEOL} \cdot w_{BEOL}$. Note that, FO_t is the number of fan-out branches of P in the top tier. Similarly, the total BEOL wire resistances, R^t_{BEOL} and R^b_{BEOL} (corresponding to top and bottom tiers, respectively), of the functional path are given as: $R^x_{BEOL} = FO_x \cdot r_{BEOL} \cdot w_{BEOL}$. Here, $x \in \{t, b\}$ and r_{BEOL} is average BEOL wire resistance per unit length. All fan-out branches in the top and bottom tiers are loaded with minimum-width inverters.

The signal $V_{IN,V}$ to the input of a driver gate in the bottom tier drives the victim ILV. To evaluate the impact of coupling on signal transition, we observe the signal propagated by the victim ILV by tapping the input of one of the load inverters (V_{out}) in the top tier. Both left and right aggressor ILVs are tied to signal $V_{IN,A}$ in order to maximize the impact of coupling on the victim ILV for worst-case analysis.

Table 2 summarizes the notations of the parameters used in the ILV-to-ILV and ILV-to-active device coupling models, and the corresponding values used for HSPICE simulations of the models. The chosen parameter values are based on 45 nm technology node [19, 29, 47].

Table 2. Notations for Parameters Used in the Coupling Models and Parameter Values Used for Simulations

Parameter notation	Description	Parameter value used for HSPICE simulations
t_{ILD}	ILD thickness	110 nm
t_{Si}	Thickness of silicon layer in top tier	30 nm
w	Width of ILV's square-shaped base	70 nm
t_W	Height of ILV portion containing filler metal	125 nm
ρ_{db}, ρ_W, ρ_{Si}	Resistivity of {diffusion barrier, filler, silicon}	250 Ω-nm (TiN), 56 Ω-nm (W), 6.4×10^{11} Ω-nm
ϵ_{ILD}	Permittivity of ILD	$3.9\epsilon_o$ F/m (silicon dioxide)
r_{BEOL}, c_{BEOL}	Average resistance and capacitance per unit BEOL wirelength	0.24 Ω/μm, 0.35 fF/μm
w_{BEOL}	Average BEOL wirelength of a fan-out branch	1 μm
FO_t, FO_b	Fan-outs of ILV path in top and bottom tiers	{1, 2, 3}, {1, 2, 3}
d	Separation between adjacent ILVs	≥70 nm
d_{Si}	Separation between ILV and body contact	≥71 nm
w_b	Width of square-shaped body contact	70 nm

ϵ_o: permittivity of free space, $\epsilon_o = 8.85 \times 10^{-12}$ F/m.

We next use a mesh-based high-frequency simulator ADS (Advanced Design System software from Keysight Technologies) to validate the RC connectivity and derivation used in the lumped model. We use the "Momentum RF" solver of ADS to build the model for ILV-to-ILV coupling and perform S-parameter analysis. We then compare the coupling S-parameters of the Momentum extracted model with that of our lumped model. As vias made of composite materials (such as tungsten filler enclosed within TiN diffusion barrier) cannot be modeled in ADS, we carry out a piece-wise verification of the three segments (A, B, and C) of our lumped model by only considering barrier-to-barrier coupling and ignoring the filler metal, which does not appear in the path of the coupling current. Figure 7(a) illustrates the four-port via-to-via coupling model implemented in ADS. Each via is tapered cylindrical and made of TiN. The vias have smaller diameter D_l (in nm) and larger diameter $D_h = D_l + 10$. Square-shaped landing pads of width $(D_h + 10)$ are added on either end of a via. The height of each via is L_{ILV} which equals the ILV height in a given segment. As fabricated ILVs are cuboidal in shape and ADS vias are cylindrical, D_l is set such that the effective volume of the ADS via equals that of the TiN barrier in an ILV in a given segment. For example, in Segments B and C, volume of the TiN barrier equals $L_{ILV} \cdot (w^2 - (w - t^s_{db})^2)$. Hence, the diameter D_l of an equivalent cylindrical via of height L_{ILV} is calculated as: $D_l = 2\sqrt{\frac{w^2-(w-t^s_{db})^2}{\pi}}$. For Segment A, $D_l = 2\sqrt{\frac{w^2}{\pi}}$. The dielectric medium between the vias is chosen as Si or SiO_2 depending on the segment under verification. We analyze the four-port network comprising adjacent vias to obtain the S-parameters S_{13} and S_{14} (in dB) for the backward and forward coupling powers, respectively. For S-parameter analysis, a termination load of 50 Ω is added to each of the four ports. The operating frequency is swept from 0.5 GHz to 3 GHz.

Figure 7(b)–(c) shows the comparison in coupling S-parameters between the ADS extracted model and our lumped model for Segments A and C. The discrepancy in the coupling power estimate between the extracted and lumped models ranges between 5 dB and 15 dB. Such a discrepancy arises from differences in the physical shape of an ILV assumed by extracted and lumped models. Given the extremely low magnitudes of the coupling powers estimated by both the models, a discrepancy of <15 dB is acceptable. For Segment B, the lumped model estimates a low magnitude of the coupling power with discrepancy ~5 dB. Therefore, the above segment-wise verification validates the connectivity and derivation procedure of the RC parasitics in our lumped model. As the three segments are connected in series in the lumped model and the coupling paths primarily include the diffusion barrier, segment-wise verification is sufficient to verify the entire lumped model.

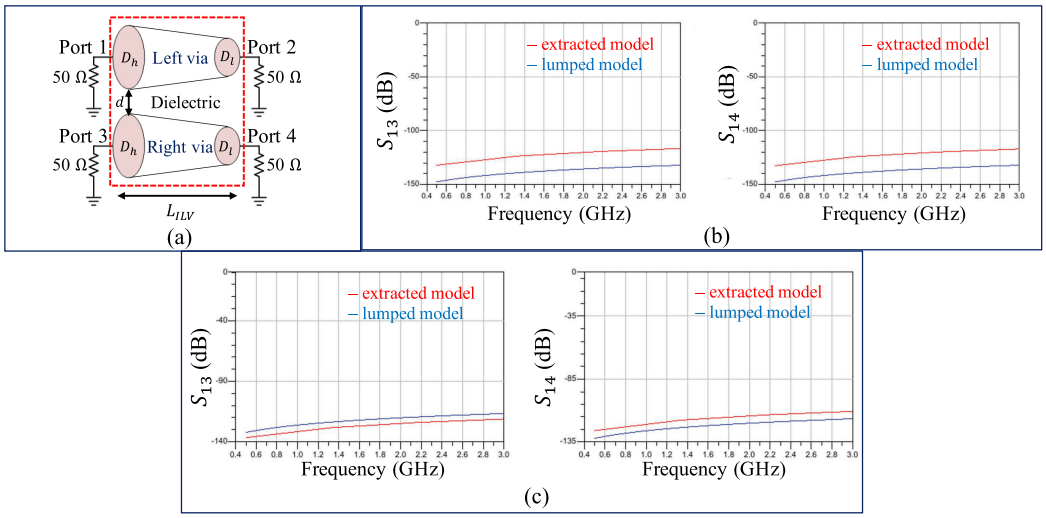


Fig. 7. (a) Via-to-via coupling model implemented in ADS; (b) S-parameter comparison for Segment A; and (c) S-parameter comparison for Segment C.

6.1.4 Impact of ILV-to-ILV Coupling on Timing. Table 3(a) gives the impact of ILV-to-ILV coupling on the delay of the functional path through the victim ILV for different choices of FO_b, FO_t, and driver gate. The driver gates considered are inverter (INVX1), buffer (BUFX2), and 2-input NAND (NANDX1) from Nangate 45 nm open-cell library. The low-to-high (high-to-low) propagation delay of the victim ILV when the aggressor ILVs undergo low-to-high (high-to-low) transition is denoted by t_r^r (t_f^f). The high-to-low (low-to-high) propagation delay of the victim ILV when the aggressor ILVs undergo low-to-high (high-to-low) transition is denoted by t_f^r (t_r^f). The symbol Δ before the delay terms indicates the change in the victim's propagation delay due to coupling; positive (negative) polarity of Δ implies increase (decrease) in the delay. All of the above delay-related terms are expressed as percentages of the functional clock period T_{clk}. Here, T_{clk} is 1 ns, and the rise and fall delays of $V_{IN,V}$ and $V_{IN,A}$ are set at 50 ps. The impact of coupling is seen to be negligible.

Table 3(b) gives the impact of ILV-to-ILV coupling in the presence of an open defect in the victim ILV for $FO_b = FO_t = 3$ and BUFX2 driver. The open defect is modeled as a resistance R_d, which is inserted in series with $R_{db}^A/2$ in Segment A of the ILV model (see Figure 6). The open increases the propagation delay of the victim ILV. Moreover, the adverse impact of the open on ILV delay is further amplified by the delay introduced by coupling. In the presence of coupling, larger-sized defects with higher values of R_d tend to have a higher impact on the victim ILV's propagation delay.

Figure 8 shows the impact of increasing ILV separation d on the timing impact of coupling denoted by Δt_f^r (expressed as a percentage of the functional clock period T_{clk}=1 ns). Here, Δt_f^r indicates the change in the high-to-low propagation delay of the victim ILV when the aggressor ILVs undergo a low-to-high transition. The victim ILV is driven by BUFX2 and has a fan-out of six. Note that the nominal separation between two ILVs is 70 nm for the technology node under consideration. We observe that the increase in ILV delay drops sharply when the separation is increased beyond 100 nm.

6.1.5 ILV-to-Active Device Coupling. If an ILV is fabricated in close proximity to an active device, such as the transistor, in the top tier, coupling current can flow between the ILV and the body

Table 3. Impact of ILV-to-ILV Coupling on the Delay of Victim ILV

(a) Open defect absent

Driver	FO_b	FO_t	Δt_r^r (%)	Δt_f^f (%)	Δt_r^f (%)	Δt_f^r (%)
INVX1	1	1	0.02	0.04	0.003	−0.021
	2	2	−0.001	0.002	0.01	0.005
	3	3	0.008	−0.008	0.006	0.012
BUFX2	1	1	0.002	−0.002	0.002	0.001
	2	2	0.002	−0.001	0.002	0.002
	3	3	0.001	0	0.001	0.2
NANDX1	1	1	0.001	−0.002	−0.002	−0.005
	2	2	0	0.01	0.003	−0.03
	3	3	0	0.03	0.006	0.001

(b) Open defect present

R_d (KΩ)	No coupling		Coupling		Impact	
	t_r^f (%)	t_f^r (%)	t_r^f (%)	t_f^r (%)	Δt_r^f (%)	Δt_f^r (%)
5	3.45	3.52	3.57	3.53	0.11	0.01
10	4.78	4.64	4.82	4.83	0.04	0.2
20	7.73	7.86	7.91	7.92	0.2	0.06
40	13.71	14.12	13.91	14.2	0.2	0.1
50	17.81	15.91	16.82	17.11	−1	1.2

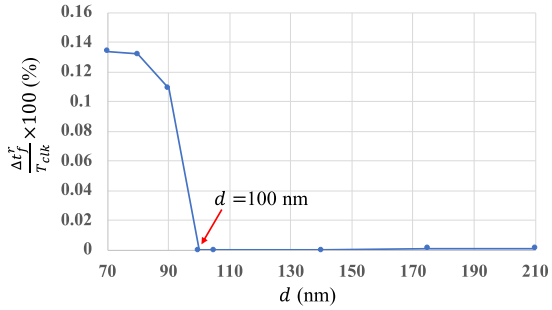


Fig. 8. Effect of increasing ILV-to-ILV separation on the timing impact of coupling.

contact of the transistor through the silicon layer [10]. Assuming a p-type substrate for the silicon epitaxial layer (epi-layer) in the top tier, the body contacts of n-type transistors are especially susceptible to crosstalk. The p-type transistors are enclosed in an n-well and are naturally shielded from crosstalk. All n-type transistors of a standard cell share one body contact that is tied to the ground.

When the side-walls of an ILV and its adjacent body contact are aligned face-to-face, the nature of the electrostatic field lines between the contact and the ILV is similar to that of the fringing field lines between a metal interconnect's side-wall and the ground plane, when rotated anti-clockwise by 90°; see Figure 9(a–c). Therefore, the coupling capacitance C_{Si} between ILV and the n-type transistor's body contact is: $C_{Si} = \frac{\pi \epsilon_{Si} w_b}{\ln d_{Si}/w_b}$. Here, w_b is the width of the square-shaped contact and d_{Si} is the ILV-to-contact separation; $d_{Si} > w_b$ such that the capacitance is a positive value. As the body contact is placed on top of the epi-layer, a large proportion of the coupling current flows near the silicon surface. Therefore, the expression for the coupling resistance through silicon can be adapted from that of the semiconductor sheet-resistance as measured by four-point probe method [20]. The coupling resistance R_{Si} is given as: $R_{Si} = \frac{\rho_{Si}}{\pi t_{Si}} \cdot \ln(\sinh \frac{t_{Si}}{d_{Si}} / \sinh \frac{t_{Si}}{2d_{Si}})$. Figure 9(d) illustrates the ILV-to-active device lumped coupling model for an inverter.

Table 4 gives the impact of ILV-to-active device coupling on the ILV delay in the presence of an open defect. The propagation delay of a rising (falling) transition on the ILV is denoted by t_r (t_f) and is expressed as a percentage of T_{clk}; T_{clk} is 1 ns. The impact of coupling is significantly higher in the presence of larger-sized defects (or higher values of R_d).

We also investigate the effect of increasing ILV-to-active device separation on ILV timing. According to the standard design rules for any technology node, the separation must exceed the contact width, i.e., $d_{Si} > w_b$. As the contact width is 70 nm, we set the minimum value of d_{Si}

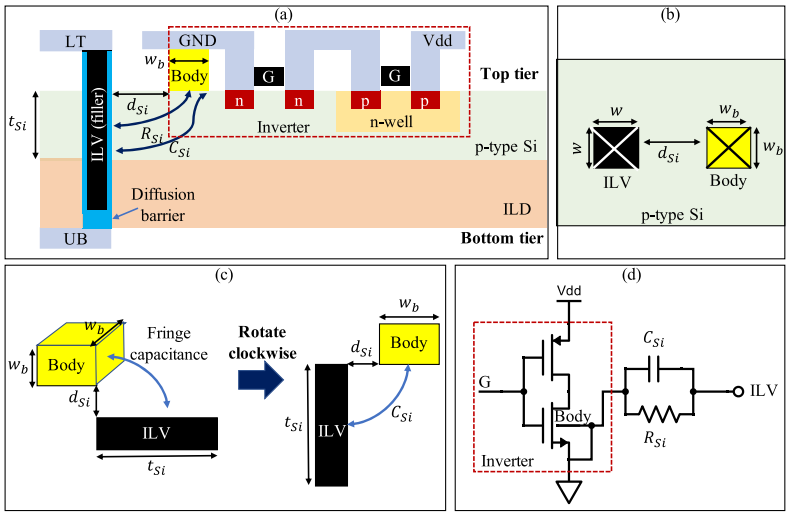


Fig. 9. ILV-to-active device lumped coupling model: (a) side view of a two-tier M3D–IC showing the ILV and substrate (body) contact of an inverter in close proximity; (b) top view of the placement of ILV and body contacts in the top tier, that is considered for analyzing coupling; (c) illustration of the derivation of C_{Si} from fringe-capacitance model; and (d) schematic of the lumped coupling model involving the ILV and the inverter.

Table 4. Impact of ILV-to-active Device Coupling on the Delay of Victim ILV

R_d (KΩ)	No coupling		Coupling		Impact	
	t_r (%)	t_f (%)	t_r (%)	t_f (%)	Δt_r (%)	Δt_f (%)
0	2.18	2.05	2.26	2.2	0.07	0.15
5	3.45	3.52	4.05	3.97	0.6	0.45
10	4.78	4.64	6	5.94	1.2	1.3
20	7.73	7.86	10.2	10.13	2.47	2.27
40	13.71	14.12	18	18.33	4.3	4.21

to 71 nm. Figure 10 shows the impact of increasing d_{Si} on the coupling-induced delay in an ILV during low-to-high transition. The increase in the low-to-high propagation delay of the victim ILV due to coupling is denoted by Δt_r and is expressed as a percentage of T_{clk}. We observe that the timing impact drops sharply as d_{Si} is increased beyond 90 nm. As the coupling capacitance between ILV and active device has a logarithmic dependence on d_{Si}, d_{Si} must be significantly increased to further mitigate the timing impact of coupling.

6.2 Impact of Coupling on Statistical Delay Quality Level

The **Statistical Delay Quality Level** (**SDQL**) is a commonly used surrogate metric for SDD coverage and test pattern grading [53]; a lower value of SDQL signifies higher effectiveness of the SDD test patterns. To simulate the effectiveness of test patterns under process variations, a delay-defect distribution function $F(z)$ is considered, where z is the delay fault size. Therefore, $F(z)$ denotes the probability that either a rising or a falling transition delay fault of size z exists at a random node. Note that $F(z)$ is typically obtained from empirical data or a process test chip; some examples of

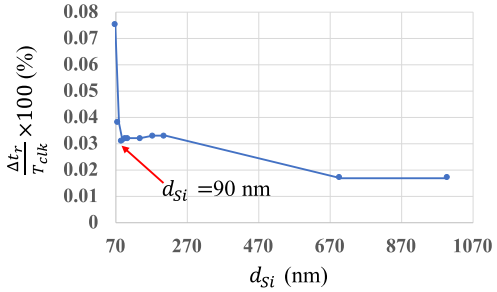


Fig. 10. Effect of increasing ILV-to-transistor separation on the timing impact of coupling.

such distributions can be found in [42]. SDD detection becomes more relevant when more defects are of small size, i.e., $F(z)$ decreases rapidly with increasing z.

Consider an SDD fault X and a corresponding test pattern TP_X for it. Let the timing margin of the longest (minimum-slack) path through X in the absence of any parameter variations be given by $T_m^\star$. Similarly, let the nominal slack through the path sensitized by TP_X be $T_d^\star$. Therefore, faults of size $z < T_m^\star$ are redundant whereas all faults of size $z > T_d^\star$ can be detected be TP_X. The probability that an irredundant fault of size z at X remains undetected is therefore given by $P_X^\star = \int_{T_m^\star}^{T_d^\star} F(z)dz$. The nominal SDQL for a test pattern set is then given by

$$\theta^\star = \sum_{j=1}^{2N} P_{X_j}^\star = \sum_{j=1}^{2N} \int_{T_m^\star}^{T_d^\star} F(z)dz, \tag{1}$$

where the number of nodes (delay faults) in the netlist is $N(2N)$. The probability of SDD escape decreases with decreasing $\theta^\star$. Conventional SDD ATPG tools attempt to generate a test-pattern set that minimizes $\theta^\star$.

As given in Table 3(a), the presence of coupling among neighboring ILVs has a small impact on the propagation delay of the affected paths when no resistive open or SDD is present in those paths. Therefore, it is likely that coupling alone will not result in transition delay faults. However, as the slacks on the critical paths change, already existing SDDs can escape detection. Consider a distribution of SDDs as shown in Figure 11. $T_m^\star$ ($T_d^\star$) denotes the timing margin of the longest (sensitized) path through a SDD, and T_m (T_d) is the corresponding slack in the presence of coupling. Clearly, $T_m \leq T_m^\star$ and $T_d \leq T_d^\star$. The SDQL in the presence of coupling is then given by

$$\theta = \sum_{j=1}^{2N} \int_{T_m}^{T_d} F(z)dz = \theta^\star + \sum_{j=1}^{2N} \left(\int_{T_m^\star}^{T_m} F(z)dz - \int_{T_d^\star}^{T_d} F(z)dz \right) = \theta^\star + \sum_{j=1}^{2N} \Delta\theta_j. \tag{2}$$

Here, $\Delta\theta_j$ denotes the change in the area under the curve for the j^{th} SDD. Note that three cases can arise here:

- Case-1: Both the critical and sensitized path through the j^{th} SDD are equally affected by coupling faults. In this case, $T_m - T_m^\star = T_d - T_d^\star$. As $F(z)$ is monotonically decreasing, $\Delta\theta_j > 0$.
- Case-2: The critical path is more affected by coupling faults, while the sensitized path is not (or less) affected. In this case, $T_m - T_m^\star > T_d - T_d^\star$. Similar to Case-1, $\Delta\theta_j > 0$.
- Case-3: The sensitized path is more affected by coupling faults, while the critical path is not (or less) affected. In this case, $T_m - T_m^\star < T_d - T_d^\star$. $\Delta\theta_j$ can be positive or negative depending on the value of T_m and T_d.

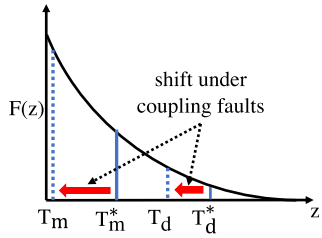


Fig. 11. Change in area under the SDQL curve due to ILV coupling.

Table 5. SDD Pattern Effectiveness in Terms of SDQL Values for Different Values of f_{couple}

Benchmark	SDQL for SDDs in ILVs ($\times 10^6$)				SDQL for all SDDs ($\times 10^6$)			
	f_{couple}				f_{couple}			
	0.0	0.1	0.15	0.2	0.0	0.1	0.15	0.2
AES-128 (400 ILVs)	4.60	5.28	5.37	5.38	48.29	48.66	48.01	49.09
AVC-Nova (316 ILVs)	0.03	0.16	0.16	0.16	12.04	35.20	35.20	35.21

The SDQL (and hence, the probability of test escape) for an SDD increases under coupling in Case-1 and Case-2, while it decreases in Case-3. However, note that Case-1 holds for all SDDs on the ILVs. Therefore, it is expected that coupling faults will result in increased test escape of SDDs on ILVs, while there is a possibility that the testing of other SDDs (not on ILVs) can be degraded as well.

Table 5 gives the impact of coupling fault on the SDD test efficiency in the AES and Nova benchmarks. We observed that, in the presence of ILV coupling, the propagation delay increases by 0.2% of the rated clock period (see Table 3(a)). Therefore, we have injected delay defects of this size on a randomly-selected subset of ILVs to simulate the effect of coupling. SDQL is then calculated for SDDs in the ILVs and in the entire design. In our experiments, we have considered that a fraction, given by f_{couple}, of all ILVs is affected by coupling. We present results for four values of $f_{couple} - 0$ (no coupling), 0.1, 0.15, and 0.2. To simplify the model, we have considered all the injected coupling faults to be of similar size. From Table 5, we observe that for SDDs in ILVs, the SDQL (test efficiency) increases (decreases) as f_{couple} increases. The impact of such SDDs, however, depend on the circuit topology. For example, the SDQL for AES, under SDDs in ILVs, increases at most by 16.96 % (for $f_{couple} = 0.2$). On the other hand, for Nova, the SDQL increases to 5× its nominal value. Note also that the impact of coupling on the SDQL does not increase significantly as f_{couple} increases. This can be attributed to the constant size of the coupling fault.

Recall that SDD fault sites not located on ILVs can belong to any of the three cases mentioned above. Therefore, the SDQL of such faults can either increase or decrease due to coupling. This is observed for the AES benchmark for $f_{couple} = 0.15$ (see Table 5) when SDQL for all SDDs are considered; in this case the SDD test efficiency improves under coupling. However, for the other cases, SDQL increases from its baseline value. Therefore, it is possible that SDD test efficiency of faults not on the ILVs can also be affected due to ILV coupling. Clearly, SDD testing in M3D–ICs becomes unpredictable in the presence of ILV-to-ILV coupling. Emerging technologies such as M3D–ICs are prone to SDDs due to inevitable process variations and manufacturing defects. A targeted BIST approach for detecting coupling-induced faults in ILVs is necessary to avoid SDD test escape.

6.3 Test Generation for Coupling-Induced Faults

Faults arising due to coupling between adjacent ILVs, i.e., victim-aggressor pair, are glitch and transition (delay) faults. The **maximal aggressor fault** (**MAF**) model is used to generate test

Table 6. Coverage Matrix for Mapping Test Patterns in T to the Corresponding Faults Covered in an ILV Pair

ILV	Fault model	Pattern transitions (ordered pair) in T											
		P^1, P^2	P^2, P^3	P^3, P^4	P^4, P^5	P^5, P^6	P^6, P^7	P^7, P^8	P^8, P^9	P^9, P^{10}	P^{10}, P^{11}	P^{11}, P^{12}	P^{12}, P^{13}
Left	Delay (coupling)		✓			✓		✓	✓	✓	✓	✓	✓
	Glitch (coupling)	✓		✓	✓		✓						
	Open (delay)		✓			✓		✓	✓	✓	✓	✓	✓
	Stuck-at-0	✓				✓	✓		✓		✓		✓
	Stuck-at-1		✓	✓	✓			✓		✓		✓	
	Shorted to right ILV	✓	✓		✓	✓			✓			✓	
Right	Delay (coupling)	✓	✓	✓	✓	✓	✓	✓			✓		
	Glitch (coupling)								✓	✓		✓	✓
	Open	✓	✓	✓	✓	✓	✓	✓			✓		
	Stuck-at-0		✓		✓		✓				✓	✓	✓
	Stuck-at-1	✓		✓		✓		✓	✓	✓			
	Shorted to left ILV	✓	✓		✓	✓			✓			✓	

patterns for detecting coupling-induced faults between interconnects by considering one interconnect as the victim and all neighboring interconnects as aggressors [11]. The same fault model can be adapted for test generation for coupling-induced faults in ILVs. When the victim ILV is propagating a signal $s \in \{0, 1\}$ and the aggressor ILV is undergoing a rise or fall transition, the potential corresponding to s can fluctuate leading to a glitch, and cause circuit misbehavior and increased switching-power dissipation. On the other hand, if signals on both victim and aggressor ILVs simultaneously undergo transitions, the signal propagated by the victim ILV can potentially suffer from a delay fault, leading to incorrect logic propagation.

We next determine the sequence consisting of a minimum number of test patterns to detect all possible coupling-induced faults in a 1D ILV array where the "victim" and "aggressor" terms can be interchangeably associated with any given ILV. In other words, the same ILV can act as both victim and aggressor in different cycles of the test sequence. Theorem 3 in the appendix shows that the number of test patterns (both necessary and sufficient) required to detect all possible coupling-induced faults (based on the MAF model) in a 1D ILV array is 13.

The test sequence T for detecting coupling-induced faults in the 1D ILV array is:

$$T = (11, 10, 01, 00, 01, 10, 11, 00, 10, 00, 11, 01, 11).$$

Here, the ith pattern P^i $(1 \leq i \leq 13)$ in T is represented using two bits q_1^i and q_0^i, where q_1^i (q_0^i) is the logic value applied to the left (right) ILV. As each ILV (when considered as the victim) has at most two aggressors on its left and right, alternate ILVs in the ILV bus receive the same pattern bit q_j^i $(j \in \{0, 1\})$ in test mode. In addition to coupling-induced faults, the patterns in T also cover stuck-at faults, shorts, and opens in the ILVs. Table 6 maps the transitions in T to the corresponding faults covered in adjacent (left and right) ILVs.

7 SHARED-BIST ARCHITECTURE FOR FAULT DETECTION AND LOCALIZATION

7.1 Shared-BIST Architecture

The Shared-BIST architecture is designed to enable both detection and localization of faults in the ILVs under test. The architecture has two main components: (i) BIST Launch engine and (ii) BIST Capture engine. The BIST engine tests up to n ILVs concurrently where all ILVs propagate signal in the same direction (up-going or down-going). Here, n is referred to as the *width* of the BIST engine and is typically a power of 2, i.e., $n = 2^k$ where k is a non-negative integer. For up-going (down-going) ILVs traveling from bottom (top) to top (bottom) tier, the BIST Launch is in the bottom (top) tier and the BIST Capture is in the top (bottom) tier. Without loss of generality, we focus on up-going ILVs for the rest of the description of the Shared-BIST architecture.

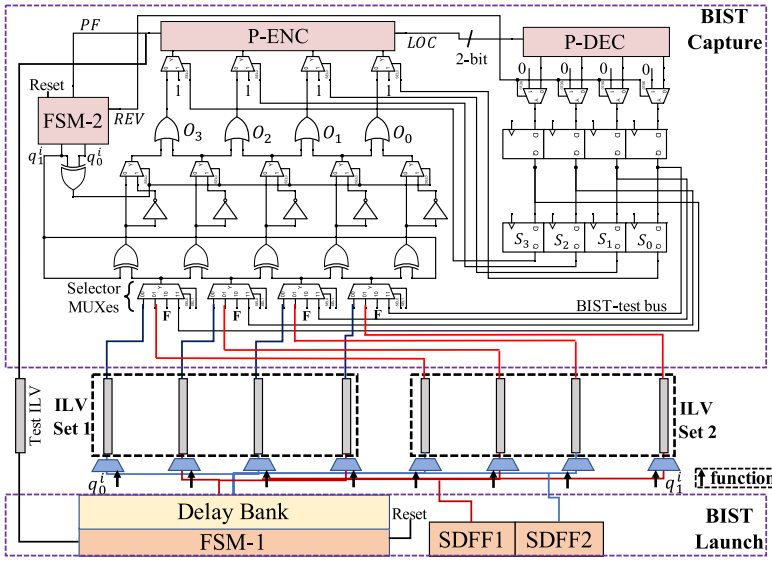
In BIST Launch, a finite-state machine, referred to as FSM-1, generates the test sequence T. The FSM-1 block produces two outputs, q_1^i and q_0^i, that are tied to the inputs of two delay banks. The delay bank adds additional delay to the test path through the ILV-under-test for testing SDDs. The delayed pattern is then passed to the test-mode inputs of 2:1 MUXes feeding alternate up-going ILVs.

The BIST Capture has 2^k *selector* MUXes tied to the outputs of the up-going ILVs. To minimize area overhead, the Shared-BIST architecture enables multiple ILV buses (or sets) to share one BIST engine via time-multiplexing. For time-multiplexing the test of multiple ILV sets, the selector MUXes have at least as many inputs as the number of ILV sets sharing one BIST engine. If there are m BIST engines and s sets of up-going ILVs, where each set contains at most 2^k ILVs, each selector MUX has at least $\lceil s/m \rceil$ inputs. The outputs of adjacent selector MUXes feed an XOR gate. For 2^k MUXes, there are $2^k + 1$ XOR gates. The left (right) pin of the leftmost (rightmost) XOR gate is tied to the output of a second FSM block, referred to as FSM-2. The FSM-2 block also generates the test sequence T synchronously with FSM-1 but its outputs are directly tied only to the leftmost and rightmost XOR gates in the BIST Capture engine. The $2^k + 1$ XOR outputs are also inverted to produce $2^k + 1$ XNOR outputs. There are 2^k OR gates following the XOR/XNOR layer. The inputs of each OR gate are tied to the outputs of adjacent XOR/XNOR gates via 2:1 MUXes. The select signal of these MUXes is driven by FSM-2. If $q_1^i \oplus q_0^i = 1(0)$, the XOR (XNOR) output is passed to the OR gate. The outputs of OR gates feed a priority encoder block, referred to as P-ENC, via 2:1 MUXes. Therefore, the P-ENC block has 2^k inputs and produces $k + 1$ outputs.

The **least significant bit** (**LSB**) of the P-ENC output bus, *PF*, indicates if the input bus of P-ENC contains at least one "0"-carrying bit. The remaining k output bits, collectively called *LOC*, indicate the position of a "0" bit in the 2^k-bit input bus of P-ENC in the binary format, considering the output of the leftmost OR gate as the **most significant bit** (**MSB**). For example, for $k = 2$, the input bus to the P-ENC has four bits. If the MSB of the input bus is "0", the P-ENC produces "111" as the output: *PF* is "1" and *LOC* is "11".

The *LOC* output of P-ENC feeds a priority decoder, referred to as P-DEC. The P-DEC block has k inputs and 2^k outputs. The outputs are tied to a 2^k-bit wide **first-in-first-out** (**FIFO**) block of depth 2. The first (second) stage of the FIFO is referred to as *ping* (*pong*) register. The outputs of P-DEC are connected to the ping register via 2^k 2:1 MUXes. The select lines of these MUXes are driven by a signal *REV* produced by FSM-2. The 2^k parallel outputs of the pong register feed the select inputs of the 2^k 2:1 MUXes at the input of P-ENC. The output bus of the pong register serves as *bypass-select* lines. A "0" select-input passes the OR outputs into the P-ENC; a "1" select-input bypasses the OR outputs to send "1" to the P-ENC. The *PF* output of P-ENC is fed back to FSM-1 and FSM-2 blocks to generate the next test pattern. Note that a test ILV is added to propagate *PF* from BIST Capture in the top tier to FSM-1 in the bottom tier. Figure 12 illustrates the Shared-BIST architecture for detection and on-chip localization of ILV faults.

The 2^k-bit BIST-test bus from ping register to selector MUX inputs enables isolated testing of the BIST Capture engine. Dedicated flops—SDFF1 and SDFF2—are added to the q_1^i and q_0^i outputs of the delay bank for isolated testing of BIST Launch. All registers in the Shared-BIST engine are stitched into a scan chain that enables isolated testing of the Shared-BIST logic, including the test ILV, in *BIST-test* mode before using the BIST to test functional ILVs. The selector MUX has $m + 2$ inputs to switch between ILV sets, BIST-test mode, and functional mode. The select line of the MUX is driven by an on-chip test controller. In the functional mode when the ILVs are not tested, the MUX output is tied to a constant value F (F $\in \{0, 1\}$) to reduce power dissipation by preventing the BIST Capture logic from switching; the BIST Launch is also frozen by activating the Reset input to FSM-1.



Truth Table for P-ENC

O_3	O_2	O_1	O_0	$LOC[1]$	$LOC[0]$	PF
1	1	1	1	X	X	0
1	1	1	0	0	0	1
1	1	0	X	0	1	1
1	0	X	X	1	0	1
0	X	X	X	1	1	1

Truth Table for P-DEC

$LOC[1]$	$LOC[0]$	PF	S_3	S_2	S_1	S_0
X	X	0	0	0	0	0
0	0	1	1	1	1	1
0	1	1	1	1	1	0
1	0	1	1	1	0	0
1	1	1	1	0	0	0

Fig. 12. Block diagram of the Shared-BIST architecture.

7.2 Fault Detection and Localization

7.2.1 Pattern Generation using FSM. The output PF of P-ENC indicates if the ILV set under test "passed" or "failed" the test when a forward pattern transition $P^{i-1} \rightarrow P^i$ is applied by BIST Launch. If no fault is detected by the transition, $PF = 0$; otherwise $PF = 1$. The location of the detected fault is indicated by the LOC output of P-ENC. The LOC provides the location of the highest-priority bit carrying 0 in the input bus of P-ENC, which is the output of the OR layer. As a result, in a single test cycle, the location of only one 0-carrying bit is returned by P-ENC.

It is possible that the same transition $P^{i-1} \rightarrow P^i$ detects multiple faults in the ILV set, leading to multiple 0s being propagated to P-ENC. Therefore, if a fault is detected (i.e., $PF = 1$) by $P^{i-1} \rightarrow P^i$ and its location recorded, the same pattern transition must be re-applied to detect other faults that cause 0s to propagate to the lower-priority input bits of P-ENC. When the same pattern transition is re-applied, the FIFO shifts out P-DEC-generated signals to bypass all OR outputs before, and including, the most-recently localized 0-carrying input of P-ENC. The FIFO has a depth of 2 to store the LOC values corresponding to the two most-recent pattern transitions.

In order to re-apply $P^{i-1} \rightarrow P^i$, the FSM block generates P^{i-1} after P_i; this transition sets the REV signal of P-ENC to 1. The backward transition $P^i \rightarrow P^{i-1}$ may detect additional fault(s) forcing $PF = 1$. However, we disregard the faults detected by a backward transition as it is already covered by one of the forward transitions in T. Therefore, if $PF = 1$ and $REV = 1$ after applying P^{i-1}, the FSM block next generates P^i (and not P^{i-2}) to re-apply the transition $P^{i-1} \rightarrow P^i$ for potential detection and localization of other faults. The bypass-select output of P-DEC is also disregarded if the fault is detected by a backward pattern transition. Hence, $REV = 1$ bypasses the P-DEC output and feeds an all-0 input to the ping register.

If $PF = 1$ and $REV = 0$, the FSM knows that the detected fault is the result of a forward pattern transition and therefore, generates a backward transition in order to re-apply the same forward transition that produced a "fail" earlier. If no fault is detected after applying a pattern transition (forward or backward), the FSM always generates a forward transition in the next test cycle.

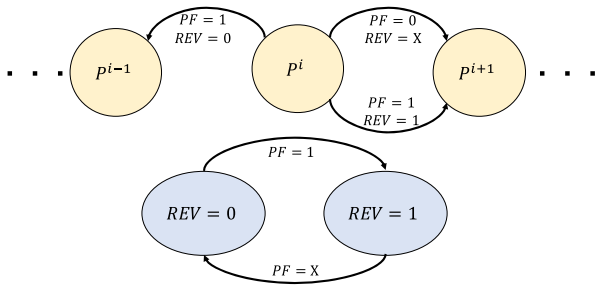


Fig. 13. State transition diagram for Mealy machine-based FSM-1 and FSM-2.

If $REV = 1$ in the present test cycle due to a backward pattern transition in the previous cycle, REV is reset to 0 in the next cycle as the present cycle is always a forward transition. On the contrary, if $REV = 0$ in the present test cycle, it is set to 1 in the next cycle if and only if $PF = 1$ at the end of the present cycle, which forces a backward transition in the next cycle. Figure 13 illustrates the different state transitions carried out by Mealy machine-based FSM-1 and FSM-2 blocks.

7.2.2 Detection and Localization of Single and Multiple Faults. There are four unique test patterns—$U1, U2, U3$, andU4—in T that are applied to the ILVs in a set:

- $U1$: all-0s; $U1$ corresponds to P^4, P^8, and P^{10} in T;
- $U2$: all-1s; $U2$ corresponds to P^1, P^7, P^{11}, and P^{13} in T;
- $U3$: alternating 1s and 0s with the leftmost ILV receiving 1; $U3$ corresponds to P^2, P^6, and P^9 in T;
- $U4$: alternating 0s and 1s with the leftmost ILV receiving 0; $U4$ corresponds to P^3, P^5, and P^{12} in T.

When $U1$ and $U2$ are applied by BIST Launch to the ILVs under test, the OR layer in BIST Capture gets its inputs from XNOR gates. If no fault is present in the ILVs, the XNOR outputs are all 1s which, in turn, produce all 1s at the output of the OR layer. Consequently, the PF output of P-ENC is 0. When $U3$ and $U4$ are applied by the BIST Launch, the OR layer gets its inputs from XOR gates. The multiplexing of XOR and XNOR outputs to the OR layer is controlled by the select signal produced by FSM-2. If no fault is present in the ILVs, the XOR outputs are all 1s which, in turn, produce all 1s at the output of the OR layer. As a result, $PF = 0$. Therefore, $PF = 0$ for a test pattern P^j indicates that no shorts, opens, stuck-at faults, or coupling-induced faults are detected by the transition $P^j \rightarrow P^{j+1}$ $(1 \leq j \leq 12)$.

Consider four ILVs in an 1D array: I_1, I_2, I_3, and I_4 (I_1 is the leftmost ILV). Let the corresponding XOR gates be X_1, X_2, X_3, X_4, and X_5. Here, X_1 is the leftmost XOR gate connected between the q_0^i output of FSM-2 and I_1, and X_5 is the rightmost XOR gate connected between I_4 and the q_1^i output of FSM-2. The OR gate connected between X_i and X_{i+1} $(i \in \{1, 2, 3, 4\})$ is denoted by O_i.

A short present between adjacent ILVs is detected by $U3$ or $U4$. Let I_2 be shorted to I_3. If $U3$ is applied, I_1 receives 1, I_2 receives 0, I_3 receives 1, and I_4 receives 0. Due to the short between I_2 and I_3, two possible scenarios arise: (A) both I_2 and I_3 propagate 0 if the logical drive of I_2 is stronger than that of I_3; (B) both I_2 and I_3 propagate 0 if the logical drive of I_2 is stronger than that of I_3. In scenario (A), the outputs of X_3 and X_4 are 0. Therefore, the output of O_3 is 0, which generates $PF = 1$, thereby indicating the presence of a fault. The location of the short in the 1D ILV array is indicated by LOC output of P-ENC; $LOC =$ "$01'$ considering I_1 as the MSB of the ILV set under

test. The *LOC* value is then read into local memory for post-processing. The scenario (B) is also detected by U_3 with the outputs of X_2 and X_3 being 0. Consequently, the output of O_2 is 0, which generates $PF = 1$ and $LOC = `10'$. Essentially, *LOC* points to the location of the highest-priority input of P-ENC, that is carrying a "0". Multiple shorts produce multiple 0s in P-ENC's input bus when a pattern transition is applied during a test cycle. The P-ENC localizes at most one 0-input per test cycle; the P-DEC then bypasses the localized input for potential detection and localization of other lower-priority 0's produced by the same pattern transition. Similarly, it can be shown that multiple stuck-at faults and delay faults are detected and localized by the P-ENC/P-DEC pair; the corresponding pattern transitions in T are given in Table 6.

Note that the output of O_i is generated from the signals propagated by I_{i-1}, I_i, and I_{i+1}, where $i \in \{2, 3\}$. The output of O_1 (O_4) is generated from the signals propagated by I_1 (I_3) and I_2 (I_4). Therefore, if the output of the leftmost (rightmost) OR gate is "0", a fault is detected and localized to the two leftmost (rightmost) ILVs via P-ENC. If the output of any other OR gate is "0", the detected fault is localized to a candidate set of three ILVs.

7.3 Design-Space Exploration for BIST Insertion

7.3.1 Assignment of ILVs to BIST-Capture Engines. After the 2D design is partitioned and placed-and-routed, the ILVs form the input/output pins of the M3D tiers. Consequently, the ILV locations (i.e., co-ordinates) are available from the DEF file, which is one of the outputs of the place-and-route process. The separation between two adjacent ILVs in any bus adheres to the minimum-allowed interconnect pitch designated by the design rules of the technology node. We divide ILVs, that propagate signal in the same direction, into two categories based on the information available after synthesis, tier-partitioning, and place-and-route:

- **Paired ILV:** The ILVs placed in close vicinity of each other are referred to as *paired* ILVs. Any two adjacent ILVs in the same bus are paired ILVs. If the separation between two adjacent ILV buses is less than or equal to the maximum separation between two ILVs in the same bus, the rightmost ILV of the left bus and the leftmost ILV of the right bus are paired ILVs.
- **Unpaired ILV:** If an ILV is not paired, it is considered to be an *unpaired* ILV.

Paired ILVs belonging to the same pair must be concurrently tested for bridging and coupling-induced faults by connecting the ILVs to adjacent inputs of the BIST Capture engine. Note that adjacent inputs of the BIST Capture engine feed an XOR gate that can detect a short or coupling-induced fault. We form sets of ILVs where each set contains at most n ILVs; here, n is the width of the BIST Capture engine. All ILVs in a set are tested concurrently by tying to the same Capture engine. A single BIST Capture engine can test multiple ILV sets, propagating signal in the same direction, across different test iterations by using the selector MUXes. If one of the ILVs (I_m) in a pair cannot be accommodated in the same set as that of the other ILV (I_n), a second set must contain both I_m and I_n with both being tied to adjacent inputs of BIST Capture. A set can contain both paired and unpaired ILVs. For example, let $n = 4$. Consider four ILV pairs to be present—$(I_1, I_2), (I_2, I_3), (I_3, I_4)$, and (I_4, I_5). Consider two unpaired ILVs—I_6 and I_7. The first set contains $\{I_1, I_2, I_3, I_4\}$ and the second set contains $\{I_4, I_5, I_6, I_7\}$. Therefore, the seven ILVs are tested in two test iterations.

After assigning ILVs to different sets and obtaining the total set count, DSE of the BIST architecture is carried out prior to BIST insertion in order to balance the trade-off between area and test-time overheads. The DSE is performed separately for up-going and down-going ILVs as they will be tested by separate BIST engines. If multiple BIST engines are inserted, more sets of ILVs can be tested concurrently leading to fewer test iterations. The width of the selector MUX in BIST Capture is also reduced leading to MUX area savings, but at the cost of increased gate count

(excluding that of selector MUXes) in the multiple Capture engines. We next describe the DSE methodology.

7.3.2 Overhead Minimization of BIST Insertion. To determine the optimum number m of n-bit BIST engines for insertion via DSE, we derive a closed-form expression to evaluate the BIST area overhead. Let the number of ILV sets (consisting of only up-going or only down-going ILVs) be s. Let the 2-input NAND-equivalent (NAND2-EQ) gate count of the BIST Capture engine (excluding the gate count of the selector MUXes) be A_n^C. Let the NAND2-EQ gate count of the BIST Launch engine be A_n^L. For m Capture engines and s sets, the required number of test iterations t is: $t = \lceil \frac{s}{m} \rceil$. Therefore, in a given test iteration, the selector MUXes select one out of a maximum of t input sets. Each selector MUX also has two additional inputs for selecting BIST-test and functional modes. Hence, the total number of inputs of a selector MUX is $t + 2$. Setting the input pin count of selector MUX to the nearest power of 2, the total NAND2-EQ gate count of the n selector MUXes is $n \times \frac{2^{\lceil \log_2(t+2) \rceil} - 1}{4}$. The area overhead Δ_{SB} (in terms of NAND2-EQ gate count) of the Shared-BIST architecture with m BIST engines shared by s sets of ILVs is:

$$\Delta_{SB} = m \times \left(A_n^L + A_n^C + n \times \frac{2^{\lceil \log_2(t+2) \rceil} - 1}{4} \right).$$

We define the normalized test cost TC of the Shared-BIST architecture as a weighted sum of test-time and area overheads that are normalized with respect to their corresponding maximums—t_{max} and $\Delta_{SB,max}$, respectively—across the different values of m explored. The test cost TC that needs to be minimized is: $TC = \alpha \cdot \frac{t}{t_{max}} + (1 - \alpha) \cdot \frac{\Delta_{SB}}{\Delta_{SB,max}}$, where α is the weight allotted to test-time minimization ($0 \leq \alpha \leq 1$).

Among the explored values of m, the value producing the minimum TC is selected as the optimum number of BIST engines to be inserted for testing the corresponding ILV sets. Note that we obtain two values of m—m_u and m_d—after carrying out DSE separately for up-going and down-going ILVs, respectively. Therefore, the total number of BIST engines inserted in the M3D design is $m_u + m_d$. The search space for DSE of an n-bit wide BIST architecture is bounded as: $1 \leq m_u \leq s_u$ and $1 \leq m_d \leq s_d$, where s_u (s_d) is the number of sets of up-going (down-going) ILVs. Note that all BIST engines are n-bit wide, irrespective of up-going or down-going ILVs.

The DSE is carried out for different values of n ($n \geq 2$). The value of n yielding the minimum combined test cost for up-going and down-going ILVs is selected. Here, n is set to be a power of 2, i.e., $n = 2^k$ (k is a positive integer), and $n \leq n_B$ where n_B is the maximum bus width in the M3D design. As a result, the worst-case time complexity of the search is $O(\log n_B \cdot (s_u + s_d))$.

7.4 Delay Bank Design

To test for a wider range of resistive or **small-delay defects** (**SDDs**) in ILVs, additional delay stages are added to the ILV inputs, as discussed in Section 5. These delay stages constitute the delay bank in BIST Launch. The test-path delay Δ_{tp} of an ILV equals the flop-to-flop propagation delay of the datapath through the ILV in the test mode; a test path starts at the flop outputs of FSM-1 in BIST Launch and ends at the output of the P-ENC in BIST Capture. The test-path delays of all ILVs are approximately equal. Let the delay of the critical functional path through an ILV I_i be $\Delta_{fp,i}$. If $\Delta_{fp,i} > \Delta_{tp}$, an additional delay of $(\Delta_{fp,i} - \Delta_{tp})$ needs to be added to the input of I_i for SDD testing; the addition of the delay stage to match test-path and functional-path (critical) delays is referred to as *delay equalization*. If $\Delta_{fp,i} \leq \Delta_{tp}$, delay equalization is not required for I_i.

As the ILV count in a block-level partitioned M3D design is typically in the order of thousands, such a high number of delay stages will introduce a prohibitively large area overhead. Therefore, we present a clustering-based approach to select M delay stages for delay equalization of N ILVs,

where $M << N$. The N ILVs are clustered into M clusters based on the $(\Delta_{fp,i} - \Delta_{tp})$ values of the ILVs $(1 \leq i \leq N)$; the information on functional-path slack is available after place-and-route. After clustering, the largest delay $(\Delta^*_{fp,i} - \Delta_{tp})$ in a cluster C_j $(1 \leq j \leq M)$ is conservatively added to the inputs of all ILVs in C_j for preventing test escape of SDDs. Here, $\Delta^*_{fp,i}$ is the functional-path delay of the *centroid* ILV of C_j. As a result, all ILVs in C_j, except the ILV with the highest $(\Delta_{fp,i} - \Delta_{tp})$, are being aggressively tested as their new test-path delays (after delay addition) exceed their functional-path delays. Aggressive testing of I_i can potentially lead to *yield loss*, YL_i, which is quantified as:

$$YL_i = \int_{s'_{tp,i}}^{s_{fp,i}} F(z)dz,$$

where $F(z)$ is the probability density function of defect size z, $s'_{tp,i} = T_{clk} - \Delta'_{tp,i}$, T_{clk} is the functional clock period, $\Delta'_{tp,i}$ is the new test-path delay, and $s_{fp,i} = T_{clk} - \Delta_{fp,i}$. During aggressive at-speed testing of I_i, the functional-path slack $s_{fp,i}$ is higher than the new test-path slack $s'_{tp,i}$. Therefore, an SDD whose (delay) size is smaller than $s_{fp,i}$ but larger than $s'_{tp,i}$ will be detected by the BIST and the chip is flagged as faulty. However, such an SDD is actually benign to the chip's functionality because the delay introduced by the SDD is smaller than $s_{fp,i}$. Consequently, the SDD-induced error cannot propagate to a flop in the design. This results in more chips getting unnecessarily marked as faulty leading to increased yield loss.

The objective of clustering is to select M clusters such that the yield loss YL associated with SDD testing of ILVs is minimized. Let the functional-path slack of the centroid ILV of cluster C_j be s^j_{fp}. Let the variable $\beta_{i,j}$ indicate whether ILV I_i is in cluster C_j. If I_i is in C_j, $\beta_{i,j} = 1$; otherwise $\beta_{i,j} = 0$. Let YL^j_i denote the yield loss associated with SDD testing of I_i in C_j: $YL^j_i = \int_{s^j_{fp}}^{s_{fp,i}} F(z)dz$. The objective of clustering can therefore be expressed as:

$$\begin{aligned} \min_{\beta_{i,j}} \quad & YL = \sum_{j=1}^{M} \sum_{i=1}^{N} \beta_{i,j} \cdot YL^j_i \\ \text{s.t.} \quad & \sum_{j=1}^{M} \beta_{i,j} = 1 \end{aligned} .$$

We have adopted the heuristic clustering algorithm for delay-stage selection from K-means clustering [23] with modifications to the "distance" function and the policy of updating centroids. In any given iteration of clustering, the distance L between an ILV I_m and a centroid ILV I_n $(1 \leq m, n \leq N)$ is calculated as: $L(s_{fp,m}, s_{fp,n}) = \int_{s_{min}}^{s_{max}} F(z)dz$, where $s_{min} = \min(s_{fp,m}, s_{fp,n})$ and $s_{max} = \max(s_{fp,m}, s_{fp,n})$. Following the distance-driven assignment of ILVs to M clusters, the centroid update takes place to select new centroid ILVs before the next iteration of clustering begins. For a cluster C_j, the ILV having the minimum functional-path slack (or maximum functional-path delay) among all ILVs in C_j is selected as the new centroid of C_j. The exit conditions of the algorithm are: (i) centroids remain the same after centroid-update, (ii) YL increases as a result of the current iteration, and (iii) maximum-allowed iteration count B is reached. Figure 14 describes the pseudo-code of the clustering algorithm, which is implemented in Python. The worst-case time complexity of the algorithm is $O(B \cdot N \cdot M)$.

8 EXPERIMENTAL RESULTS

8.1 Detection of Hard and Resistive Faults with Dual-BIST

We evaluated the detection of resistive faults through HSPICE simulations using the Nangate 45 nm open-cell library. Our test-bench consists of a representative set of 11 ILVs to demonstrate the

Input: B, N, M, S_{fp} // S_{fp}: list of critical (functional)-path slacks ($s_{fp,i}$) through ILVs (I_i)
Output: C, β // C is list of critical-path slacks of centroid ILVs, $\beta = \{\beta_{i,j} | 1 \leq i \leq N, 1 \leq j \leq M\}$
Sort S_{fp}, initialize $\beta_{i,j} \leftarrow 0$ for $1 \leq i \leq N, 1 \leq j \leq M$;
Initialize $C \leftarrow$ slacks picked randomly from sorted S_{fp};
Initialize $YL_{new} \leftarrow 0$, $iter \leftarrow 0$, $exit \leftarrow 0$; // algorithm stops when $exit$ is 1
while $exit == 0$ **do**
  Initialize $YL_{old} \leftarrow YL_{new}$, update $iter \leftarrow iter + 1$;
  for $s_{fp,i} \in S_{fp}$ **do**
    Initialize $L_i \leftarrow []$; // empty list
    for $s_{fp}^j \in C$ **do**
      Append $L(s_{fp,i}, s_{fp}^j)$ to L_i;
    end
    Assign s_{fp}^j to I_i for which $L(s_{fp,i}, s_{fp}^j)$ is minimum: $\beta_{i,j} \leftarrow 1$;
  end
  for $j \in [1, M]$ **do**
    Update $s_{fp}^j \leftarrow$ minimum critical-path slack of ILV with $\beta_{i,j} = 1$; // updating C
  end
  Calculate $YL_{new} \leftarrow \sum_{j=1}^{M} \sum_{i=1}^{N} \beta_{i,j} \cdot YL_i^j$;
  if C *unchanged after update* **or** $YL_{new} > YL_{old}$ **or** $iter == B$ **then** $exit \leftarrow 1$;
end

Fig. 14. Pseudo-code for the implementation of clustering-based delay selection.

functionality of the proposed BIST architecture via integration with the SPICE models of the ILVs. We considered three test clock frequencies—500 MHz, 1 GHz, and 2 GHz—and a power supply voltage of 1 V.

Four scenarios (I–IV) were chosen to validate the effectiveness of our enhanced Dual-BIST solution. Scenario I is the fault-free scenario; Scenario II contains two hard shorts (3 Ω) between adjacent ILVs of two distinct pairs; Scenario III contains two hard opens (1 MΩ) in two distinct ILVs; and Scenario IV contains a resistive short (5 KΩ) and a resistive open (20 KΩ) in two distinct ILVs. The signatures Y_1 and Y_2 are captured at the rising edge of the second clock cycle. The test clock frequency is 2 GHz. Figure 15 shows that an error is detected for each multiple-fault scenario.

Table 7(a) shows the minimum detectable resistive open ($R_{o,min}$) for different buffer sizes and clock frequencies (0.5 GHz, 1 GHz, and 2 GHz). The results confirm that for all three clock frequencies, the detection range of resistive opens increases with the decrease in transistor widths of the buffer—we are able to detect a resistive open as small as 25 KΩ with a clock frequency of 2 GHz. The impact of buffer size on the maximum detectable resistive short (R_S) is presented in Table 7(b). The range of detectable shorts is independent of the clock frequency, and it expands with a decrease in buffer size. Hence, we are able to detect a resistive short as large as 12 KΩ. The transistor lengths are fixed at 50 nm for both open- and short- detection experiments. The maximum lengths of appropriately-sized inverter chains in the input segment that provided detection at 0.5 GHz, 1 GHz, and 2 GHz were 21, 16, and 12, respectively.

8.2 Design and Synthesis Flow of BIST-inserted M3D IC

We evaluated the impact of BIST on the **power-performance-area** (**PPA**) metrics of four M3D benchmarks; see Table 8. Synthesis was performed using the Nangate 45 nm library. Figure 16 shows the integrated M3D design flow based on *Shrunk-2D* [38], which is used to synthesize and

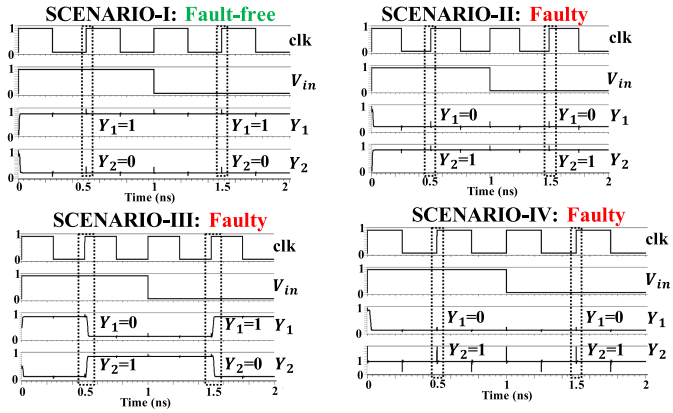


Fig. 15. Detection results for multiple-fault scenarios using Dual-BIST.

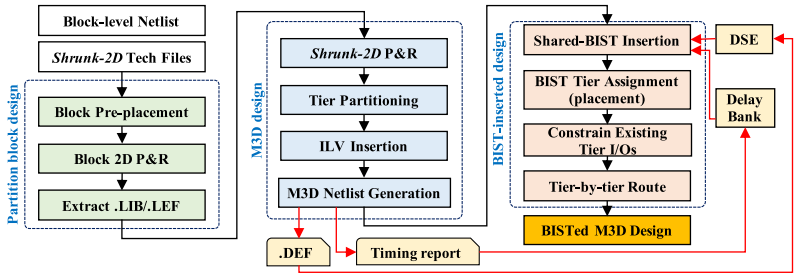


Fig. 16. Integration of Shared-BIST insertion with M3D design flow.

Table 7. Impact of Buffer (and Buffer Sizing) on (a) Resistive Open Detection and (b) Resistive Short Detection

(a)

Frequency (GHz)	Width (μm/μm) (PMOS/NMOS)	$R_{o,min}$ (KΩ)
0.5	Without buffer	118
	2/1	112
	1/0.5	110
1	Without buffer	56
	2/1	54
	1/0.5	52
2	Without buffer	29
	2/1	26
	1/0.5	25

(b)

Width (μm/μm) (PMOS/NMOS)	R_S (KΩ)
Without buffer	0.13
1/0.5	0.9
0.5/0.25	1.8
0.2/0.1	5
0.1/0.05	12

Table 8. Block-level Partitioned M3D Benchmarks used for SDQL Evaluation and PPA Comparison

Benchmark	Operational frequency (MHz)	Footprint (μm $\times$ μm)	Gate count	ILV count
AES-128 [1]	482.4	370 $\times$ 370	120,352	400
AVC-Nova [1]	144.1	500 $\times$ 500	148,410	316
Rocketcore (I) [1]	130.7	770 $\times$ 770	330,385	1,073
Rocketcore (II) [1]	110.7	770 $\times$ 770	300,227	1,062

Table 9. PPA Comparison of 2D and M3D N-BI Designs

Circuit	Metric	2D	M3D N-BI	Δ (%)
AES-128	Gate count	121,190	120,352	−0.69
	Cell area (μm^2)	168,878.61	167,662.73	−0.71
	Wirelength (m)	2.01	1.94	−3.54
	t_{pd}^{crit} (ns)	2.02	2.07	2.71
	Functional power (mW)	55.26	48.25	−12.69
AVC-Nova	Gate count	149,892	148,410	−0.98
	Cell area (μm^2)	275,495.14	273,982.13	−0.54
	Wirelength (m)	4.16	3.98	−4.44
	t_{pd}^{crit} (ns)	7.24	6.94	−4.34
	Functional power (mW)	25.11	24.30	−3.2
Rocketcore (I)	Gate count	330,890	330,385	−0.15
	Cell area (μm^2)	732,398.55	732,287.36	−0.01
	Wirelength (m)	7.64	7.63456699	−0.12
	t_{pd}^{crit} (ns)	6.55	7.65	14.36
	Functional power (mW)	80.27	78.14	−2.65
Rocketcore (II)	Gate count	302,665	300,227	−0.8
	Cell area (μm^2)	703,531.69	702,213.93	−0.18
	Wirelength (m)	7.29	7.29	0
	t_{pd}^{crit} (ns)	8.81	9.03	2.46
	Functional power (mW)	74.22	64.59	−12.97

implement the BIST-inserted benchmark designs. First, we start with a block-level netlist and *Shrunk-2D* technology files. Before the M3D design stage, we pre-place blocks/partitions in the corresponding tiers and perform **place-and-route** (**P&R**) in normal 2D fashion. Then, we extract **library** (**LIB**) and **library exchange format** (**LEF**) files of pre-designed blocks for M3D design stage.

We consider two major factors, i.e., the design hierarchy and area balancing while doing block partitioning. Considering the design hierarchy, we have grouped logic blocks having large number of connections between them and assigned them in the same partition; therefore, they can be placed close together. Moreover, in tier partitioning, area balancing is the one of key steps used to improve the performance of overall design. Therefore, we generated the area report of the synthesized netlist for area-balanced block and tier partitioning as well. It is clear that partitioning affects the performance of design and the number of ILVs. The two Rocketcore designs in Table 8 use different partitions, resulting in different numbers of ILVs and different operating frequencies.

In M3D design stage, we conduct *Shrunk-2D* P&R which cells are half-sized. From *Shrunk-2D* design, we blow-up cells to their original sizes and partition them into top and bottom tiers. When all cells are well-partitioned into corresponding tiers, we insert ILVs, which connect top and bottom tiers. Finally, we generate M3D netlist set, which contains the netlists of top-level M3D design, top tier design and bottom tier design for BIST insertion. Table 9 analyzes the PPA of 2D and M3D **non-BISTed** (**N-BI**) designs. The critical-path delay is denoted by t_{pd}^{crit}. The M3D designs achieve significant reduction in cell area, wirelength, and power consumption compared to their 2D counterparts.

In BIST-inserted design stage, we insert BIST circuitry for ILV testing. For both Dual-BIST and Shared-BIST insertion, the ILV sets are formed based on the proximity information that is obtained from the DEF file. The width of the Dual-BIST engine is set to 16 bits, i.e., a maximum of 16 ILVs can be tested concurrently by a single engine. Assuming the probability of fault occurrence to be 0.01 for ILVs and BIST logic, the fault-masking probability ($P[S_3 \cap M_3]$) for a 16-bit wide Dual-BIST engine is $\sim 2.7 \times 10^{-4}$. The number of Dual-BIST engines inserted equals the total number of up-going and down-going ILV sets.

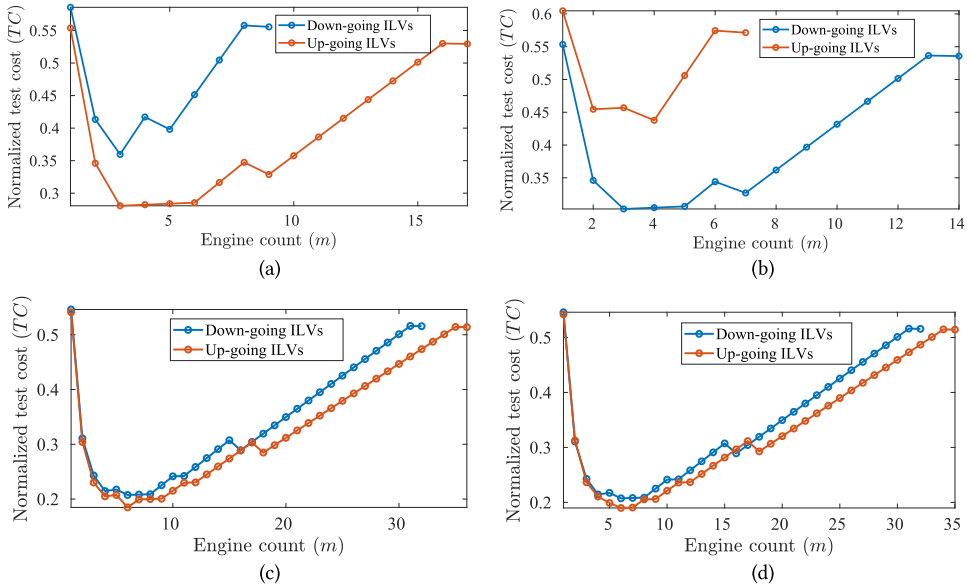


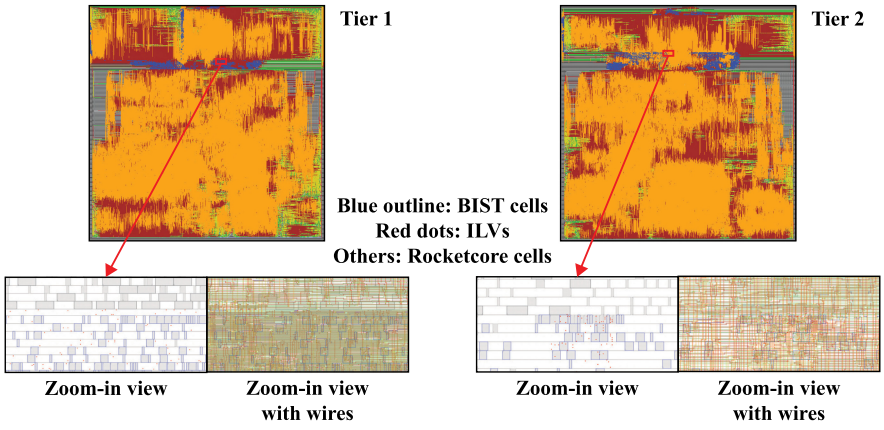
Fig. 17. DSE for area-time co-optimization of: (a) AES-128 ($s_d = 9, s_u = 17$), (b) AVC-Nova ($s_d = 14, s_u = 7$), (c) Rocketcore (I) ($s_d = 32, s_u = 36$), and (d) Rocketcore (II) ($s_d = 32, s_u = 35$).

Table 10. Maximum Operating Frequency for Shared-BIST

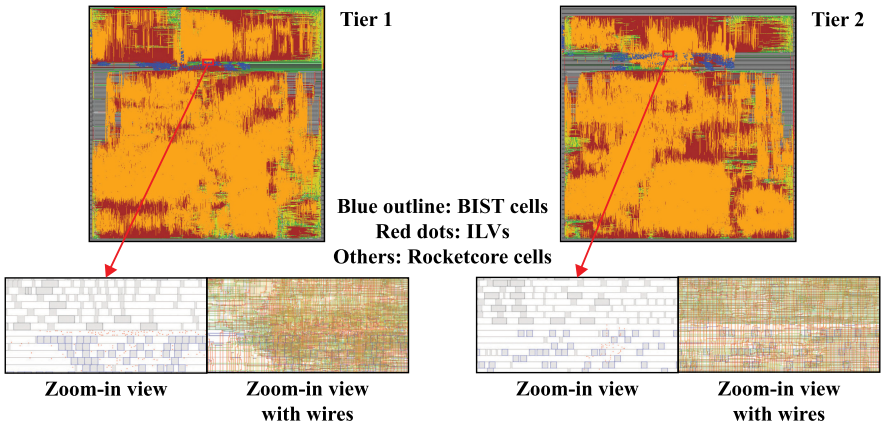
Engine width (bits)	Maximum test frequency (MHz)
8	787
16	502
32	263
64	137
128	77

To enable at-speed testing with Shared-BIST, DSE can be carried out to select the Shared-BIST configuration with a maximum operating frequency that exceeds the functional clock frequency of the M3D design. Table 10 presents the maximum Shared-BIST frequency as the width of the BIST-capture engine is increased, i.e., more ILVs tested concurrently by an engine. Given that the functional clock frequency is below 500 MHz for all four of our synthesized M3D benchmarks, we select the 16-bit wide Shared-BIST for insertion as its maximum operating frequency is 502 MHz.

For Shared-BIST insertion, the DSE for up-going and down-going ILV sets is carried out to select the optimal Shared-BIST configuration for insertion such that the normalized test cost TC is minimized. Figure 17 shows the DSE plots for selecting the optimum number of 16-bit wide Shared-BIST engines ($m_d + m_u$) to be inserted in AES-128, AVC-Nova, and Rocketcore, with the objective of co-optimizing area and test-time overheads. Note that the DSE search space for Rocketcore (I) and (II) are identical. The x-coordinate corresponding to the minima in a DSE curve indicates the engine count that minimizes TC. The values of m_d (m_u) for AES-128, AVC-Nova, Rocketcore (I), and Rocketcore (II) are 3 (3), 3 (4), 6 (6), and 6 (6), respectively. The delay bank is then designed using the critical-path slack information of ILVs, and integrated with the BIST Launch engine. The finalized BIST Capture engines for up-going (down-going) ILVs are inserted in the top (bottom)



(a) Dual-BISTed Rocketcore (I)



(b) Shared-BISTed Rocketcore (I)

Fig. 18. BISTed Rocketcore layouts.

tier and the BIST Launch engines for up-going (down-going) ILVs are inserted in the bottom (top) tier.

After the BIST insertion, we place the BIST logic in the corresponding tiers and perform tier-by-tier routing to finalize the BIST-inserted (BISTed) M3D design. The pin locations (i.e., ILVs) of the two tiers are constrained to remain unchanged during the post BIST-insertion P&R. Figure 18 illustrates the BIST-inserted layouts of Rocketcore (I).

8.3 Overhead for BIST

The 16-bit Dual-BIST and Shared-BIST architectures are inserted for testing 384, 288, and 1056 ILVs in AES-128, AVC-Nova, and Rocketcore (I and II), respectively. The BIST launch and capture engines are inserted in a non-intrusive manner and they do not affect the logical behavior of the original circuit. The MUX at the ILV input switches between the functional path and the connection to the BIST-launch engine. The ILV output is tapped and connected to the input of the BIST-capture engine. As a result, the BIST logic does not interfere with the original netlist logic.

We analyze the impact on PPA of N-BI M3D designs due to the insertion of: (i) a baseline DfT scheme, (ii) Dual-BIST, (iii) area-optimized Shared-BIST, and (iv) area-time co-optimized Shared-BIST. The baseline DfT scheme is based on the IEEE 1838 3D test standard where dedicated launch flops (control points) and capture flops (observation points) are added to those input or output ends of ILVs that are not connected to any functional flop [35]. If the input or output end of an ILV is already connected to a functional flop, the baseline DfT considers the functional flop to be a "shared flop" and does not add a dedicated flop for ILV test. Note that, unlike Shared-BIST, the baseline DfT scheme does not employ on-chip logic for fault localization. In baseline DfT, the test response from ILV testing must be scanned out and processed off-chip for debug and diagnosis.

The objective of area-optimized Shared-BIST insertion is to minimize the area overhead of the inserted Shared-BIST circuitry. We set $\alpha = 0$ in the expression for test cost TC (see Section 7.3) to obtain the optimum number of Shared-BIST engines to be inserted in each tier. For all four designs, area-optimized Shared-BIST scheme inserted one BIST engine per tier. The number of Shared-BIST engines inserted as a result of area-time co-optimization ($\alpha = 0.5$) is the same as the ones obtained via the DSE discussed in Section 8.2. Table 11 presents the overhead comparison between baseline DfT and the three BIST schemes. The overhead Δ for a PPA metric is calculated as the percentage change in the metric when a test architecture is inserted in the N-BI M3D design. The impact of test-logic insertion on circuit timing is measured as the change in the critical-path delay (t_{pd}^{crit}) of the original circuit. The test power refers to the power consumption in the test mode when only test logic is switching; the increase (Δ) in test power consumption over N-BI design is given in mW. Note that the test power is not indicated for baseline DfT and Dual-BIST schemes as the test logic cannot be turned off in the functional mode. Hence, the functional powers reported for baseline DfT and Dual-BIST schemes include the switching power consumption of both functional and test logic. On the contrary, the functional powers reported for the Shared-BIST schemes indicate the switching power consumption of the functional logic only as the Shared-BIST is turned off in the functional mode; the test powers indicate the switching power consumption of the BIST logic only in test mode. The PPA analysis includes the MUXes inserted as part of the BIST-launch and BIST-capture engines.

Owing to single engines inserted per M3D tier, the cell area and test power overheads of area-optimized Shared-BIST are generally lower than that of Dual-BIST and area-time co-optimized Shared-BIST. The functional power and wirelength overheads of both Shared-BIST schemes are comparable and are significantly lower than Dual-BIST. This is because the number of BIST engines inserted is higher for Dual-BIST, followed by area-time co-optimized Shared-BIST and area-optimized Shared-BIST. For this very reason, the timing impact (measured as the change in t_{pd}^{crit}) of Dual-BIST is worse compared to the Shared-BIST schemes. The timing impact is generally higher due to increased routing congestion and increased gate count. The small number of BIST engines shared by all the ILVs in area-optimized Shared-BIST leads to routing congestion, whereas the large number of Dual-BIST engines increases the gate count. The number of engines inserted via area-time Shared-BIST scheme is smaller than that of Dual-BIST but higher than that of area-optimized Shared-BIST. As a result, area-time co-optimized Shared-BIST leads to the least timing impact as it tends to reach the optimal point in the trade-off between gate count and routing congestion. The timing degradation due to area-time co-optimized Shared-BIST is ~10–13% for AES and Rocketcore (I) and as low as 0.4% for Rocketcore (II). The BIST power consumption is less than 3.5% of the functional power consumption. Therefore, the power consumption of BIST does not exceed the power budget of the M3D IC.

For the larger Rocketcore designs, the area, wirelength, and power consumption overheads of Shared-BIST are significantly lower than that of the baseline DfT scheme. The timing impact is

Table 11. Impact of Baseline DfT, Dual-BIST, and Shared-BIST (SBIST) on PPA of M3D N-BI Designs

Circuit	Metric	N-BI	Baseline Δ (%)	Dual-BIST Δ (%)	SBIST (area) Δ (%)	SBIST (area-time) Δ (%)
AES-128	Gate count	120,352	0.46	1.87	1.07	1.92
	Cell area (μm^2)	167,662.73	1.07	1.84	1.29	2.38
	Wirelength (m)	1.94	0.44	8.69	5.97	4.55
	t_{pd}^{crit} (ns)	2.07	4.94	116.05	34.10	13.29
	Functional power (mW)	48.25	48.14	15.02	8.45	9.02
	Test power (mW)	0	-	-	1.44 mW	1.78 mW
AVC-Nova	Gate count	148,410	0.71	1.14	0.75	1.79
	Cell area (μm^2)	273,982.13	1.24	0.84	0.68	1.64
	Wirelength (m)	3.98	0.68	1.17	1.8	1.83
	t_{pd}^{crit} (ns)	6.94	8.78	15.84	1.74	1.64
	Functional power (mW)	24.3	1.66	5.53	−25.15	−24.24
	Test power (mW)	0	-	-	0.26 mW	0.44 mW
Rocketcore (I)	Gate count	330,385	1.29	1.87	0.76	1.58
	Cell area (μm^2)	732,287.36	1.86	1.16	0.56	1.24
	Wirelength (m)	7.63	5	7.33	13.66	4.36
	t_{pd}^{crit} (ns)	7.65	12.75	55.47	124.8	10.76
	Functional power (mW)	78.14	1.83	1.89	1.41	1.28
	Test power (mW)	0	-	-	1.20 mW	0.91 mW
Rocketcore (II)	Gate count	300,227	1.41	2.06	0.83	1.74
	Cell area (μm^2)	702,213.93	1.92	1.21	0.58	1.29
	Wirelength (m)	7.29	5.28	8.09	6.43	4.48
	t_{pd}^{crit} (ns)	9.03	−0.29	42.01	20.03	0.42
	Functional power (mW)	64.59	1.83	1.83	0.63	1.24
	Test power (mW)	0	-	-	0.65 mW	0.78 mW

comparable for baseline and Shared-BIST schemes across the four designs as both schemes add fan-out loads to almost every ILV for reading the signals propagated by them for potential fault detection. Despite the added functionalities of on-chip fault localization and testing coupling-induced faults, the overall impact of Shared-BIST on the PPA is lower than that of baseline DfT and Dual-BIST schemes.

Currently, we fix or constrain all ILV locations so that the shift in ILV locations during post BIST-insertion P&R does not lead to new short possibilities and the inserted BIST remains fully effective. The timing overhead of Shared-BIST can be mitigated by not constraining selected ILVs whose shift in location does not generate new possibilities of shorts. Such a mitigation technique will involve multiple iterations of physical-design optimization and BIST insertions to help recover a significant portion of the affected timing. To further reduce the impact on timing, inductive fault analysis can be carried out in order to insert BIST only on those ILVs that are in functionally critical regions of the IC and are relatively more susceptible to defects and parametric variations.

9 CONCLUSION

We have presented a low-cost XOR-BIST architecture that requires only two test patterns to detect opens, stuck-at faults, and bridging faults (shorts) in ILVs. We have extended XOR-BIST to the Dual-BIST architecture to minimize ILV fault masking due to multiple BIST faults. We have investigated the impact of coupling between adjacent ILVs in block-level partitioned designs. Based on the analysis, we have presented the Shared-BIST architecture enabled with the functionality of detecting and localizing single and multiple faults, including coupling-induced faults. We have also presented a clustering-based methodology for designing a delay bank, as a component of the Shared-BIST Launch engine, for testing SDDs in ILVs. Evaluation of PPA for four two-tier M3D benchmarks show the effectiveness of the proposed BIST solutions.

A APPENDIX

We present below theorems and proofs that highlight the fault detection properties of the proposed BIST approach.

Theorem 1. *A set of ILVs contains no hard faults if and only if Y_1 is 1 in both clock cycles.*

Proof. If Y_1 is 1 in both the clock cycles, it implies that all the XOR outputs must be 1 in these two cycles. This implies that every XOR gate gets opposite values at its two inputs. Therefore, there is no hard short present between adjacent pair(s) of ILVs. If an ILV is stuck-at-d (s/d, $d \in \{0,1\}$), the XOR output will flip to 0 in the input cycle when the adjacent ILV is given d as input, which will force Y_1 to 0. If the ILV contains a hard open, the ILV output will be held at d for both cycles; this causes the XOR output to be 0 in the input cycle when the adjacent ILV is given d as input, which forces Y_1 to 0. Therefore, if Y_1 is 1 in both cycles, no hard faults are present in the ILVs. We next prove the "only if" part of the theorem. If Y_1 is 0 in either one or both cycles, we can infer that at least one of the XOR outputs is 0 in each cycle. If a hard short is present, it will force Y_1 to 0 in both cycles. If there is a stuck-at fault or hard open, Y_1 will be 0 in either of the two cycles, as explained earlier. If multiple concurrent faults are present, they will also be detected since their effects propagate through the XOR present between every adjacent ILV pair to Y_1 in a mutually exclusive manner; a hard short will force Y_1 to 0 in both cycles and a hard open or stuck-at fault will force Y_1 to 0 in one of the cycles, irrespective of the presence of other faults. □

Theorem 2. *The ILVs under test and the Dual-BIST engine are fault-free if and only if $Y_1 = 1$ and $Y_2 = 0$ in both cycles.*

Proof. We present the proof by explicitly enumerating all possible scenarios. Consider the case where $Y_1 = 0$ in either or both cycles. This observation implies at least one of the following three scenarios: (i) two or more ILVs are shorted; (ii) there is a stuck-at-0 fault in either BIST-A or BIST-B; and (iii) one or more ILVs are either stuck-at-0/stuck-at-1 or have an open fault. It is clear that all the above cases are indicative of faults in the ILVs and/or the Dual-BIST architecture. Next we consider the case where $Y_1 = 1$ in both the cycles. Let us take any two neighbouring ILVs, say V_1 and V_2. Suppose V_1 and V_2 are the two inputs to the XOR gate and its counterpart XNOR gate denoted by X_1 and XN_1, respectively. The possible events that can give rise to this scenario are as follows:

- *Event I:* V_1 and V_2 are shorted (or both are s/d; $d \in \{0, 1\}$); there is a node s/1 in the path from X_1's output to Y_1. This implies that BIST-A is faulty and it masks the short between V_1 and V_2.
- *Event II:* V_1 (V_2) is s/d. This implies that BIST-A is faulty since the other input of X_1, namely, V_2 (V_1) is s/$\bar{d}$.
- *Event III:* V_1 is s/d and V_2 is s/$\bar{d}$. Therefore, the two inputs of X_1 and XN_1 are s/d and s/$\bar{d}$, respectively.
- *Event IV:* BIST-A is fault free; V_1 and V_2 are fault free.

If *Event I* occurs, error detection is provided by Y_2. Since there is a fault in BIST-A, BIST-B is fault-free based on our single fault assumption for BIST. Hence, the short will produce a 1 at XN_1's output and $Y_2 = 1$ in both cycles. Thus, $Y_1.Y_2 = 1$ in both cycles. If *Event II* occurs, error detection is provided by Y_2 again. Similar to the previous case, BIST-B is fault-free due to the single fault assumption. The open/stuck-at will produce a 1 at XN_1's output and $Y_2 = 1$ in the two cycles since V_2 (V_1) will toggle with the applied input pattern despite V_1 (V_2) being unchanged. Thus, $Y_1.Y_2 = 1$ in the two cycles. For the analysis of *Event III*, consider V_0 and V_3 to be the ILVs adjacent to V_1 and V_2 respectively. If the ILVs are fault-free, they can propagate the effect of open/stuck-at faults. In

Event IV, BIST-A is fault-free, thus we can consider a single fault in BIST-B. *Event IV* may arise in two possible scenarios:

— Any node in BIST-B is s/1, hence $Y_2 = 1$ in both cycles. Therefore, $Y_1.Y_2 = 1$ in both cycles.
— BIST-B contains either no fault or a s/0 fault, and thus $Y_2 = 0$ in both cycles. No fault masking occurs since V_1 and V_2 are fault-free.

This proves that the ILVs and BIST are fault-free if and only if $Y_1 = 1$ and $Y_2 = 0$ in both cycles. Thus, we can detect ILV faults in the presence of a single BIST fault. □

Theorem 3. *The number of test patterns (both necessary and sufficient) required to detect all possible coupling-induced faults (based on the MAF model) in a 1D ILV array is 13.*

Proof. To obtain the minimum number of test patterns in the sequence, we count the minimum number of signal transitions to force on either or both left and right ILVs in an adjacent ILV-pair. The number of ways in which the signal on the left ILV undergoes transition (rise or fall), while the right ILV carries a fixed logic value (0 or 1) are $2 \times 2 = 4$. The number of ways in which the signal on the right ILV undergoes transition, while the left ILV carries a fixed logic value are $2 \times 2 = 4$. The number of ways in which signals on both left and right ILVs undergo transition (rise or fall) simultaneously is $2 \times 2 = 4$. Therefore, 12 transitions, i.e., 13 test patterns, are necessary and sufficient to cover all possible coupling-induced faults in the 1D ILV array. □

REFERENCES

[1] 1999. OpenCore Benchmark Suite. Retrieved from http://www.opencores.org/.
[2] 2009. Small-delay-defect Testing. Retrieved from https://www.edn.com/small-delay-defect-testing-2/.
[3] 2019. Monolithic 3D Roadmap. Retrieved from https://www.3dincites.com/2019/03/coolcube-more-than-a-true-3d-vlsi-alternative-to-scaling/.
[4] P. Batude, C. Fenouillet-Beranger, L. Pasini, V. Lu, F. Deprat, L. Brunet, B. Sklenard, F. Piegas-Luce, M. Cassé, B. Mathieu, O. Billoint, G. Cibrario, O. Turkyilmaz, H. Sarhan, S. Thuries, L. Hutin, S. Sollier, J. Widiez, L. Hortemel, C. Tabone, M.-P. Samson, B. Previtali, N. Rambal, F. Ponthenier, J. Mazurier, R. Beneyton, M. Bidaud, E. Josse, E. Petitprez, O. Rozeau, M. Rivoire, C. Euvard-Colnat, A. Seignard, F. Fournel, L. Benaissa, P. Coudrain, P. Leduc, J.-M. Hartmann, P. Besson, S. Kerdiles, C. Bout, F. Nemouchi, A. Royer, C. Agraffeil, G. Ghibaudo, T. Signamarcheix, M. Haond, F. Clermidy, O. Faynot, and M. Vinet. 2015. 3DVLSI with CoolCube process: An alternative path to scaling. In *Proceedings of the Symposium on VLSI Technology*.
[5] P. Batude, L. Brunet, C. Fenouillet-Beranger, F. Andrieu, J.-P. Colinge, D. Lattard, E. Vianello, S. Thuries, O. Billoint, P. Vivet, C. Santos, B. Mathieu, B. Sklenard, C.-M. V. Lu, J. Micout, F. Deprat, E. Avelar Mercado, F. Ponthenier, N. Rambal, M.-P. Samson, M. Cassé, S. Hentz, J. Arcamone, G. Sicard, L. Hutin, L. Pasini, A. Ayres, O. Rozeau, R. Berthelon, F. Nemouchi, P. Rodriguez, J.-B. Pin, D. Larmagnac, A. Duboust, V. Ripoche, S. Barraud, N. Allouti, S. Barnola, C. Vizioz, J.-M. Hartmann, S. Kerdiles, P. Acosta Alba, S. Beaurepaire, V. Beugin, F. Fournel, P. Besson, V. Loup, R. Gassilloud, F. Martin, X. Garros, F. Mazen, B. Previtali, C. Euvrard-Colnat, V. Balan, C. Comboroure, M. Zussy, Mazzocchi, O. Faynot, and M. Vinet. 2017. 3D sequential integration: Application-driven technological achievements and guidelines. In *Proceedings of the 2017 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 1–3.
[6] Nicola Campregher, Y. K. Peter, George A. Constantinides, and Milan Vasilko. 2005. Analysis of yield loss due to random photolithographic defects in the interconnect structure of FPGAs. In *Proceedings of the 2005 ACM/SIGDA 13th International Symposium on Field-Programmable Gate Arrays*.
[7] John Carulli. February 2021. GLOBALFOUNDRIES. Private correspondence. February 2021.
[8] A. Chaudhuri, S. Banerjee, H. Park, B. W. Ku, K. Chakrabarty, and S.-K. Lim. 2019. Built-in self-test for inter-layer vias in monolithic 3D ICs. In *Proceedings of the IEEE European Test Symposium*.
[9] Jonghyun Cho, Jongjoo Shim, Eakhwan Song, Jun So Pak, Junho Lee, Hyungdong Lee, Hyungdong Lee, and Joungho Kim. Active circuit to through silicon via (TSV) noise coupling. In *Proceedings of the 2009 IEEE 18th Conference on Electrical Performance of Electronic Packaging and Systems*.
[10] Jonghyun Cho, Kihyun Yoon, Jun So Pak, Joohee Kim, Woojin Lee, Taigon Song, Kiyeong Kim, Junho Lee, Hyungdong Lee, Kunwoo Park, Seungtaek Yang, Minsuk Suh, Kwangyoo Byun, and Joungho Kim. 2011. Modeling and analysis of through-silicon via (TSV) noise coupling and suppression using a guard ring. *IEEE Transactions on Components, Packaging and Manufacturing Technology* 1, 2 (2011), 220–233.

[11] M. Cuviello, S. Dey, Xiaoliang Bai, and Yi Zhao. 1999. Fault modeling and simulation for crosstalk in system-on-chip interconnects. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design. Digest of Technical Papers*.

[12] Rainer Dudek, Birgit Bramer, Roland Irsigler, and Bernd Michel. 2009. Thermo-mechanical reliability assessment for 3D through-Si stacking. In *Proceedings of the EuroSimE 2009-10th International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems*.

[13] Dominik Erb, Karsten Scheibler, Matthias Sauer, Sudhakar Reddy, and Bernd Becker. 2015. Multi-cycle circuit parameter independent ATPG for interconnect open defects. In *Proceedings of the 2015 IEEE 33rd VLSI Test Symposium (VTS)*.

[14] Yassine Fkih, Pascal Vivet, Bruno Rouzeyre, Marie-lise Flottes, and Giorgio Di Natale. 2013. A 3D IC BIST for pre-bond test of TSVs using ring oscillators. In *Proceedings of the 11th IEEE International New Circuits and Systems Conference*.

[15] Thomas Frank, Cédrick Chappaz, Patrick Leduc, Lucile Arnaud, Stéphane Moreau, Aurélie Thuaire, Rebha El Farhane, and Lorena Anghel. 2010. Reliability approach of high density through silicon via (TSV). In *Proceedings of the 12th Electronics Packaging Technology Conference*.

[16] Robert M. Geffken and Stephen E. Luce. 1999. Method of Forming a Self-aligned Copper Diffusion Barrier in Vias. US Patent 5,985,762.

[17] Wei Guo et al. 2012. Impact of through silicon via induced mechanical stress on fully depleted bulk FinFET technology. In *Proceedings of the 2012 International Electron Devices Meeting*.

[18] Richard F. Hafer, Oliver D. Patterson, Roland Hahn, and Hong Xiao. 2015. Full-wafer voltage contrast inspection for detection of BEOL defects. *IEEE Transactions on Semiconductor Manufacturing*, 28, 4 (2015), 461–468.

[19] Ron Ho, Kenneth W. Mai, and Mark A. Horowitz. 2001. The future of wires. *Proceedings of the IEEE* 89, 4 (2001), 490–504.

[20] Miccoli Ilio, Frederik Edler, Herbert Pfnür, and Christoph Tegenkamp. 2015. The 100th anniversary of the four-point probe technique: The role of probe geometries in isotropic and anisotropic systems. *Journal of Physics: Condensed Matter* 27, 22 (2015), 223201.

[21] A Jutman. 2004. Shift register based TPG for at-speed interconnect BIST. In *Proceedings of the 24th International Conference on Microelectronics*.

[22] Sukeshwar Kannan, Abhishek Koneru, and Krishnendu Chakrabarty 2020. Testing Monolithic Three Dimensional Integrated Circuits. US Patent 10,775,429.

[23] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. 2002. An efficient K-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 7 (2002), 881–892.

[24] Johann Knechtel, Igor L. Markov, and Jens Lienig. 2012. Assembling 2-D blocks into 3-D chips. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 31, 2 (2012), 228–241.

[25] Abhishek Koneru, Sukeshwar Kannan, and Krishnendu Chakrabarty. 2016. Impact of wafer-bonding defects on Monolithic 3D integrated circuits. In *Proceedings of the 2016 IEEE 25th Conference on Electrical Performance of Electronic Packaging and Systems*.

[26] Abhishek Koneru, Sukeshwar Kannan, and Krishnendu Chakrabarty. 2017. Impact of electrostatic coupling and wafer-bonding defects on delay testing of monolithic 3D integrated circuits. *ACM Journal on Emerging Technologies in Computing Systems* 13, 4 (2017), 1–23.

[27] Abhishek Koneru, Sukeshwar Kannan, and Krishnendu Chakrabart. 2018. A design-for-test solution based on dedicated test layers and test scheduling for monolithic 3D integrated circuits. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 38, 10 (2018), 1942–1955.

[28] Rajeev Kumar and Sunil P. Khatri 2013. Crosstalk avoidance codes for 3D VLSI. In *Proceedings of the 2013 Design, Automation & Test in Europe Conference & Exhibition*.

[29] Young-Joon Lee, Patrick Morrow, and Sung Kyu Lim. 2012. Ultra high density logic designs using transistor-level monolithic 3D integration. In Proceedings of the *2012 IEEE/ACM International Conference on Computer-Aided Design*. 539–546.

[30] Werner Liebsch et al. 2009. Process characterisation of DuPont MXA140 dry film for high resolution microbump application. In *EMPC*.

[31] Chang Liu, Taigon Song, Jonghyun Cho, Joohee Kim, Joungho Kim, and Sung Kyu Lim. 2011. Full-chip TSV-to-TSV coupling analysis and optimization in 3D IC. In *Proceedings of the 2011 48th ACM/EDAC/IEEE Design Automation Conference (DAC)*. IEEE, 783–788.

[32] Chang Liu and Sung Kyu Lim. 2012. A design tradeoff study with monolithic 3D integration. In *Proceedings of the 13th International Symposium on Quality Electronic Design*.

[33] Jian-Qiang Lu, Ken Rose, and Susan Vitkavage. 2007. 3D Integration: Why, what, who, when?. In *Proceedings of the Future Fab Int.*

[34] Luke Maresca. 2012. On the design partitioning of 3D monolithic circuits. (2012).

[35] Erik Jan Marinissen, Teresa McLaurin, and Hailong Jiao. 2016. IEEE Std P1838: DfT standard-under-development for 2.5 D-, 3D-, and 5.5 D-SICs. In *Proceedings of the 21th IEEE European Test Symposium*.
[36] Shreepad Panth, Kambiz Samadi, Yang Du, and Sung Kyu Lim. 2013. High-density integration of functional modules using monolithic 3D-IC technology. In *ASP-DAC*.
[37] Shreepad Panth, Kambiz Samadi, Yang Du, et al. 2014. Design and CAD methodologies for low power gate-level monolithic 3D ICs. In *Proceedings of the 18th Asia and South Pacific Design Automation Conference*. 171–176.
[38] Shreepad Panth, Kambiz Samadi, Yang Du, and Sung Kyu Lim. 2017. Shrunk-2D: A physical design methodology to build commercial-quality monolithic 3D ICs. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 36, 10 (2017), 1716–1724.
[39] Oliver D. Patterson, Horatio Wildman, Doron Gal, and Kevin Wu. 2006. Detection of resistive shorts and opens using voltage contrast inspection. In *Proceedings of the 17th Annual SEMI/IEEE ASMC 2006 Conference*.
[40] Rajesh Pendurkar, Abhijit Chatterjee, and Yervant Zorian. 2001. Switching activity generation with automated BIST synthesis for performance testing of interconnects. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 20, 9 (2001), 1143–1158.
[41] Janusz Rajski and Jerzy Tyszer. 2013. Fault diagnosis of TSV-based interconnects in 3-D stacked designs. In *ITC*.
[42] Yasuo Sato et al. 2005. Invisible delay quality-SDQM model lights up what could not be seen. In *Proceedings of the IEEE International Test Conference*.
[43] Max M. Shulaker, Tony F. Wu, Mohamed M. Sabry, Hai Wei, and H.-S. P. Wong. 2015. Monolithic 3D integration: A path from concept to reality. In *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition*.
[44] Taigon Song, Chang Liu, Yarui Peng, and Sung Kyu Lim. 2013. Full-chip multiple TSV-to-TSV coupling extraction and optimization in 3D ICs. In *Proceedings of the 50th Annual Design Automation Conference*.
[45] Sebastien Thuries, Olivier Billoint, Sylvain Choisnet, Romain Lemaire, Pascal Vivet, Perrine Batude, and Didier Lattard. 2020. M3D-ADTCO: Monolithic 3D architecture, design and technology co-optimization for high energy efficient 3D IC. In *Proceedings of the 2020 Design, Automation & Test in Europe Conference & Exhibition*.
[46] Menglin Tsai, Amy Klooz, Alexander Leonard, Jennie Appel, and Paul Franzon. 2009. Through silicon via (TSV) defect/pinhole self test circuit for 3D-IC. In *Proceedings of the 2009 IEEE International Conference on 3D System Integration*.
[47] Pascal Vivet. September 2020. CEA-Leti. Private correspondence. September 2020.
[48] Simon Wong, Abbas El-Gamal, Peter Griffin, Yoshio Nishi, Fabian Pease, and James Plummer. 2007. Monolithic 3D Integrated circuits. In *Proceedings of the International Symposium on VLSI Technology, Systems and Applications*.
[49] S. Simon Wong and Abbas El Gamal. 2009. The prospect of 3D-IC. In *Proceedings of the 2009 IEEE Custom Integrated Circuits Conference*.
[50] Boping Wu, Mark B. Ritter, Xiaoxiong gu, and Leung Tsang. 2011. Electromagnetic modeling of massively coupled through silicon vias for 3D interconnects. *Microwave and Optical Technology Letters* 53, 6 (2011), 1204–1206.
[51] Pooria M. Yaghini, Ashkan Eghbal, Siavash S. Yazdi, Nader Bagherzadeh, and Michael M. Green. 2015. Capacitive and inductive TSV-to-TSV resilient approaches for 3D ICs. *IEEE Transactions on Computers* 65, 3 (2015), 693–705.
[52] Wei Yao, Siming Pan, Brice Achkir, Jun Fan, and Lei He. 2013. Modeling and application of multi-port TSV networks in 3-D IC. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 32, 4 (2013), 487–496.
[53] Mahmut Yilmaz, Krishnendu Chakrabarty, and Mark M. Tehranipoor. 2010. Test-pattern selection for screening small-delay defects in very-deep submicrometer integrated circuits. *IIEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 29, 5 (2010), 760–773.
[54] Takao Yonehara. 2015. Epitaxial layer transfer technology and application. In *Proceedings of the 2015 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference*.

Received October 2020; revised March 2021; accepted May 2021