Fault Diagnosis for Resistive Random Access Memory and Monolithic Inter-Tier Vias in Monolithic 3-D Integration

Shao-Chun Hung[ID], Arjun Chaudhuri, Sanmitra Banerjee[ID], and Krishnendu Chakrabarty[ID], *Fellow, IEEE*

Abstract— Resistive random access memory (RRAM) constitutes a promising technology for next-generation memory architectures due to its simple structure, high ON/OFF ratio, and processing-in-memory ability. Its compatibility with emerging monolithic 3-D (M3D) integration enables extremely high density using monolithic inter-tier vias (MIVs). However, both RRAM and M3D are susceptible to high defect rates due to immature manufacturing processes and process variations. Research efforts have been devoted to RRAM testing, while existing test solutions predominantly focus on fault detection. Fault diagnosis for M3D-integrated RRAM and MIVs remains unexplored. In this work, we propose a diagnosis procedure to identify the fault origin when a chip fails the manufacturing test. We present a detailed characterization of RRAM faulty behaviors in the presence of concurrent process variations and manufacturing defects. Based on RRAM characteristics, we develop a diagnosis sequence by identifying appropriate reference resistance and applied voltages to efficiently distinguish fault origins. Experimental results show that the proposed solution is compatible with existing test algorithms to significantly improve diagnostic resolution. By appending the proposed sequence to test algorithms, over 90% diagnostic resolution is achieved for every type of fault considered in an M3D-integrated RRAM.

Index Terms— Fault diagnosis, resistive RAM, three-dimensional integrated circuits (ICs).

I. INTRODUCTION

RESEARCH on emerging memory technologies has accelerated the development of next-generation nonvolatile memory (NVM) architectures. Among promising NVM candidates, resistive random access memory (RRAM) offers the advantages of a simple two-terminal structure, low power consumption, high ON/OFF ratio, good scalability, and compatibility with CMOS processes [2], [3]. RRAM can be combined with monolithic 3-D (M3D) integration technology to achieve ultralow cost per bit and high memory density [4], [5].

M3D integration is an emerging technology that offers better power, performance, and area (PPA) benefits compared to traditional 2-D integrated circuits (ICs) [6]. M3D leverages fine-grained monolithic inter-tier vias (MIVs) to connect elements between device tiers [7]. Compared to the through-silicon vias (TSVs) in today's stacked 3-D ICs, MIVs are one to two orders of magnitude smaller, and the induced capacitance is negligible. These advantages allow the use of MIVs in large numbers in an M3D design, which greatly reduces the total wirelength and power consumption.

Memory-on-logic stacking is one of the major applications of M3D integration. Shulaker et al. [8] demonstrated the feasibility of stacking RRAM, carbon nanotube field-effect transistors (CNFETs), and silicon-FETs in a single design using low processing temperatures. Murali et al. [9] proposed an RRAM-based system to preserve RRAM cells at legacy nodes while scaling the peripheral logic with advanced technologies. The space freed up due to scaling allows additional analog-to-digital converters to be placed in the system, which significantly improves throughput and energy efficiency. In [10], heterogeneous RRAM arrays and peripheral circuits were separated into two tiers to improve area efficiency without affecting access latency.

The benefits of emerging technologies are invariably accompanied by new challenges. RRAM suffers from process variations during memristor fabrication [11]. For example, variability in the process steps can lead to deviation in the oxide thickness of an RRAM cell. These variations and defects induce faults, resulting in unexpected behaviors during normal operations. An unstable initialization process can also prevent RRAM cells from switching with the strength provided by the nominal write voltage.

In recent years, RRAM testing has received considerable attention. Chen et al. [12] proposed a fault modeling and testing solution for unique RRAM defects. However, physical fault origins were not fully distinguishable by the proposed method. A new algorithm was developed in [13] to detect faults in one-transistor one-memristor (1T1R) crossbar RRAM architectures, but it did not consider root-cause identification after a faulty cell was detected. In [14], design-for-testability (DfT) schemes for resistive opens were presented, but the proposed solutions can lead to the failure of good chips

(yield loss) in the presence of process variations. Programmable DfT schemes considering process variations were proposed in [15]; however, additional circuits were needed for each bitline (BL), leading to area overhead.

MIVs in M3D integration are also prone to defects, especially because they penetrate through the inter-tier dielectric. Voids produced in the dielectric due to surface roughness lead to voids in MIVs during etching, which impacts circuit operation and latency. Testing solutions for M3D designs and MIVs have been explored in recent years. Koneru et al. [16] introduced dedicated test layers to improve controllability and observability between device tiers; however, it is not always practical to insert a test layer between RRAM and CMOS tiers. Chaudhuri et al. [17] proposed a built-in self-test (BIST) structure for MIV testing and fault localization. However, the area overhead can be substantial when a large number of MIVs are placed between device tiers.

In [1], a fault diagnosis sequence was introduced to differentiate between RRAM process variations and MIV opens; however, this approach assumes the presence of only one source of variation in the defective RRAM. A new solution for M3D-integrated RRAM is therefore needed to identify the root causes of faulty cells and facilitate yield learning. An effective and practical diagnosis method should also distinguish MIV defects from RRAM process variations, without requiring additional DfT structures in the memory tiers.

In this article, we propose a novel fault diagnosis framework for M3D-integrated RRAM cells. A preliminary version of this article was presented in [1]. Compared with the preliminary version, our comprehensive method is able to generate unique signatures that can localize faults in MIVs and RRAMs in the presence of concurrent process variations, MIV opens, and MIV shorts. This capability helps in improving diagnostic resolution when these faults yield equivalent output responses during memory testing. The key contributions of this article are as follows.

1) We characterize faulty behaviors of an M3D-integrated RRAM in practical scenarios by considering concurrent process variations, MIV open defects, and MIV short defects.

2) We develop a diagnosis sequence with additional read/write operations with the appropriate reference resistance and applied voltages needed to identify the root causes of RRAM faulty behaviors.

3) We show that the proposed solution is compatible with existing test algorithms to improve diagnostic resolution without impacting fault coverage.

4) We show that no additional DfT structure around MIVs is needed for MIV diagnosis; therefore, our solution can be applied to large-size high-density M3D-integrated RRAM architectures.

The rest of this article is organized as follows. Section II provides an overview of an RRAM device and M3D integration. Section III presents a characterization of RRAM cells in the presence of process variations and MIV defects. The proposed diagnosis process is derived in Section IV. Section V shows a detailed comparison between our solution and existing test algorithms In Section VI, the generalized
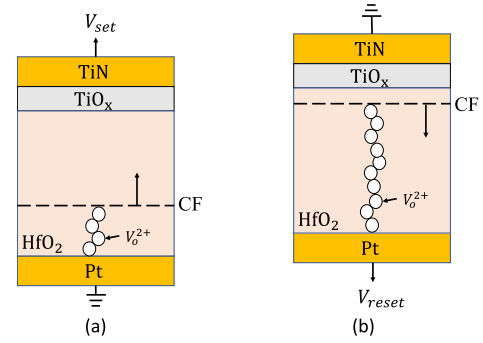

Fig. 1. Illustration of RRAM switching mechanisms. (a) SET operation and (b) RESET operation, where V_o^{2+} is the positive-charged oxygen vacancy, and V_{set} and V_{reset} refers to the applied voltage for the SET and RESET operations, respectively. The figure is adapted from [1] and [12] and redrawn.

solution for various design environments and the approaches to identify the sizes of variations and defects are provided. Finally, Section VII concludes this article.

II. BACKGROUND

A. RRAM Fundamentals

In an RRAM cell, a memristor is typically used as the storage element. The memristor is composed of a simple metal-insulator-metal (MIM) structure, which can be fabricated in the conventional back-end-of-line (BEOL). Such a memristor can be connected to a transistor to form a 1T1R RRAM device, where three terminals are connected to the BL, the source line (SL), and the word line (WL) in an RRAM array, respectively, and the transistor is a selector to control the memristor device. RRAM exhibits the memory property based on the fact that the resistance in some insulators changes with the applied electric field and retains its value when power is removed. This property has attracted attention in recent years for the development of next-generation NVM architectures.

With a voltage applied across the top and the bottom electrode, the RRAM cell switches from the high resistance state (HRS), often denoted as logic-0, to the low resistance state (LRS), often denoted as logic-1, and vice versa. The switching mechanism of an RRAM cell is based on the formation and rupture of the conduction filament (CF) between electrodes. After fabrication, an RRAM cell has an extreme HRS. A forming process, i.e., a dielectric soft breakdown is therefore necessary to initialize the RRAM cell to the LRS before normal operation. Under the high voltage in the forming process, the oxygen atoms knocked out of the lattice become oxygen ions that drift toward the anode and produce oxygen vacancies in the bulk oxide. The accumulation of the oxygen vacancies forms the CF that enables current to flow in the device; therefore, the RRAM cell switches to the LRS. This process can be visualized as writing logic-1 to the RRAM cell, denoted as the "SET" operation. To switch from the LRS to the HRS, a negative voltage is applied to force the oxygen ions to migrate back to bulk oxide and recombine with the oxygen vacancies, which ruptures the CF. This is denoted as the "RESET" operation. Fig. 1(a) and (b) illustrate the SET and RESET switching mechanism in an RRAM cell, respectively, with RRAM materials described in [18]. To read data from

an RRAM cell, a small voltage pulse, which is insufficient to change the current state, is applied to determine whether the cell is in the LRS or in the HRS. To ensure signal integrity, a safety margin is typically defined to prevent an unexpected logic value from being produced during read operations [19]. Let L be the length of an RRAM cell and $G(t)$ be the depth of CF at time t. Logic-1 is read out when $1 \geq (G(t)/L) \geq 0.6$, whereas logic-0 is read out when $0.4 \geq (G(t)/L) \geq 0$. The region corresponding to $0.6 \geq (G/L) \geq 0.4$ is defined as the undefined logic state. An RRAM cell with CF in the undefined region results in an output indistinguishable by the sense amplifier, which can be randomly interpreted as either logic-0 or logic-1 during read operations.

B. M3D Integration

M3D integration is a promising technology to continue performance improvement when Moore's Law hits physical limits. In an M3D design, all device tiers are fabricated in situ on the same wafer [20]. This has been made possible by significant breakthroughs in low-temperature manufacturing processes [21]. Fabricating upper-tier devices with low-temperature processes is necessary to prevent damage from interconnects and elements in the bottom tier. Recent work [22] has demonstrated that processing steps under the 550 °C thermal budget successfully produced reliable upper-tier devices without degrading the overall circuit performance. M3D leverages MIVs as interconnects between various tiers. The size of MIVs is of the same order as conventional BEOL vias, allowing M3D designs to achieve high alignment precision and extremely thin device tiers. Furthermore, unlike TSV-based 3-D ICs, MIVs do not induce tensile stress on surrounding elements; therefore, a massive amount of vertical interconnects can be used in M3D ICs. This advantage provides opportunities to address the communication bottleneck between logic and memory by developing heterogeneous memory-on-logic stacking architectures [9], [23].

RRAM has been used in M3D-integrated memory-on-logic designs due to its compatibility with the atomic-layer deposition (ALD) technique [4]. RRAM memory-on-logic design styles depend on the type of partitioning methodologies. Zokaee et al. [10] stacked a large unipolar RRAM array on top of a small bipolar RRAM array and peripheral circuits (e.g., row/column decoder, sense amplifier). Such a stacking method has been demonstrated to improve system performance and energy efficiency. Another partitioning style has been proposed in [5] by separating access transistors into two tiers, while all memristors are integrated into the top-tier BEOL. Compared to the planar counterparts, M3D stacking of devices achieves a 1.5× higher bit per surface ratio. Both design styles have been made possible by utilizing a large amount of MIVs to connect devices in different tiers.

However, the impact of MIV defects becomes severe in M3D-integrated RRAM designs. In contrast to conventional BEOL vias, MIVs need to penetrate through the inter-tier dielectric and a silicon layer. Therefore, MIVs are prone to defects due to voids in the fabrication process (see Section I). In M3D stacking, MIVs are used to connect the bottom electrode of a memristor (BE) of the top-tier memristors

to the drain of the bottom-tier access transistors [5]. MIV defects tend to produce an open in series with the memristor, causing individual cell misbehavior. Additionally, MIVs are fabricated close to the drain/gate/source contacts of top-tier transistors, with a distance below the maximum metal pitch in the processing technology node [5]. Large voids and pinholes can make MIVs shorted to the top-tier transistors [24], causing failure output responses during read and write operations. There is a need for a new low-cost framework based on read/write instructions to diagnose both the memory and the MIV faults. Such a framework is key to improving yield learning for emerging M3D-integrated RRAM devices and shortening the time-to-market.

C. RRAM Fault Models and Testing

Because RRAM array architectures are similar to static RRAM and dynamic RAM arrays, conventional fault models for testing RAM are applicable to RRAM testing; these fault models include stuck-at fault (SAF), transition fault (TF), state coupling fault (CFst), and address decoder fault (AF). RRAM-unique fault models have also been explored in recent years, including deep fault [25], undefined state fault (USF) [14], and overforming fault [12]. Novel test algorithms have been explored based on different RRAM fault models and RRAM-based neuromorphic architectures [12], [13], [26]. March testing schemes are popular in these algorithms due to high fault coverage and optimized test time. However, conventional March sequences are not sufficient for detecting RRAM unique faults related to the undefined region (e.g., UR and USF) because such faults produce random values in the defective cells to be read during testing.

To address this issue, Kannan et al. [25] leveraged sneak-path current to distinguish faulty cells from fault-free cells, but sneak paths lead to additional power consumption. Haron and Hamdioui [14] proposed a DfT scheme for USF detection, but the proposed scheme may cause yield loss with process variations. An enhanced March testing algorithm with multiple reference values was provided in [27] as follows:

$$\left\{ \begin{array}{l} \Uparrow (r_{\text{ref1}}, w0, w0); \Uparrow (r0, r_{\text{ref0}}, w1, w1) \\ \Downarrow (r_{\text{ref1}}, w0, r_{\text{ref0}}, w0); \Downarrow (r_{\text{ref0}}, w1, r_{\text{ref1}}, w1) \end{array} \right\} \quad (1)$$

where $w0$ ($w1$) refers to the write logic-0 (logic-1) operations, and $r0$ is the read operation with the expected logic value equal to 0. r_{ref0} and r_{ref1} are read operations with reference resistance corresponding to the lower bound and the upper bound of the undefined region, respectively. The symbol $\Uparrow$ denotes the increasing address order, while $\Downarrow$ denotes the decreasing address order. By including r_{ref1} and r_{ref0} operations in the March sequence, random values produced by faulty RRAM cells in the undefined region can be distinguished from fault-free logic-0 and logic-1; therefore, UR and USF are detected. However, fault origin identification after fault detection was not discussed in [27]. As discussed in [1], fault diagnosis is of the same importance as fault detection for emerging RRAM technology to facilitate yield learning. This motivates us to design a diagnosis framework to distinguish MIV defects from RRAM process variations. Our framework can find appropriate reference resistance values to identify
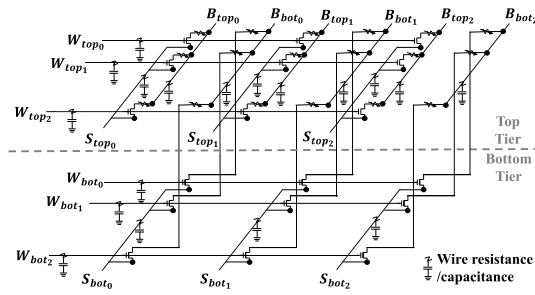
Fig. 2. Schematic of an M3D-integrated RRAM array with 3×3 cells in both tiers. The figure is adapted from [1] and redrawn.

Fig. 3. Margin of error in confidence interval analysis.

faulty behaviors caused by different RRAM process variations and MIV defects. Additional write and read operations with appropriate reference values and applied voltages are appended to the enhanced March algorithm to improve the diagnostic resolution without any loss of fault coverage. Because no DfT structure is needed for MIVs, the proposed solution is applicable to large-size high-density M3D RRAM architectures.

III. CHARACTERIZATION OF M3D-INTEGRATED RRAM WITH RRAM PROCESS VARIATIONS AND MIV DEFECTS

In this section, we present the characterization of an M3D-integrated RRAM array in the presence of RRAM process variations and MIV defects. We construct an 1T1R array with 3×3 RRAM cells in each tier using the partitioning method proposed in [5], in which access transistors are partitioned into two tiers. Access transistors are simulated using the Nangate 45-nm open cell library. The schematic of the M3D-integrated array is shown in Fig. 2.

To understand the impacts of process variations on the resistance states and switching operations of RRAM cells, we have carried out Monte Carlo simulations. Each parameter is considered to follow a Gaussian distribution centered around its nominal value, with the standard deviation equal to 15% of the nominal value. All parameters are varied simultaneously in each trial during Monte Carlo simulations, where the number of trials N has been determined using confidence interval analysis. The margin of error for a confidence interval is calculated as

$$\text{Margin of error} = \frac{Z \times \sigma}{\sqrt{N}} \quad (2)$$

where Z is the critical Z-value associated with a specific confidence interval, and σ denotes the standard deviation of the calculated variables [28]. For a 95% confidence interval, the critical Z-value Z is set to 1.96 and the maximum acceptable error is 8% [29]. Fig. 3 shows the margin of error in the confidence interval analysis for both LRS and HRS with various numbers of trials. With an increase in N, the margin of error of LRS and HRS consistently approach the values well below the acceptable threshold. As the margin of error with 10 000 trials is much lower than the maximum acceptable error, all the Monte Carlo trials reported in this article were performed with $N = 10\,000$.

To analyze the impacts of MIV defects and RRAM process variations, we conduct consecutive SET and RESET operations and read out the resistance state of the RRAM cell
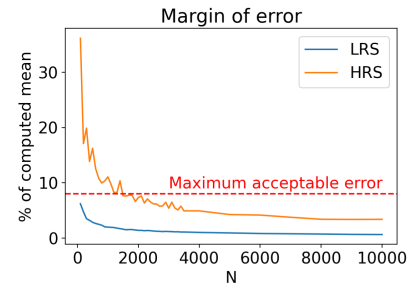
to evaluate whether the RRAM cell switches appropriately. We assume that the RRAM cells are successfully initialized during the forming process (i.e., the initial resistance state is LRS). We first perform $w0r0$ operations to switch the resistance state of each cell from LRS to HRS. The resistance value read out during $r0$ is denoted as R_H. Next, we switch back all the cells to LRS with $w1r1$ operations and obtain the corresponding resistance value, denoted as R_L. According to [13], these operations are sufficient to sensitize all SAFs and delay faults. Note that among the state-of-the-art RRAM technologies, the Stanford model [18], the voltage threshold adaptive memristor (VTEAM) model [30], and the Gonzalez-Cordero model [31] have been demonstrated to achieve high accuracy for experimental memristor devices [32]. However, the validation of the Gonzalez-Cordero model [31] is limited to experimental devices with a switch time on the microsecond scale. Given that advanced memory devices (e.g., M3D-integrated RRAM) exhibit switching behavior in nanoseconds, we carry out SPICE simulations with the Stanford model [18] and the VTEAM model [30] in our experiments.

A. Simulation Results With the Stanford Model [18]

The parameters used in our simulation with the Stanford model [18] are shown in Table I. Note that variations in cell length and the cross-sectional area have been demonstrated to impact the upper and the lower resistance limits of a memristor, determined by $G_{\max}$ and $G_{\min}$, respectively [25]. Variation in t_{ox} influences the electric field across the memristor [18], which affects the write strength during SET and RESET operations. Therefore, we vary $G_{\max}$, $G_{\min}$, and t_{ox} in our Monte Carlo simulations and characterize the cell behavior. The process variation scenarios that lead to faulty RRAM cell behavior are described as follows.

1) *Joint Variations in G_{min} and t_{ox}:* Fig. 4(a) illustrates two faulty behaviors caused by joint variations in $G_{\min}$ and t_{ox}. Variations in $G_{\min}$ influence the resistance value when the RRAM cell is in the LRS, impacting the voltage across the cell during write operations. Variations in t_{ox} affect the electric field across the memristor. If the electric field between top electrode of a memristor (TE) and BE is not sufficient to displace the CF in the insulator, the RRAM cell remains in its initial state (i.e., LRS), causing a stuck-at-1 fault. When the electric field is large enough to move the CF but not adequate to switch the RRAM's resistance state to

TABLE I
PARAMETERS USED IN SPICE SIMULATION WITH
THE STANFORD MODEL [18]

| Parameter	Value	Description
t_{ox}	12 nm	Oxide thickness
G_{max}	2 nm	Maximum gap distance between CF and TE
G_{min}	0.7 nm	Minimum gap distance between CF and TE
F_{min}	1.4×10^9 V/m	Minimum electric field to switch state
R_{th}	2.1×10^3 K/W	Thermal resistance
E_a	0.6 eV	Average activation energy for oxygen vacancy generation
$a0$	0.25 nm	Atom spacing
V_{SET}	3 V	Voltage of SET operation
V_{RESET}	-3 V	Voltage of RESET operation
V_{Read}	0.8 V	Voltage of read operation
T_{write}	13 ns	SET/RESET operation time period
T_{read}	5 ns	Read operation time period
R_{wire}	12.78 Ω	Wire resistance [33]
C_{wire}	0.046 fF	Wire capacitance [33]
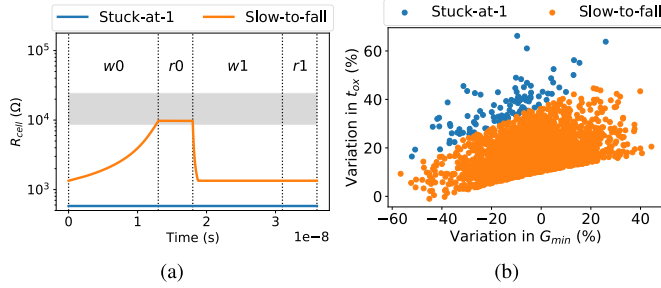
Fig. 4. Process variation scenarios leading to faulty behaviors due to joint variations in G_{min} and t_{ox}. (a) SPICE simulation results. (b) Correlation between G_{min} and t_{ox}. Note that R_{cell} is the equivalent resistance of the 1T1R cell, and the gray region denotes the undefined state.

logic-0, a slow-to-fall fault occurs. Fig. 4(b) presents the correlation between G_{min} and t_{ox} from samples leading to stuck-at-1 faults and slow-to-fall faults. Note that such samples also experience variations in G_{max}; however, the faulty behaviors illustrated in Fig. 4(a) are primarily driven by the combined impacts of G_{min} and t_{ox}, regardless of the value of G_{max}. With a constant value of G_{min}, a significant increase in t_{ox} leads to a substantial decrease in the electric field, therefore causing stuck-at-1 faults. Conversely, even a small increase in t_{ox} results in slow-to-fall faults.

2) *Decrease in G_{max}:* A reduction in G_{max} leads to a decrease in R_H of the RRAM cell. When R_H is reduced to the extent that it falls within the undefined state, the RRAM cell consistently experiences a USF following each $w0$ operation. If the reduction in R_H is even more substantial, causing it to fall below the undefined state and enter the logic-1 state, the RRAM cell always retains a logic-1 value. Fig. 5(a) and (b) show the simulation results and the distribution of variations in G_{max} that lead to stuck-at-1 faults and USFs. For samples with variations in G_{min} and t_{ox} that do not cause the RRAM stuck at its initial state or lead to a slow-to-fall fault, R_{cell} can reach its maximum state within the $w0$ period to complete the transition. However, when the reduction in G_{max} is within the range of 26.0% and 38.5%, R_H
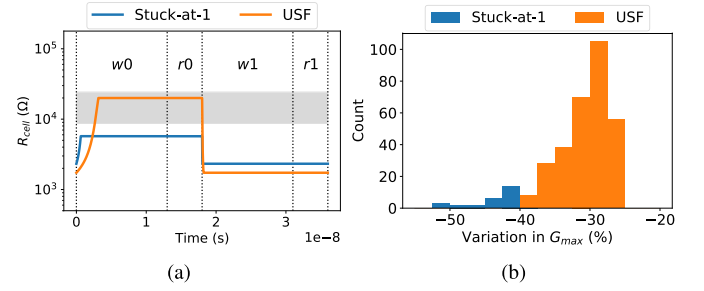

Fig. 5. Process variation scenarios leading to faulty behaviors due to decrease in G_{max}. (a) SPICE simulation results. (b) Distribution of G_{max}.
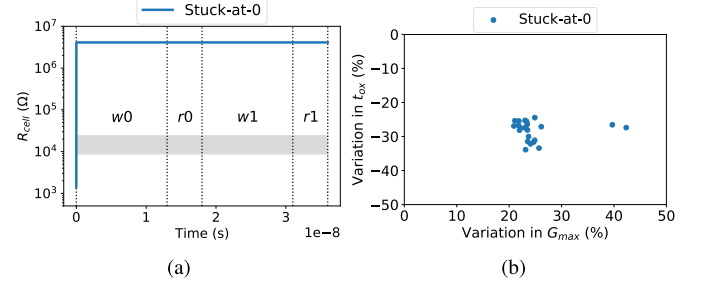

Fig. 6. Process variation scenarios leading to stuck-at-0 faults due to the increase in G_{max} and the decrease in t_{ox}. (a) SPICE simulation results. (b) Correlation between G_{max} and t_{ox}.

resides within the undefined state and causes a USF. A decrease in G_{max} that is larger than 38.5% leads to a stuck-at-1 fault.

3) *Increase in G_{max} and Decrease in t_{ox}:* An increase in G_{max} combined with a decrease in t_{ox} causes the RRAM cell to transition to HRS with an exceptionally large R_H. This results in the scenario that the RRAM cell is unable to switch back to LRS using the write strength supplied by the internal write circuit. Fig. 6(a) and (b) demonstrate the simulation results and the correlation of G_{max} and t_{ox} that lead to stuck-at-0 faults. Following the completion of the $w0$ operation, R_{cell} reaches an exceedingly high resistance value. The electric field generated by the write circuit fails to move the CF back toward BE, forcing the RRAM to be stuck in the logic-0 state after the $w1$ operation.

B. Simulation Results With the VTEAM Model [30]

The parameters used in our simulation with the VTEAM model [30] are shown in Table II. Note that process variations in a memristor are manifested as shifts in the upper and lower resistance limits. The VTEAM model [30] allows for the manipulation of these two resistance values (i.e., R_{ON} and R_{OFF}). This capability permits us to directly simulate the consequences of altering these two states and evaluate the RRAM behavior. Therefore, we vary these two parameters in our Monte Carlo simulations. The observed faulty behaviors are listed in Table III and described as follows.

1) *Decrease in R_{ON}:* In contrast to the Stanford model [18], the VTEAM model [30] relies on a threshold voltage instead of a minimum electric field to determine whether the resistance state can be switched. The voltage across

Parameter	Value	Description
R_{on}	2 kΩ	Memristor's minimum resistance
R_{off}	10 kΩ	Memristor's maximum resistance
D	3 m	Physical width of the memristor
V_{SET}	1.4 V	Voltage of SET operation
V_{RESET}	−1.4 V	Voltage of RESET operation
V_{Read}	0.1 V	Voltage of read operation
T_{write}	40 ns	SET/RESET operation time period
T_{read}	5 ns	Read operation time period

TABLE III
FAULTY BEHAVIOR DUE TO PROCESS VARIATIONS
WITH THE VTEAM MODEL

Variation	Size of variations	Fault model
Decrease in Ron	$\Delta R_{on} \geq 49.7\%$	Stuck-at-1
Decrease in $Roff$	$48.2\% > \Delta R_{off} \geq 32.1\%$	USF
	$\Delta R_{off} \geq 48.2\%$	Stuck-at-1
Decrease in R_{on} and R_{off}	$49.7\% > \Delta R_{on} \geq 40.1\%$ & $32.1\% > \Delta R_{off} \geq 23.3\%$	Slow-to-fall

the electrodes of a memristor relies on the process of charge sharing between the memristor and the equivalent resistance within the conduction path. This voltage reduces as the value of R_{ON} decreases. If the reduction in R_{ON} is greater than or equal to 49.7% of its nominal value, the voltage across TE and BE becomes insufficient to switch the state of the RRAM and leads to a stuck-at-1 fault, regardless of the variation in R_{OFF}.

2) *Decrease in R_{OFF}:* A reduction in R_{OFF} can lead to faulty behavior when it falls below the upper limit of the undefined state. In such cases, the RRAM cell cannot be identified as logic-0, even if it reaches its highest resistance value. Specifically, when the decrease in R_{OFF} lies within the range of 32.1% and 48.2% of its nominal value, a USF occurs. A Stuck-at-1 fault is induced when the reduction in R_{OFF} exceeds 48.2%. As the RRAM has successfully reached its maximum resistance value (i.e., $R_{cell} = R_{OFF}$), the variation in R_{ON} has no impact on these faulty behaviors.

3) *Joint Variations in R_{ON} and R_{OFF}:* Variations in the values of R_{ON} and R_{OFF} have a direct impact on the transition time required to switch the states of the RRAM device. A decrease in R_{ON} amplifies the difference in resistance required to complete a transition from logic-1 to logic-0. Simultaneously, it diminishes the voltage across the memristors, decelerating the movement of the CF in the insulator. Moreover, when the write strength fails to propel the CF all the way to the opposite side of the insulator, the equivalent resistance value can lie in the undefined state with a reduction in R_{OFF}. These combined effects make the RRAM cell unable to switch from logic-1 to logic-0 within a write period, leading to a slow-to-fall fault.

C. MIV Open Defects

Voids in MIVs due to manufacturing defects can lead to opens inside 1T1R RRAM cells [24]. Such open defects are
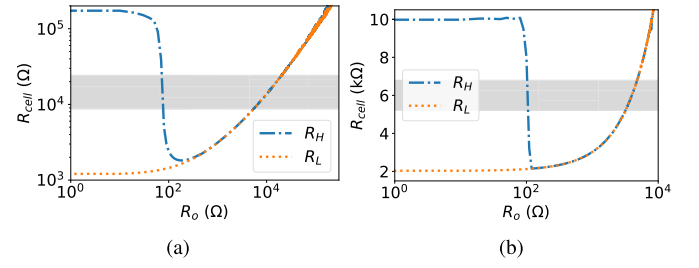

Fig. 7. Characterization of the RRAM cell in the presence of MIV open defects. (a) Stanford model [18]. (b) VTEAM model [30].

functionally equivalent to an additional resistance R_o between the drain of bottom-tier access transistors and the BE of top-tier memristors. To assess the impact of R_o on RRAM switching, we conduct SPICE simulations using models with parameters at their nominal values. We systematically increase the size of R_o in increments of 10 Ω while monitoring the resistance values during $r0$ and $r1$ operations (i.e., R_H and R_L). Fig. 7(a) and (b) shows the characterization of an RRAM cell with different sizes of R_o with the Stanford model [18] and the VTEAM model [30], respectively. Note that if R_H lies in the region above the undefined state, the cell is fault-free after the $w0$ operation. Similarly, the $w1$ operation is successful if R_L is below the undefined state. When the size of R_o is small, both $w0$ and $w1$ are fault-free. As the size of R_o increases, R_H becomes equivalent to R_L, leading to an SAF. Such a fault is caused by the voltage droop in the RRAM cell. Because the memristor is in series with R_o, the supply current flows through R_o during write operations, resulting in an additional IR-drop. If the voltage across the memristor is substantially reduced, the write strength across the device is not sufficient to move the CF in the insulator; therefore, the RRAM cell is stuck at the initial state (i.e., LRS).

Note however that the state of R_{cell} of the defective cell changes from logic-1 to logic-0 when R_o is large. This can be explained by the fact that R_o dominates the equivalent resistance inside the defective 1T1R RRAM cell. The equivalent resistance of an 1T1R cell consists of ON-resistance of the access transistor, MIV open R_o, and resistance in the memristor. The ON-resistance of the access transistor is designed to be negligible in order to minimize its impact on normal operations, while the memristor is stuck at LRS, which is relatively small compared to R_o. Therefore, if the size of R_o is larger than the undefined state, the state of the defective RRAM cell will be identified as logic-0, leading to a stuck-at-0 fault to be detected during testing.

D. MIV Short Defects

In an M3D-integrated RRAM array, MIVs serve as connections between the bottom-tier transistors and the BE of memristors located in the top-tier BEOL. Compared to the TSVs in modern 3-D technologies, MIVs are considerably smaller, and the associated capacitance is negligible. This advantage enables the fabrication of a large number of MIVs in M3D-integrated RRAM arrays. However, it also reduces the distance between MIVs and the top-tier devices within the inter-tier dielectric. Esmanhotto et al. [5] have demonstrated

TABLE IV

Faulty Behavior Caused by MIV Short Defects With the Stanford Model [18]

Defect	Scenario	Faulty behavior	Size of defects
MIV shorted to drain	(drain, $RRAM_{top}$, RESET, LRS, LRS)	$RRAM_{bot}$ switches from LRS to HRS	$R_s \leq 30\ \Omega$
	(drain, $RRAM_{top}$, RESET, HRS, LRS)	$RRAM_{bot}$ switches from LRS to HRS	$R_s \leq 70\ \Omega$
	(drain, $RRAM_{bot}$, RESET, LRS, LRS)	$RRAM_{top}$ switches from LRS to HRS	$R_s \leq 30\ \Omega$
	(drain, $RRAM_{bot}$, RESET, LRS, HRS)	$RRAM_{top}$ switches from LRS to HRS	$R_s \leq 70\ \Omega$
MIV shorted to gate	(gate, $RRAM_{top}$, RESET, LRS, LRS)	$RRAM_{bot}$ switches from LRS to HRS	$R_s \leq 340\ \Omega$
	(gate, $RRAM_{top}$, RESET, HRS, LRS)	$RRAM_{bot}$ switches from LRS to HRS	$R_s \leq 340\ \Omega$
	(gate, $RRAM_{top}$, SET, LRS, LRS)	$RRAM_{bot}$ switches from LRS to HRS	$R_s \leq 340\ \Omega$
	(gate, $RRAM_{top}$, SET, HRS, LRS)	$RRAM_{bot}$ switches from LRS to HRS	$R_s \leq 340\ \Omega$
	(gate, $RRAM_{bot}$, RESET, LRS, LRS)	$RRAM_{bot}$ slow to fall	$R_s \leq 250\ \Omega$
	(gate, $RRAM_{bot}$, RESET, HRS, LRS)	$RRAM_{bot}$ slow to fall	$R_s \leq 250\ \Omega$
MIV shorted to source	(source, $RRAM_{top}$, RESET, LRS, LRS)	$RRAM_{bot}$ switches from LRS to HRS	$R_s \leq 250\ \Omega$
	(source, $RRAM_{top}$, RESET, HRS, LRS)	$RRAM_{bot}$ switches from LRS to HRS	$R_s \leq 250\ \Omega$
	(source, $RRAM_{bot}$, RESET, LRS, LRS)	$RRAM_{bot}$ slow to fall	$R_s \leq 220\ \Omega$
	(source, $RRAM_{bot}$, RESET, HRS, LRS)	$RRAM_{bot}$ slow to fall	$R_s \leq 220\ \Omega$

TABLE V

Faulty Behavior Caused by MIV Short Defects With the VTEAM Model [30]

Defect	Scenario	Faulty behavior	Size of defects
MIV shorted to drain	(drain, $RRAM_{top}$, RESET, LRS, LRS)	$RRAM_{top}$ slow to fall	$R_s \leq 3350\ \Omega$
	(drain, $RRAM_{top}$, RESET, HRS, LRS)	$RRAM_{bot}$ switches from LRS to HRS	$R_s \leq 70\ \Omega$
	(drain, $RRAM_{bot}$, RESET, LRS, LRS)	$RRAM_{bot}$ slow to fall	$R_s \leq 3350\ \Omega$
	(drain, $RRAM_{bot}$, RESET, LRS, HRS)	$RRAM_{top}$ switches from LRS to HRS	$R_s \leq 60\ \Omega$
MIV shorted to gate	(gate, $RRAM_{top}$, RESET, LRS, LRS)	$RRAM_{bot}$ switches from LRS to HRS	$R_s \leq 1480\ \Omega$
	(gate, $RRAM_{top}$, RESET, HRS, LRS)	$RRAM_{bot}$ switches from LRS to HRS	$R_s \leq 1480\ \Omega$
	(gate, $RRAM_{top}$, SET, LRS, LRS)	$RRAM_{bot}$ switches from LRS to HRS	$R_s \leq 1480\ \Omega$
	(gate, $RRAM_{top}$, SET, HRS, LRS)	$RRAM_{bot}$ switches from LRS to HRS	$R_s \leq 1480\ \Omega$
	(gate, $RRAM_{bot}$, RESET, LRS, LRS)	$RRAM_{bot}$ slow to fall	$R_s \leq 5350\ \Omega$
	(gate, $RRAM_{bot}$, RESET, HRS, LRS)	$RRAM_{bot}$ slow to fall	$R_s \leq 5390\ \Omega$
MIV shorted to source	(source, $RRAM_{top}$, RESET, LRS, LRS)	$RRAM_{bot}$ switches from LRS to HRS	$R_s \leq 1470\ \Omega$
	(source, $RRAM_{top}$, RESET, HRS, LRS)	$RRAM_{bot}$ switches from LRS to HRS	$R_s \leq 1480\ \Omega$
	(source, $RRAM_{bot}$, RESET, LRS, LRS)	$RRAM_{bot}$ slow to fall	$R_s \leq 5370\ \Omega$
	(source, $RRAM_{bot}$, RESET, HRS, LRS)	$RRAM_{bot}$ slow to fall	$R_s \leq 5410\ \Omega$

that the distance between an MIV and the drain contact of the nearest top-tier device is approximately 200 nm. This distance is of a similar magnitude to the metal pitch in the 28-nm technology node. MIV short defects are created when the size of voids and pinholes in the dielectric is sufficiently large to connect MIVs to top-tier transistors, leading to RRAM cell misbehavior.

A schematic of an MIV shorted to drain/gate/source contacts is demonstrated in Fig. 8, with the MIV short represented as R_s. To comprehensively understand the effects of MIV shorts on RRAM switching, we carry out SPICE simulations with 48 distinct scenarios. Each scenario can be denoted as $(c, \text{target}, \text{op}, \text{RS}_{top}, \text{RS}_{bot})$, where $c \in \{\text{drain, gate, source}\}$ is the contact shorted to the MIV, target $\in \{\text{RRAM}_{top}, \text{RRAM}_{bot}\}$ is the target RRAM cell during write operations, op $\in \{\text{RESET, SET}\}$ is the operation applied to the target RRAM cell, $\text{RS}_{top} \in \{\text{LRS, HRS}\}$ is the initial resistance state of RRAM_{top}, and $\text{RS}_{top} \in \{\text{LRS, HRS}\}$ is the initial resistance state of RRAM_{bot}. For each of these scenarios, we vary the size of R_s and conduct SPICE simulations to record the resistance state of both RRAM_{top} and RRAM_{bot} following the designated operation (i.e., op).

Tables IV and V summarize the observed faulty behavior caused by MIV short defects with the Stanford model [18] and the VTEAM model [30], respectively. The detailed explanations of such faulty behavior are shown as follows.
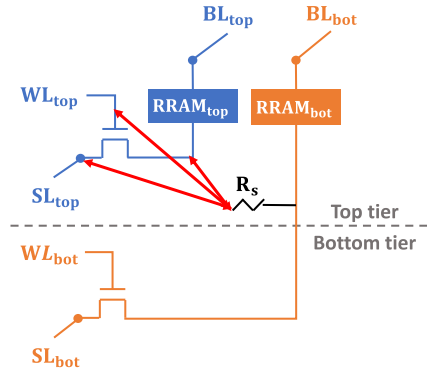
Fig. 8. Schematic of RRAM cells with MIV shorted to drain/gate/source contacts of the top-tier transistor.

1) *MIV Shorted to Drain:* An MIV short connecting the MIV to the drain contact of the top-tier transistor allows current to flow through R_s into the nontarget RRAM, creating a voltage difference between TE and BE. If this voltage is sufficiently large, it triggers an unintended transition in the nontarget RRAM cell. Moreover, the process of charge sharing between the target RRAM and the nontarget RRAM has an impact on the voltage received at the MIV. A large reduction in voltage can slow down the write operations of the target RRAM, resulting in a TF.

2) *MIV Shorted to Gate:* When the top-tier RRAM is the target during write operations (i.e., target $=$ RRAM$_{top}$), a positive voltage is applied to its gate contact to activate the access transistor. An MIV short between the gate contact and the MIV causes a current flow through R_s into the RRAM$_{bot}$, leading to a negative voltage across the memristor. RRAM$_{bot}$ switches to HRS if the voltage is sufficiently large to switch the state. However, if the bottom-tier RRAM is the target and a RESET operation is performed, R_s is connected in parallel with the RRAM$_{bot}$, thereby reducing the equivalent resistance across the device. This reduction decreases the voltage across RRAM$_{bot}$ due to charge sharing along the conduction path, which can lead to a slow-to-fall fault.

3) *MIV Shorted to Source:* MIV shorted to source leads to a current flowing from the source contact to the RRAM$_{bot}$ during the RESET operation of the top-tier RRAM. An unintended switch of the RRAM$_{bot}$ happens when R_s is low. Furthermore, when the bottom-tier RRAM is undergoing a RESET operation, charge sharing occurs among the resistances along the conduction path, the memristor, and R_s. A slow-to-fall fault is generated when the charge received at RRAM$_{bot}$ is insufficient to switch its state within the write period.

Note that the ranges of defect sizes leading to faulty behavior are significantly different for the two RRAM technologies. This discrepancy is primarily attributed to the distinct resistance values associated with the LRS and HRS in each model. Additionally, in the VTEAM model [30], the resistance state can change once the voltage across the memristor exceeds a certain threshold value. In contrast, the Stanford model [18] relies on the electric field across the RRAM cell, influenced by the location of the CF and the oxide thickness. Therefore, it is expected that the difficulty of triggering an unintended switch is relatively high for the Stanford model [18]. A severe MIV short defect (i.e., a small value of R_s) is required to induce faulty behaviors.

IV. PROPOSED DIAGNOSIS ALGORITHM

According to the characterization of RRAM behavior in the presence of RRAM process variations and MIV defects, different fault origins can lead to the same fault condition and produce equivalent output signatures during testing. Table VI shows the fault dictionary for both memristor models utilizing the testing algorithm described in (1), where r^i denotes the ith read operation, and "V" and "X" indicate that the corresponding read operation can and cannot detect the fault, respectively. Note that without loss of generality, we assign the order of ⇑ to be from the bottom tier to the top tier. Clearly, fault origins leading to the same erroneous behavior are detected by identical read operations, thus yielding equivalent output responses. Therefore, solely relying on a testing algorithm is insufficient to pinpoint the root causes of these faults.

In this section, we introduce the proposed diagnosis algorithm to help identify fault origins after a fault is detected. Note that for the stuck-at-0 fault and USF, different

TABLE VI
FAULT DICTIONARY WITH THE TESTING ALGORITHM IN (1)

Fault origin	Faulty behavior/model	Fault dictionary $r^1 r^2 r^3 r^4 r^5 r^6 r^7$
Stanford model [18]		
Joint variations in G_{min} and t_{ox}	Stuck-at-1	XVVXVVX
	Slow-to-fall	XXXXVXX
Decrease in G_{max}	Stuck-at-1	XVVXVVX
	USF	XXVXVVX
Increase in G_{max} and decrease in t_{ox}	Stuck-at-0	XXXVXXV
MIV open	Stuck-at-1	XVVXVVX
	USF	VXVVVVV
	Stuck-at-0	VXXVXXV
MIV shorted to drain	Unintended switch	VXXVXXX
MIV shorted to gate	Unintended switch	XXXVXXX
	Slow-to-fall	XXXXVXX
MIV shorted to source	Unintended switch	XXXVXXX
	Slow-to-fall	XXXXVXX
VTEAM model [30]		
Decrease in R_{on}	Stuck-at-1	XVVXVVX
Decrease in R_{off}	Stuck-at-1	XVVXVVX
	USF	XXVXVVX
Decrease in R_{on} and R_{off}	Slow-to-fall	XXXXVXX
MIV open	Stuck-at-1	XVVXVVX
	USF	VXVVVVV
	Stuck-at-0	VXXVXXV
MIV shorted to drain	Unintended switch	VXXVXXX
	Slow-to-fall	XXXXVXX
MIV shorted to gate	Unintended switch	XXXVXXX
	Slow-to-fall	XXXXVXX
MIV shorted to source	Unintended switch	XXXVXXX
	Slow-to-fall	XXXXVXX

fault origins produce distinct testing signatures, making them distinguishable during the testing process. Hence, we do not consider the fault-origin identification for the stuck-at-0 fault and USF in our proposed solution.

A. Fault-Origin Identification for Stuck-at-1 Faults

Stuck-at-1 faults can be induced by MIV open defects and RRAM process variations. However, the cell behaviors in the presence of these fault origins exhibit distinct characteristics. For cells with MIV opens, their R_L and R_H become equivalent and increase as the size of open increases, as shown in Fig. 7. If a stuck-at-1 fault is caused by joint variations in $G_{\min}$ and t_{ox} for the Stanford model [18] or a significant decrease in R_{ON} for the VTEAM model [30], the defective RRAM cell cannot be switched by the internal write circuit and remains consistently stuck in its LRS. However, when a stuck-at-1 fault is caused by the reduction in the maximum resistance value of the RRAM (i.e., decrease in $G_{\max}$ or R_{OFF}), R_H is different from R_L, even though the cell is in the logic-1 state after $w0$ operations. Leveraging these characteristics, we design a March sequence to identify fault origins for a stuck-at-1 fault as follows:

$$\{\Updownarrow (w1, r_{MIV}, w0, r_{MIV})\} \tag{3}$$

where r_{MIV} is the read operation with the reference resistance equal to the minimum resistance value of defective cells with MIV opens. When logic values $(0, 0)$ are observed during two
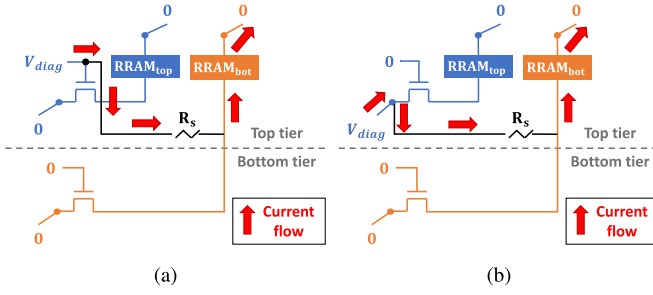
Fig. 9. Schematic of fault-origin identification for MIV short defects. (a) MIV shorted to gate. (b) MIV shorted to source. Note that V_{diag} denotes the voltage for diagnosis.

r_{MIV} operations, the cell under diagnosis has an MIV open. Logic values $(1, 1)$ correspond to joint variations in G_{min} and t_{ox} or a decrease in R_{ON}, while logic values $(1, 0)$ are generated by defective cells experiencing a decrease in G_{max} or R_{OFF}.

B. Fault-Origin Identification for MIV Short Defects

An MIV short defect can lead to an unintentional state change of a nontarget RRAM (i.e., its state changes when the access transistor is not switched on). Leveraging this characteristic, the location of the short can be identified by applying the appropriate voltage to BL/WL/SL and observing the state changes of RRAM cells. Fig. 9(a) illustrates a schematic for distinguishing defects caused by an MIV short connected to the gate contact from other defects. This is achieved by applying V_{diag} specifically to the gate contact of the top-tier transistor while turning off all other wires. If R_s is not present, both RRAM_{top} and RRAM_{bot} maintain their resistance states because there is no voltage difference between their TE and BE. However, if the short defect connects the MIV to the gate contact, current flows through R_s into the bottom-tier RRAM, causing RRAM_{bot} to switch from LRS to HRS. Similarly, when V_{diag} is applied to the top-tier SL while grounding all other wires, a switch in RRAM_{bot} only occurs if there is an MIV short connected to the source contact, as shown in Fig. 9(b).

Based on these properties, we derive a March sequence to identify fault origins for an MIV short defect as follows:

$$\left\{ \Downarrow w1;\ \Uparrow D_g;\ \Uparrow r0;\ \Uparrow D_s;\ \Uparrow r0 \right\} \tag{4}$$

where D_g and D_s are the diagnosis process for the MIV defects shorted to the gate contact and shorted to the source contact, respectively. Initially, all RRAM cells in the array are set to LRS by the first $w1$ operation. D_g and D_s are carried out to identify the location of the MIV short after an unintentional switch is detected during testing. If the output of the first $r0$ is 0, the MIV short is connected to the gate contact; the MIV short is connected to the source contact if the output of the second $r0$ is 0. The MIV short is connected to the drain contact when both $r0$ operations output logic-1. The value of V_{diag} can be determined through SPICE simulations. In this work, we set V_{diag} as 4 V to ensure that all MIV short defects listed in Tables IV and V can be appropriately identified.

C. Fault-Origin Identification for Slow-to-Fall Faults

A slow-to-fall fault can be caused by both RRAM process variations and MIV short defects. Note that the defects due to MIV shorts to the gate and source contacts contact can be identified using the same sequence as described in (4). Therefore, it suffices to distinguish process variations from MIV shorts connected to the drain contact.

As slow-to-fall faults arising from MIV shorts to the drain contact are primarily due to charge sharing (as detailed in Section III-D), they can be rectified by introducing additional charges during write operations. The March sequence to identify the origins of a slow-to-fall fault is formulated as follows:

$$\left\{ \Updownarrow w1;\ \Updownarrow \left(w0^*, r_{\text{ref0}}\right) \right\} \tag{5}$$

where $w0^*$ is a distinct RESET operation employed during diagnosis. During $w0^*$, in addition to applying V_{RESET} to the SL of the target RRAM, the same voltage is also applied to the BL of the nontarget RRAM. If the slow-to-fall fault is a consequence of process variations, the voltage applied to the BL has no impact and the slow-to-fall fault still occurs. Therefore, the output of r_{ref0} is equal to logic-1. However, if the fault is induced by MIV short defects, the voltage applied to the BL introduces additional charges that help to switch the state of the target RRAM. Hence, the output of r_{ref0} becomes logic-0.

D. Proposed March Algorithm for Diagnosis

By combining (3)–(5), the proposed March algorithm for diagnosis is shown as follows:

$$\left\{ \begin{array}{l} \Updownarrow (w1, r_{\text{MIV}});\ \Updownarrow (w0^*, r_{\text{MIV}}, r_{\text{ref0}}); \\ \Downarrow w1;\ \Uparrow D_g;\ \Uparrow r0;\ \Uparrow D_s;\ \Uparrow r0 \end{array} \right\}. \tag{6}$$

The repeated $w1$ and $w0$ operations are removed to reduce the diagnosis runtime. Table VII provides the responses with the proposed March diagnosis algorithm in the presence MIV opens and RRAM process variations. Clearly, every fault origin leading to the same faulty behavior has its distinct output responses. Although there are overlaps in logic values between fault models in the diagnosis process, these fault models have different signatures during testing. Combined with the testing signatures, the proposed March diagnosis sequence guarantees to distinguish MIV defects from RRAM process variations. Moreover, this diagnosis sequence is compatible with any March test algorithm to improve diagnostic resolution without any adverse impact on fault coverage. As the diagnosis process is conducted after the manufacturing test in order to facilitate yield learning, there is no test time overhead for the fault-free devices.

V. EXPERIMENTAL RESULTS

The proposed solution is compatible with existing algorithms to help in identifying fault origins after faults are detected. To assess its efficacy, we append the sequence in (6) to existing test algorithms [13], [27] and conduct experiments on simulated samples. We carry out SPICE simulations with 68 000 samples, including 10 000 samples from Monte Carlo

TABLE VII

OUTPUT RESPONSES WITH THE PROPOSED MARCH DIAGNOSIS ALGORITHM

Fault model/ behavior	Logic values					Fault origin
	r_{MIV}	r_{MIV}	r_{ref0}	r_0	r_0	
Stuck-at-1	0	0	1	1	1	MIV open
	1	0	1	1	1	a decrease in G_{max} with [18] or a decrease in R_{off} with [30]
	1	1	1	1	1	joint variations in G_{min} and t_{ox} with [18] or a decrease in R_{on} with [30]
Unintended switch	1	0	0	1	1	MIV shorted to drain
	1	0	1	0	1	MIV shorted to gate
	1	0	1	1	0	MIV shorted to source
Slow-to-fall	1	0	1	1	1	joint variations in G_{min} and t_{ox} with [18] or joint variations in R_{on} and R_{off} with [30]
	1	0	0	1	1	MIV shorted to drain
	1	0	1	0	1	MIV shorted to gate
	1	0	1	1	0	MIV shorted to source

simulations, 10 000 samples featuring MIV open defects, and 48 000 samples involving MIV short defects, with 10 000 samples for each MIV short scenario. We employ two metrics to evaluate the effectiveness of the proposed diagnosis algorithm, namely diagnostic resolution and diagnosability. *Diagnostic resolution* quantifies the percentage of samples for which the fault origin can be successfully distinguished among all samples having the same faulty behavior; *diagnosability* measures the percentage of samples that can be accurately diagnosed among all samples with the same fault origin.

Table VIII presents the results of diagnostic resolution. Without the proposed diagnosis sequence, fault origins for the stuck-at-1 fault, slow-to-fall fault, and unintended switch cannot be identified, as multiple fault origins can yield the same response during testing. However, with the introduction of the proposed diagnosis sequence in (6), additional write and read operations with appropriate voltages and reference resistance generate distinct signatures for these fault origins. The diagnostic resolution is therefore significantly improved by at least 96.3%.

Compared to the results obtained from the diagnosis sequence in [1], employing the proposed sequence in (6) substantially improves the diagnostic resolution for both the slow-to-fall fault and the unintended switch. This is because the diagnosis sequence in [1] lacks coverage for MIV short defects. When a slow-to-fall fall is detected during the manufacturing test, the diagnosis sequence in [1] is insufficient to distinguish process variations from MIV shorts. Moreover, for MIV short defects, the contact shorted to the MIV is unable to be identified. With the proposed sequence in (6), the diagnostic resolution for both the slow-to-fall fault and the unintended switch can achieve 100%, addressing the limitations encountered by the sequence in [1].

Note that there is a 3.7% decrease in diagnostic resolution for the stuck-at-1 fault with the Stanford model [18] and a 1.1% decrease with the VTEAM model [30] when appending the proposed sequence to the test algorithm [27]. This reduction is primarily due to misidentifying samples with MIV opens that lead to an R_L value below the reference resistance of the r_{MIV} operation. However, such a loss can be compensated through the subsequent physical failure analysis. The output responses generated by our diagnosis sequence aim to provide early feedback prior to more extensive and destructive physical failure analysis. A 93.8% diagnostic

resolution is adequate for the foundry to conduct volume diagnoses and to review the immature M3D and RRAM manufacturing processes. Note that the test algorithm in [13] does not consider the USF, which may cause defective chips to pass the manufacturing test (i.e., lead to test escape). As such defective chips will not proceed to the diagnosis process, the diagnostic resolution cannot be improved by the proposed diagnosis sequence.

Regarding the results obtained with the VTEAM model [30], all test algorithms can achieve 100% diagnostic resolution for the stuck-at-0 fault because only one defect among MIV defects and RRAM process variations leads to a stuck-at-0 fault, as shown in Table VI. The distinct output responses during testing for stuck-at-0 faults are sufficient to distinguish this defect from other fault origins. However, for other fault models and types of faulty behavior, the proposed diagnosis sequence becomes essential to correctly identify the root-cause fault origins.

Table IX provides the diagnosability results for test algorithms, both with and without the proposed diagnosis sequence. Diagnosability serves as a valuable metric for evaluating the effectiveness of the diagnosis process for each defect or variation. For example, the MIV open defect is crucial to M3D-integrated RRAM architectures, but existing test algorithms have 0% diagnosablility with both models because all faults induced by MIV opens can also be caused by RRAM process variations. Simply detecting these faults during testing is insufficient for diagnosing them accurately and identifying their correct origin. Note that the diagnosability of [27] standalone is up to 90.3% with the VTEAM model in [30] when RRAM process variations are present. This is because defective cells with RRAM process variations primarily generate USFs, which can yield distinct output signatures during testing. However, diagnosability for the MIV defects is worse than 0.2%; the proposed diagnosis sequence remains indispensable for identifying root-cause fault origins. Moreover, compared to standalone test algorithms and those with the diagnosis sequence in [1], the proposed solution is the only approach that can identify MIV short defects and accurately pinpoint the contact shorted to the MIV. The robust diagnosability across all fault origins demonstrates the effectiveness of our solution for diagnosing faults due to manufacturing defects and process variations, which is key for yield learning.

TABLE VIII

EVALUATION OF DIAGNOSTIC RESOLUTION FOR TEST ALGORITHMS WITH AND WITHOUT THE PROPOSED DIAGNOSIS SEQUENCE

Test algorithm	Stanford model [18]					VTEAM model [30]				
	Stuck-at-0	Stuck-at-1	Slow-to-fall	USF	Unintended switch	Stuck-at-0	Stuck-at-1	Slow-to-fall	USF	Unintended switch
[13] standalone	100.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%
[13] with diagnosis sequence [1]	100.0%	96.3%	0.0%	0.0%	0.0%	100.0%	99.1%	0.0%	0.0%	0.0%
[13] with Equation (6)	100.0%	96.3%	100.0%	0.0%	100.0%	100.0%	99.1%	100.0%	0.0%	100.0%
[27] standalone	100.0%	0.0%	0.0%	100.0%	0.0%	100.0%	0.0%	0.0%	100.0%	0.0%
[27] with diagnosis sequence [1]	100.0%	96.3%	0.0%	100.0%	0.0%	100.0%	99.1%	0.0%	100.0%	0.0%
[27] with Equation (6)	100.0%	96.3%	100.0%	100.0%	100.0%	100.0%	99.1%	100.0%	100.0%	100.0%

TABLE IX

EVALUATION OF DIAGNOSABILITY FOR TEST ALGORITHMS WITH AND WITHOUT THE PROPOSED DIAGNOSIS SEQUENCE

Test Algorithm	Stanford model [18]			VTEAM model [30]		
	RRAM process variations	MIV open	MIV short	RRAM process variation	MIV open defect	MIV short defect
[13] standalone	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
[13] with diagnosis sequence [1]	5.8%	98.5%	0.0%	8.6%	99.8%	0.0%
[13] with Equation (6)	88.3%	98.5%	100.0%	8.6%	99.8%	100%
[27] standalone	11.4%	1.5%	0.0%	90.3%	0.2%	0.0%
[27] with diagnosis sequence [1]	16.4%	100.0%	0.0%	98.9%	100.0%	0.0%
[27] with Equation (6)	99.7%	100.0%	100.0%	98.9%	100.0%	100.0%

The runtime of March test schemes depends on the total number of read and write operations multiplied by the number of cells under test. In (6), ten operations are included in the proposed diagnosis sequence; therefore, the runtime is increased by $10N$, where N is the number of RRAM cells in an array. Although the proposed diagnosis process leads to an increase in the runtime, it is conducted only on chips that fail the manufacturing test. There is no test-time overhead for fault-free devices. Moreover, the proposed solution can distinguish MIV defects from every source of RRAM process variation/defect, which is important for M3D-integrated architectures to facilitate yield learning and shorten time-to-market. Because no additional DfT structures are needed in the proposed algorithm, the proposed solution can be applied to any large-size high-density array without impacting the benefits of M3D integration.

VI. DISCUSSION

A. Reference Resistance for the r_{MIV} Operation

In the proposed diagnosis sequence, five read operations with different reference resistance values are required to help distinguish between root-cause fault origins. As discussed in [1], reference values for both r_{ref0} and r_0 are inherent properties of memristors, while additional steps are needed to identify the appropriate value for r_{MIV}. Conducting SPICE simulation is the most accurate and straightforward way to obtain these resistance values. However, different technology nodes and design environments have unique features (e.g., supply voltage). Simulations have to be repeated when features are changed, leading to extra test cost and runtime. Therefore, it is important to derive a generalized solution that is compatible with multiple design environments.

The reference resistance value for r_{MIV} depends on the minimum R_o to cause stuck-at-1 faults (see Section IV-A). A large R_o leads to a decrease in voltage across the defective memristor due to the voltage droop, making the electric field between TE and BE insufficient to switch the resistance state. The relationship between the electric field and the voltage across the memristor in [18] is shown as follows:

$$\left(\gamma_0 - \beta \times g^3\right) \times \frac{|V|}{t_{\text{ox}}} \geq F_{\text{min}} \tag{7}$$

where V is the voltage across the memristor, g is the tunneling gap distance between CF and TE, γ_0 and β are fitting parameters, and t_{ox} and F_{min} are the parameters as described in Table I. The current-voltage (I–V) relationship of the memristor is provided as

$$I = I_0 \times \exp\left(-\frac{g}{g_0}\right) \times \sinh\left(\frac{V}{V_0}\right) \tag{8}$$
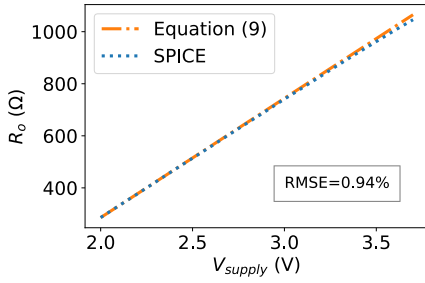
Fig. 10. Comparison between values calculated with (10) and values from SPICE simulation for the minimum R_o needed to cause a SAF, where RMSE is the normalized root-mean-square error between the calculated values and simulation results. This figure is adapted from [1].

Fig. 11. RRAM behavior analysis with samples following uniform distribution. (a) Correlation between variation in $G_{\min}$ and variation in t_{ox}. (b) Correlation between R_L and variation in t_{ox}.

where I is the current flow through the memristor, and I_0, g_0, V_0 are fitting parameters. The impact of voltage droop on V can be modeled as

$$V = V_{\text{supply}} - I \times (R_w + R_T + R_o) \qquad (9)$$

where V_{supply} is the nominal supply voltage, R_w is the equivalent resistance along the conduction path, and R_T refers to the ON-resistance of the access transistor. If the inequality in (7) does not hold because V is significantly reduced by the voltage droop, the memristor can no longer switch, leading to a SAF.

To identify the reference values for the r_{MIV} operation, we need to consider the condition where the equality in (7) holds. Therefore, V can be rewritten as $[F_{\min}t_{ox}(\gamma_0 - \beta \times g^3)]$ and substituted into (8) and (9), and the following relationship can be obtained:

$$R_w + R_T + R_o = \frac{(\gamma_0 - \beta \times g^3) V_{\text{supply}} - F_{\min}t_{ox}}{(\gamma_0 - \beta \times g^3) I_0 \times \exp\left(-\frac{g}{g_0}\right) \times \sinh\left(\frac{F_{\min}t_{ox}}{(\gamma_0 - \beta \times g^3)V_0}\right)}. \qquad (10)$$

For the r_{MIV} operations, the memristor is assumed to be fault-free and successfully initialized to the nominal LRS. Hence, by replacing g with $G_{\min}$ in Table I, we can obtain the minimum R_o that can cause an SAF under different supply voltages. Note that R_w and R_T are estimated values from the design kit for the RRAM array. Fig. 10 presents the comparison between the minimum R_o calculated using (10) and the minimum R_o obtained from SPICE simulation. The RMSE between calculation results and simulation results is less than 1%. The difference between calculated values and simulation results is due to the equivalent ON-resistance of the access transistor. We utilize a fixed R_T for approximation in the generalized solution, while the ON-resistance of access transistors varies with the change of supply voltage. However, the RMSE is below 1%, which is acceptable for the derivation of the reference resistance. Therefore, (10) can be utilized as a generalized solution to identify the appropriate reference resistance for the r_{MIV} operations without conducting any simulation.

B. Identification of Sizes of Process Variations

In addition to pinpointing the root-cause fault origins, the RRAM behavior characteristics combined with the diagnosis
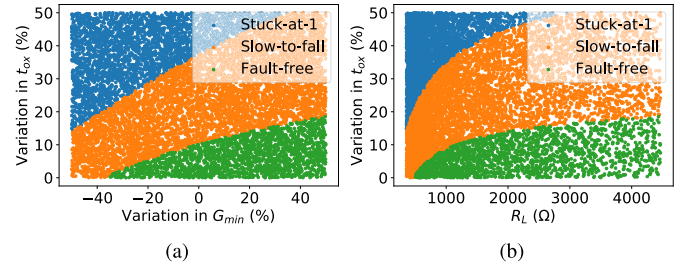
sequence can serve as a tool for estimating the range of process variations that lead to faulty behavior. For example, when a USF is detected and the diagnosis signatures indicate that the fault is attributed to RRAM process variations, the range of the decrease in $G_{\max}$ with the Stanford model [18] typically falls between 26.0% and 38.5%, as illustrated in Fig. 5(b). Similarly, for the VTEAM model [30], a decrease in R_{OFF} can be estimated to range from 32.1% to 48.2%.

However, for the faulty cells experiencing joint variations in $G_{\min}$ and t_{ox}, it becomes challenging to identify the variations in these two parameters solely based on the output responses generated by our diagnosis sequence. This complexity arises because both decreases and increases in $G_{\min}$, when combined with specific variations in oxide thickness (as discussed in Section III-A), can lead to faulty behavior.

To gain deeper insights into the effects of joint variations in $G_{\min}$ and t_{ox} on RRAM behavior, we conducted SPICE simulations using samples following a uniform distribution. In this distribution, $G_{\min}$ ranges from a 50% decrease to a 50% increase compared to its nominal value, while t_{ox} ranges from 0% to 50% of its nominal value. The results of this RRAM characterization are shown in Fig. 11(a).

Similar to the trends observed in Fig. 4(b), with a fixed variation in $G_{\min}$, a significant increase in t_{ox} leads to a stuck-at-1 fault, while a minor increase in t_{ox} causes a slow-to-fall fault. An increase in $G_{\min}$ requires a larger increase in t_{ox} to induce faulty behavior. This is reasonable because an increase in $G_{\min}$ enlarges the resistance value at the LRS (i.e., R_L). Therefore, the RRAM has an increased voltage across its TE and BE due to charge sharing with equivalent resistance along the conduction path. When a stuck-at-1 fault is caused, a large increase in t_{ox} is anticipated to make the electric field across the device insufficient for switching the state. This relationship can be observed by corresponding the variation in $G_{\min}$ to the physical values of R_L, as demonstrated in Fig. 11(b). When the fault origin is identified as joint variations in $G_{\min}$ and t_{ox} through the proposed diagnosis sequence, Fig. 11(b) can be employed to determine the range of variations in t_{ox} based on the value of R_L. The sizes of other RRAM process variations that lead to faulty behavior can be determined by referring to the characteristics analyzed during Monte Carlo simulations in Section III.

C. Sizes of MIV Short Defects

To determine the location of the MIV short, we introduce two additional operations, denoted as D_g and D_s, into the
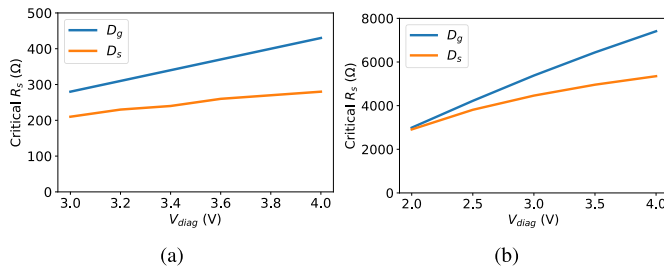
Fig. 12. Critical R_s leading to RRAM switching during D_g and D_s operations with various V_{diag}. (a) Stanford model [18]. (b) VTEAM model [30].

proposed diagnosis sequence. These operations are designed by deactivating all access transistors and applying a voltage, labeled as V_{diag}, to the WL and SL of the target top-tier RRAM, respectively. During D_g, if an MIV inside a bottom-tier RRAM is shorted to the gate contact of the target cell, a current can pass through the MIV short R_s to switch the state of such a RRAM. Similarly, the state of a bottom-tier RRAM is changed during the D_s operation when the corresponding MIV is shorted to the source contact of the top-tier transistor. Because whether the state of the bottom-tier RRAM is changed depends on the charge sharing between R_s and the memristor, we can alter the applied voltage to identify the range of the size of R_s at present.

Fig. 12(a) and (b) show the correlation between critical value of R_s and V_{diag} for the Stanford model [18] and the VTEAM model [30], respectively, where critical R_s is the maximum R_s leading to a RRAM state change. Note that the difference of critical R_s between D_g and D_s operations is due to the wire resistance. As V_{diag} increases, the critical R_s also increases because additional charges are provided by the voltage source. Conversely, a small V_{diag} requires a more server MIV short defect (i.e., a smaller R_s) to induce a state change in the RRAM. Leveraging this property, we can manipulate V_{diag} to narrow down the range of possible values for R_s. This approach helps in analyzing the severity of MIV short defects within an M3D-integrated RRAM array, which is important for yield learning.

VII. Conclusion

We have presented a characterization of RRAM cells in the presence of MIV defects and concurrent RRAM process variations. Based on the cell behaviors, we have proposed a diagnosis sequence to identify root-cause fault origins with additional read/write operations with appropriate reference resistance and applied voltage. We have demonstrated that the proposed solution is compatible with existing March test algorithms for improving diagnostic resolution and diagnosability. We have provided a generalized solution to derive reference resistance under different environments without conducting any SPICE simulation. We have also discussed approaches for identifying the sizes of process variations and MIV shorts based on the RRAM characteristics and the applied voltages. As no additional DfT structure is required, the proposed solution is applicable to large-size high-density M3D-integrated RRAM architectures to help distinguish MIV defects from RRAM process variations.

References

[1] S.-C. Hung, A. Chaudhuri, S. Banerjee, and K. Chakrabarty, "Fault diagnosis for resistive random-access memory and monolithic inter-tier vias in monolithic 3D integration," in *Proc. IEEE Int. Test Conf. (ITC)*, Sep. 2022, pp. 118–127.

[2] A. Chen, "A review of emerging non-volatile memory (NVM) technologies and applications," *Solid-State Electron.*, vol. 125, pp. 25–38, Nov. 2016.

[3] Y. S. Chen et al., "Challenges and opportunities for HfOX based resistive random access memory," in *IEDM Tech. Dig.*, Dec. 2011, p. 31.

[4] C. Xu, D. Niu, Y. Zheng, S. Yu, and Y. Xie, "Impact of cell failure on reliable cross-point resistive memory design," *ACM Trans. Des. Automat. Electron. Syst.*, vol. 20, no. 4, p. 63, 2015.

[5] E. Esmanhotto et al., "High-density 3D monolithically integrated multiple 1T1R multi-level-cell for neural networks," in *IEDM Tech. Dig.*, Dec. 2020, p. 36.

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

[7] S. Panth, S. Samal, Y. S. Yu, and S. K. Lim, "Design challenges and solutions for ultra-high-density monolithic 3D ICs," in *Proc. SOI-3D-Subthreshold Microelectron. Technol. Unified Conf. (S3S)*, Oct. 2014, pp. 1–2.

[8] M. M. Shulaker et al., "Monolithic 3D integration of logic and memory: Carbon nanotube FETs, resistive RAM, and silicon FETs," in *IEDM Tech. Dig.*, Dec. 2014, p. 27.

[9] G. Murali, X. Sun, S. Yu, and S. K. Lim, "Heterogeneous mixed-signal monolithic 3-D in-memory computing using resistive RAM," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 29, no. 2, pp. 386–396, Feb. 2021.

[10] F. Zokaee, M. Zhang, X. Ye, D. Fan, and L. Jiang, "Magma: A monolithic 3D vertical heterogeneous reram-based main memory architecture," in *Proc. 56th ACM/IEEE Design Autom. Conf. (DAC)*, Jun. 2019, pp. 1–6.

[11] D. Niu, Y. Chen, C. Xu, and Y. Xie, "Impact of process variations on emerging memristor," in *Proc. 47th Design Automat. Conf.*, Jun. 2010, pp. 877–882.

[12] C.-Y. Chen et al., "RRAM defect modeling and failure analysis based on March test and a novel squeeze-search scheme," *IEEE Trans. Comput.*, vol. 64, no. 1, pp. 180–190, Jan. 2015.

[13] P. Liu, Z. You, J. Kuang, Z. Hu, H. Duan, and W. Wang, "Efficient March test algorithm for 1T1R cross-bar with complete fault coverage," *Electron. Lett.*, vol. 52, no. 18, pp. 1520–1522, Jun. 2016.

[14] N. Z. Haron and S. Hamdioui, "DfT schemes for resistive open defects in RRAMs," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2012, pp. 799–804.

[15] S. Hamdioui, M. Taouil, and N. Z. Haron, "Testing open defects in memristor-based memories," *IEEE Trans. Comput.*, vol. 64, no. 1, pp. 247–259, Jan. 2015.

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

[17] A. Chaudhuri et al., "Built-in self-test and fault localization for inter-layer vias in monolithic 3D ICs," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 18, no. 1, p. 22, 2021.

[18] Z. Jiang, S. Yu, Y. Wu, J. H. Engel, X. Guan, and H.-S.-P. Wong, "Verilog-A compact model for oxide-based resistive random access memory (RRAM)," in *Proc. Int. Conf. Simulation Semiconductor Processes Devices (SISPAD)*, Sep. 2014, pp. 41–44.

[19] Y. Ho, G. M. Huang, and P. Li, "Dynamical properties and design analysis for nonvolatile memristor memories," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 58, no. 4, pp. 724–736, Apr. 2011.

[20] S. Wong et al, "Monolithic 3D integrated circuits," in *Proc. Int. Symp. VLSI Technol., Syst. Appl. (VLSI-TSA)*, Apr. 2007, pp. 1–4.

[21] L. Brunet et al., "Breakthroughs in 3D sequential technology," in *IEDM Tech. Dig.*, Dec. 2018, p. 7.

[22] C.-M. V. Lu et al., "Key process steps for high performance and reliable 3D sequential integration," in *Proc. Symp. VLSI Technol.*, Jun. 2017, pp. T226–T227.

[23] Y. Yu and N. K. Jha, "SPRING: A sparsity-aware reduced-precision monolithic 3D CNN accelerator architecture for training and inference," *IEEE Trans. Emerg. Topics Comput.*, vol. 10, no. 1, pp. 237–249, Jan. 2022.

[24] A. Koneru, S. Kannan, and K. Chakrabarty, "Impact of electrostatic coupling and wafer-bonding defects on delay testing of monolithic 3D integrated circuits," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 13, no. 4, p. 54, 2017.

[25] S. Kannan, J. Rajendran, R. Karri, and O. Sinanoglu, "Sneak-path testing of crossbar-based nonvolatile random access memories," *IEEE Trans. Nanotechnol.*, vol. 12, no. 3, pp. 413–426, May 2013.

[26] C. Münch and M. B. Tahoori, "Testing resistive memory based neuromorphic architectures using reference trimming," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Feb. 2021, pp. 1592–1595.

[27] P. Liu, Z. You, J. Wu, B. Liu, Y. Han, and K. Chakrabarty, "Fault modeling and efficient testing of memristor-based memory," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 11, pp. 4444–4455, Nov. 2021.

[28] D. Gilliland and V. Melfi, "A note on confidence interval estimation and margin of error," *J. Statist. Educ.*, vol. 18, no. 1, 2010, doi: 10.1080/10691898.2010.11889474.

[29] *Web Link*. Accessed: Sep. 8, 2023. [Online]. Available: https://www.zoho.com/survey/margin-of-error.html

[30] S. Kvatinsky, M. Ramadan, E. G. Friedman, and A. Kolodny, "Vteam: A general model for voltage-controlled memristors," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 62, no. 8, pp. 786–790, Aug. 2015.

[31] G. González-Cordero, J. B. Roldan, F. Jiménez-Molinos, J. Suñé, S. Long, and M. Liu, "A new compact model for bipolar RRAMs based on truncated-cone conductive filaments—A verilog-a approach," *Semicond. Sci. Technol.*, vol. 31, no. 11, 2016, Art. no. 115013.

[32] D. Panda, P. P. Sahu, and T. Y. Tseng, "A collective study on modeling and simulation of resistive random access memory," *Nanosc. Res. Lett.*, vol. 13, no. 1, p. 8, Dec. 2018.

[33] *International Technology Roadmap for Semiconductors*, ITRS, Denver, CO, USA, 2011.

Arjun Chaudhuri received the bachelor's degree in electronics and electrical communication engineering from IIT Kharagpur, Kharagpur, India, in 2017, and the Ph.D. degree in electrical and computer engineering from Duke University, Durham, NC, USA, in 2022.

He is currently a DFT Methodology Engineer at NVIDIA Corporation, Santa Clara, CA, USA. His research interests include design-for-testability and fault tolerance of machine learning hardware and monolithic 3-D integrated circuits.

Sanmitra Banerjee received the B.Tech. degree from IIT Kharagpur, Kharagpur, India, in 2018, and the M.S. and Ph.D. degrees from Duke University, Durham, NC, USA, in 2021 and 2022, respectively.

He is currently a Senior DFX Methodology Engineer at NVIDIA Corporation, Santa Clara, CA, USA. His research interests include machine-learning-based DFX techniques, and fault modeling and optimization of emerging AI accelerators under process variations and manufacturing defects.

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

He was the John Cocke Distinguished Professor and the Chair of Electrical and Computer Engineering at Duke University, Durham, NC, USA. He is currently the Fulton Professor of Microelectronics at the School of Electrical, Computer and Energy Engineering, Arizona State University (ASU), Tempe, AZ, USA. His current research interests include design-for-testability of 2.5-D/3-D integrated circuits, hardware security, failure prediction using AI/ML, AI accelerators, microfluidic biochips, AI for healthcare, and neuromorphic computing systems.

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

He was an Intern with NVIDIA Corporation, Santa Clara, CA, USA, and Cadence Design Systems, Austin, TX, USA. His current research interests include reliability, testing, and diagnosis of monolithic 3-D integrated circuits.