

Field- and Temperature-Induced Drift in Multi-Level Al-Doped HfO₂ RRAM

Andrea Baroni, Jianan Wen[✉], Eduardo Pérez[✉], Keerthi Dorai Swamy Reddy[✉], Max Uhlmann[✉], *Member, IEEE*, Stefan Pechmann, Christian Wenger[✉], Daniele Ielmini[✉], *Fellow, IEEE*, and Cristian Zambelli, *Member, IEEE*

Abstract—Drift of multi-level conductance states is a key reliability concern for resistive random access memory (RRAM) used in multi-level storage and in-memory computing applications. This work experimentally quantifies how repeated read operations and elevated-temperature storage reshape multiple conductive states in Al5%:HfO₂-based RRAM cells. Systematic read-disturb and multi-temperature retention experiments are performed on one-transistor-one-resistor (1T1R) arrays, and their evolution is analyzed through intra-level transition matrices and a distance-weighted reshuffling factor F_r that jointly capture internal reshuffling dynamics beyond median and spread. Read-disturb exhibits a pronounced voltage dependence, with increasing reshuffling as V_{RD} rises, whereas retention shows a two-phase thermal evolution dominated by an early warm-up reshuffling followed by a slower isothermal regime. Comparing incremental step pulse and verify algorithm (ISPVA) and one-shot programming (OSP) under identical targets and read-stress conditions further shows that both algorithms experience redistribution, but with different directionality and level dependence. These results clarify the distinct statistical signatures of field- and temperature-induced drift and identify the programming algorithm, read-voltage budget, and warm-up handling as key design knobs to improve the number of reliable levels in multi-level RRAM arrays.

Index Terms—RRAM, HfO₂, aluminum doping, Multi-level cell (MLC), Read disturb, drift, retention.

Received 24 March 2026; accepted 3 June 2026. Date of publication 5 June 2026; date of current version 12 October 2026. This work was supported in part by the Ministry of Science, Research and Culture of the State of Brandenburg within the Program “Strengthening of Technological and Applied Research” (StaF) under Grant 86004028; in part by European Research Council (ERC) under European Union’s Horizon 2020 Research and Innovation Program under Grant 648635; in part by the Deutsche Forschungsgemeinschaft (DFG), German Research Foundation under Project PE 2657/4-1; and in part by the Spoke 1 “FutureHPC and BigData” of Italian Research Center on High-Performance Computing, Big Data and Quantum Computing (ICSC), funded by MUR Missione 4-Next Generation EU (NGEU). (*Corresponding author: Eduardo Pérez.*)

Andrea Baroni, Jianan Wen, Keerthi Dorai Swamy Reddy, and Max Uhlmann are with the IHP - Leibniz Institute for High Performance Microelectronics, 15236 Frankfurt (Oder), Germany.

Eduardo Pérez and Christian Wenger are with the IHP - Leibniz Institute for High Performance Microelectronics, 15236 Frankfurt (Oder), Germany, and also with BTU Cottbus-Senftenberg, 03046 Cottbus, Germany (e-mail: perez@ihp-microelectronics.com).

Stefan Pechmann is with the Chair of Micro- and Nanosystems Technology, Technical University of Munich, 85748 Munich, Germany.

Daniele Ielmini is with the Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano and IU.NET, 20133 Milan, Italy.

Cristian Zambelli is with the Dipartimento di Ingegneria, Università degli Studi di Ferrara, 44122 Ferrara, Italy.

Digital Object Identifier 10.1109/TDMR.2026.3700862

I. INTRODUCTION

RESISTIVE Random Access Memory (RRAM) is a promising candidate for next-generation non-volatile memories and in-memory computing accelerators thanks to its scalability, low-voltage operation, and fast switching [1], [2], [3], [4]. In particular, filamentary RRAM arrays enable energy-efficient vector-matrix multiplication (VMM) for artificial neural networks (ANNs) by storing synaptic weights directly in the device conductance [5], [6], [7], [8]. Multi-level cell (MLC) operation further increases the effective bit-per-cell density, but also amplifies the impact of device-to-device (D2D) and cycle-to-cycle (C2C) variability, and of conductance drift over time [9], [10], [11]. These phenomena can severely degrade inference accuracy and system-level reliability [8], [12], [13], [14], [15], where a key limitation lies in the progressive overlap between programmed conductance levels distribution. Short- and long-timescale drift are known to depend on operating conditions such as time, temperature, and target resistive state [13], [16], [17], [18], [19]. In filamentary RRAM, drift originates from thermally activated migration and redistribution of oxygen vacancies within the conductive path and its surrounding region. As a result, both the mean conductance and the spread of the multi-level distributions evolve with stress, blurring the boundaries between levels. Material engineering and programming algorithms have been widely explored to mitigate these issues. In particular, Al doping in HfO₂ has been shown to stabilize grain boundaries and suppress lateral vacancy diffusion, improving endurance and retention while reducing variability [20], [21], [22], [23], [24], [25]. In our previous work [26], we compared pure HfO₂ with Al-doped HfO₂ devices in a 1T1R 130 nm CMOS technology, demonstrating that Al5%:HfO₂ offers a near-optimal trade-off between forming yield, MLC variability, endurance, and high-temperature retention.

However, several open questions remain. First, the evolution of the full multi-level distributions is often summarized by median and standard deviation, which do not capture which devices populate the distribution tails and whether tail membership is preserved or reshuffled over time. Second, read-disturb is usually discussed in terms of unwanted soft-programming of the conductive states, whereas read-induced drift toward lower conductance at intermediate read voltages has not been explicitly characterized [14], [15], [27]. Finally, the impact of the programming algorithm itself on drift mechanisms is rarely quantified.

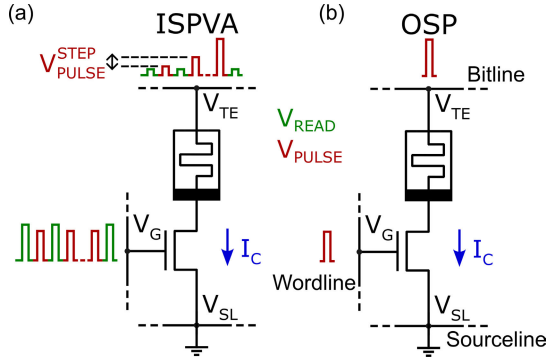


Fig. 1. Schematic representation of the ISPVA (a) and OSP (b) programming schemes.

In this work, we address these gaps by systematically studying how repeated read operations and temperature stress reshape the programmed multi-level distributions in Al5%:HfO₂ RRAM devices integrated in a 1T1R architecture. We introduce statistical metrics to distinguish reshuffling from distribution shift in field-biased and thermally active environments, and to evaluate the role of the programming algorithm.

II. DEVICES AND EXPERIMENTAL METHODS

A. Al5%:HfO₂ 1T1R Cell Structure

The devices are 1T1R RRAM cells fabricated in a 130-nm CMOS technology at IHP [28], [29]. Each cell integrates an nMOS selector ($L=130$ nm, $W=1000$ nm) in series with a BEOL MIM stack (footprint of 600×600 nm²) placed between Metal 2 and Metal 3. The stack is TiN/Al5%:HfO₂ (4 nm)/Ti/TiN, with Al:HfO₂ deposited by ALD. 1T1R cells are integrated in 4-kbit matrices (64×64). The 1T1R configuration enables individual cell selection and ensures precise control of the compliance current I_C via gate-voltage V_G regulation, which is critical for accurate MLC operation. All measurements in this work were carried out at wafer level on 200 mm substrates using a Cascade Microtech PA200 probe station connected to NplusT TestMesh100 automated test system.

B. Multi-Level Programming Schemes

We explore the performance of seven conductance states (one HRS and six LRS). They are defined by target read currents at the reference voltage $V_{\text{read}} = 0.2$ V: $5 \mu\text{A}$ for HRS, and 10 – $60 \mu\text{A}$ for LRS1–LRS6 in $10 \mu\text{A}$ increments.

Two programming schemes are considered in this work.

1) *Incremental Step Pulse and Verify Algorithm (ISPVA)*: The baseline algorithm is the ISPVA, applied to forming, Set, and Reset operations, similar to [29], [30], and [31]. For each Set (Reset) operation, a sequence of voltage pulses with fixed width and gradually increasing amplitude is applied to the bitline (sourceline), while the wordline sets the target compliance current. After each pulse, a verify read at V_{read} is performed. The programming-pulse amplitude is increased until the measured current reaches the target value or a maximum number of pulses is reached. A schematic of the ISPVA flow is shown in Fig. 1(a). Thanks to the closed-loop

verify step, ISPVA provides tight control of the programmed conductance and limits over-programming, at the cost of multiple write pulses per operation and increased write latency.

2) *One-Shot Programming (OSP)*: To assess the impact of the programming algorithm on drift mechanisms, we also employ an OSP scheme (Fig. 1(b)). Each level is programmed using a single voltage pulse with fixed amplitude and width, chosen to achieve the same nominal target currents as ISPVA but without intermediate verify steps. The gate voltage again controls the compliance current. After the program pulse, the algorithm ends. Then a single verify read at V_{read} is performed to record the resulting conductance.

C. Read-Disturb Stress Protocol

Read-disturb experiments are performed on 512 devices per level, programmed using either ISPVA or OSP. After programming, a reference read at V_{read} is carried out to characterize the starting distributions and ensure an operating yield above 90%. In this work, read-disturb stress consists of sequences of $N_{\text{RD}} = 10^5$ read pulses with duration $t_{\text{RD}} = 10 \mu\text{s}$ and amplitude V_{RD} in the 0.1 – 0.5 V range, achieving a total stress time of 1 s. The pulses are applied to the bitline while the gate voltage is set to the same value used for standard read operations. For each stress condition, the following sequential procedure is adopted:

- 1) Program all cells to the target conductance.
- 2) Verify the reference current at V_{read} .
- 3) Measure the initial current I_0 at V_{RD} in a single cell.
- 4) Apply all N_{RD} read pulses at fixed V_{RD} on the cell.
- 5) Move to the following cell and repeat from step 3.

D. Temperature Stress and Retention Protocol

Retention experiments are carried out on ISPVA-programmed cells. For each temperature and level, 512 devices are programmed to the target state at room temperature (RT) at $\approx 25^\circ\text{C}$. The samples are then subjected to isothermal storage at temperatures $T \in \{RT, 85, 105, 125, 150\}^\circ\text{C}$ in a temperature-controlled chuck. After a stabilization period at the target temperature lasting 30 min, the current of each cell is measured at the reference V_{read} without additional programming. This is referred to as the 0 h time point. Then, at each temperature, the following retention times are considered: $t \in \{0, 1, 3, 10, 30, 100\}$ h; and the conductance levels measured at V_{read} . The resulting multi-level current distributions are used to analyze distribution shift, tails evolution, and intra-level reshuffling as a function of temperature and time.

III. DRIFT AND RESHUFFLING METRICS

To quantify the evolution of the multi-level conductance distributions under field- and temperature-induced stress, we introduce a set of statistical metrics that quantify the internal reshuffling of device populations. This section describes the common framework used throughout the paper. The analysis is performed separately for each programmed level (HRS, LRS1–LRS6) and for each stress condition.

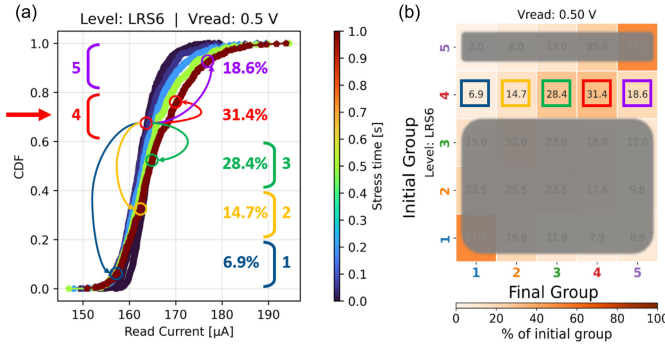


Fig. 2. Intra-level binning and transition matrix example (LRS6, ISPVA). (a) CDF split into $K = 5$ equiprobable bins. (b): transition matrix P_{ij} .

A. Intra-Level Binning and Transition Matrices

For a given level L , we build the empirical cumulative distribution function (CDF) of the read current at an initial reference time t_0 . The CDF is then divided into K equiprobable groups, with $K = 5$ in this work. The resulting bins correspond approximately to the CDF intervals $[0, 0.2)$, $[0.2, 0.4)$, $[0.4, 0.6)$, $[0.6, 0.8)$, and $[0.8, 1]$, which we label by an integer index $i \in \{1, \dots, K\}$ from the lowest- to the highest-current region. At a later time t_f (after stress), each device is again assigned to a bin index $j \in \{1, \dots, K\}$ based on its rank within the current CDF at t_f . In this way, the bin index describes the relative position of a device within the level distribution at a given time, and its evolution reveals how the internal structure of the distribution is reshaped under stress, as illustrated in Fig. 2(a).

To describe how devices migrate between different regions of the distribution, we construct a transition matrix for each level and stress condition, as shown in Fig. 2(b). Let i and j denote the bin indices of a device at the initial time t_0 and at the final time t_f , respectively. By tracking all devices in the population, we count the number of transitions N_{ij} from initial bin i to final bin j . We also define the population of the initial bin i as:

$$N_i = \sum_{j=1}^K N_{ij}. \quad (1)$$

The row-normalized transition matrix is then obtained as:

$$P_{ij} = \frac{N_{ij}}{N_i}, \quad \sum_{j=1}^K P_{ij} = 1, \quad (2)$$

so that P_{ij} represents the conditional probability that a device initially belonging to bin i migrates to bin j after stress. The main diagonal terms P_{ii} quantify the fraction of devices that remain in the same region of the distribution, corresponding to a preserved internal ordering. Off-diagonal elements capture changes in relative position within the distribution: transitions to higher-index bins ($j > i$) indicate that a device ends up in a more conductive region of the final CDF, whereas transitions to lower bins ($j < i$) indicate a move toward less conductive regions. However, since the binning is defined in terms of percentiles at t_0 and t_f , these transitions reflect changes in rank rather than the sign of the absolute current drift, which must

be evaluated by directly comparing I_f and I_0 . By examining the full $K \times K$ matrix, we can identify which initial regions (e.g., tails vs. core) are more prone to migration and which final regions they tend to populate.

In the read-disturb analysis of Section IV, the transition matrices provide a compact representation of level stability and intra-level reshuffling under different read voltages and programming schemes. For the retention experiments discussed in Section V, the same concept is used, while the temperature-time dependence of the tails is further analyzed in terms of failure fractions and MTTF.

B. Distance-Weighted Reshuffling Factor

While transition matrices provide detailed information on the migration between bins, it is convenient to summarize the overall amount of reshuffling into a single scalar metric. To this end, we define a distance-weighted reshuffling factor F_r .

Let w_i denote the population weight of the initial bin i , i.e., the fraction of devices that belong to bin i at time t_0 , such that $\sum_i w_i = 1$. We also define a normalized distance $d(i, j)$ between bins i and j as:

$$d(i, j) = \frac{|i - j|}{K - 1} \in [0, 1], \quad (3)$$

which is 0 if a device remains in the same bin and 1 only for the largest possible jump between the lowest and highest bins. The reshuffling factor is then defined as:

$$F_r = \sum_{i=1}^K w_i \sum_{j=1}^K P_{ij} d(i, j) = \sum_{i=1}^K \sum_{j=1}^K \frac{N_{ij}}{N} d(i, j), \quad (4)$$

where the second form follows from $w_i = N_i/N$ and N is the total number of devices for a given level. For each initial bin i , the inner sum $\sum_j P_{ij} d(i, j)$ represents the average normalized jump distance of devices starting from that bin. The outer average with weights w_i yields the mean jump distance over the entire level population.

By construction, F_r is bounded in $[0, 1]$, and its numerical values depend on the chosen number of bins K . We therefore summarize the main interpretation guidelines below:

- $F_r = 0$ if and only if all devices remain in their initial bin (no reshuffling, fully preserved internal ordering);
- F_r increases with both the probability of leaving the diagonal (P_{ij} with $i \neq j$) and the distance of the jumps;
- For $K = 5$ (used in this work), a fully random reassignment across bins ($P_{ij} \approx 1/K$) yields $F_r \approx 0.4$, while highly non-local reshuffling corresponds to larger F_r values approaching unity.

It is important to note that F_r is agnostic to the direction of the drift and focuses purely on the magnitude of internal redistribution. As a result, two stress conditions can exhibit similar F_r values but qualitatively different drift patterns (dominant soft-programming vs. dominant relaxation), which are distinguished by inspecting the imbalance between upward and downward transitions in the transition matrices and the CDF evolution.

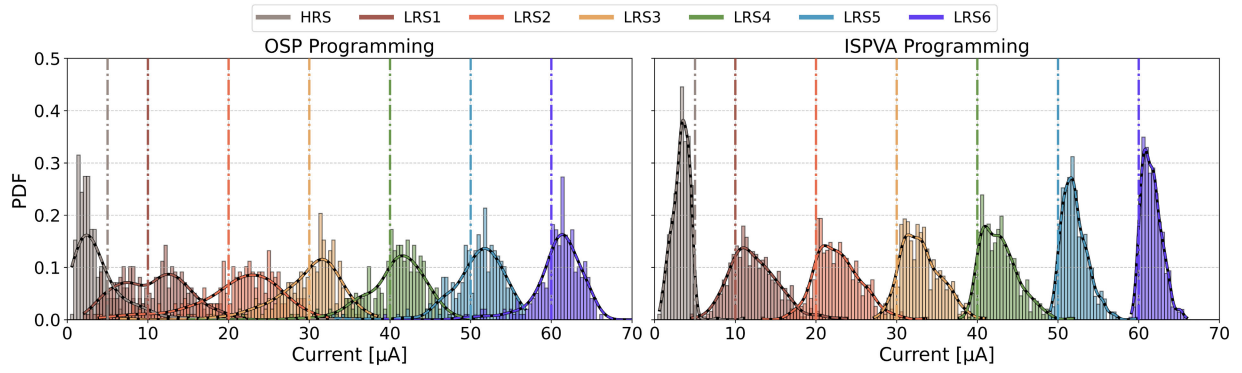


Fig. 3. Initial multi-level distributions at room temperature for OSP and ISPVA. Dashed vertical lines indicate the levels target current.

IV. READ DISTURB

This section reports the read-disturb evolution and reshuffling statistics, first for ISPVA as baseline and then comparing ISPVA vs. OSP. The multi-level distributions obtained at the end of the programming process with both approaches are shown in Fig. 3. Fig. 4 provides an overview of median, interquartile range (IQR), and 5th–95th percentiles evolution vs. cumulative stress time for a few representative levels (HRS, LRS2, LRS4 and LRS6).

A. Baseline Behavior of ISPVA-Programmed Cells

Fig. 5 illustrates the evolution of the read currents for ISPVA-programmed cells under up to 10^5 read pulses at $V_{RD} = 0.1$ – 0.5 V. For each level, the initial current I_0 at V_{RD} is plotted on the horizontal axis, while the final current I_f after stress is reported on the vertical axis. The dashed diagonal corresponds to $I_f = I_0$. At low read voltages ($V_{RD} = 0.1$ – 0.2 V), the scatter points remain tightly clustered around the diagonal for all LRS levels, indicating that only limited drift occurs and that the internal ordering of devices within each distribution is largely preserved. In this regime, the transition matrices are strongly dominated by the main diagonal, with only small probabilities in the nearest off-diagonal entries (Fig. 6 (a)). Consistently, the distance-weighted reshuffling factor F_r remains near 0.1 for most levels.

When the read voltage is increased to $V_{RD} = 0.3$ V, a qualitatively different behavior emerges. The scatter plots begin to show an asymmetric spread away from the diagonal, with a larger fraction of devices drifting toward lower currents while the most conductive part of each level remains almost unchanged. In terms of CDFs, the lower tail of the distribution progressively moves to lower current values, whereas no clear growth of the upper tail is visible. We refer to this regime as induced drift, namely a read-stress activated downward redistribution of conductance values. While the present data do not allow us to unambiguously disentangle electric-field and self-heating contributions, the observed preferential drift toward lower currents is consistent with a thermally assisted relaxation of conductive filaments plausibly assisted by self-heating (Joule dissipation) during repeated read pulses (see Fig. 7 for a representative example on LRS3). In the corresponding transition matrices, the diagonal entries P_{ii} are still

dominant, but higher weights appear in the lower-left off-diagonal elements, corresponding to an average F_r of about 0.15 across the programmed levels.

Further increasing the read voltage up to $V_{RD} = 0.5$ V amplifies and qualitatively modifies this trend. As shown in Fig. 6(b), the transition matrices display substantial probability mass away from the diagonal, involving jumps not only to nearest-neighbor bins but also between core and tail regions. In this high-bias regime, induced drift toward lower currents persists, but it is increasingly accompanied by soft-programming events that push part of the core population toward higher conductance values. The net effect is a pronounced broadening of the distributions in both directions and a marked loss of rank correlation inside each level. Consistently, the average reshuffling factor F_r grows monotonically with V_{RD} and reaches values close to 0.3 at 0.5 V, reflecting the growing average jump distance in the CDF.

B. Impact of Programming Algorithm on Read-Disturb

To assess how the programming algorithm influences read-disturb robustness, we compare ISPVA with OSP under identical conditions: the same nominal target currents are used for all conductive levels and the read-disturb protocol of Section II-C is applied.

The results are summarized in Fig. 8, which focuses on the directionality of read-induced changes by quantifying the incidence of soft-programming events, i.e., the fraction of devices that gain conductivity after the stress compared to their initial values. To make the classification robust against readout noise, we introduce a symmetric guard band and label a device as stable when $|I_f - I_0| \leq \delta$, with $\delta = 1$ μ A. Fig. 8(a) reports the soft-programming fraction for each level (HRS and LRS1–LRS6) at increasing read voltages V_{RD} , showing a strong growth with V_{RD} across the conductive states and indicating that read stress progressively biases the population toward upward current shifts.

Beyond this global trend, the programming routine systematically modulates the probability of upward transitions in a level- and voltage-dependent manner: at $V_{RD} = 0.3$ V, OSP tends to yield more soft-programming than ISPVA across most LRS levels, with a noticeable increase for low-/intermediate levels (LRS2–LRS4); at $V_{RD} = 0.4$ V, the OSP–ISPVA gap

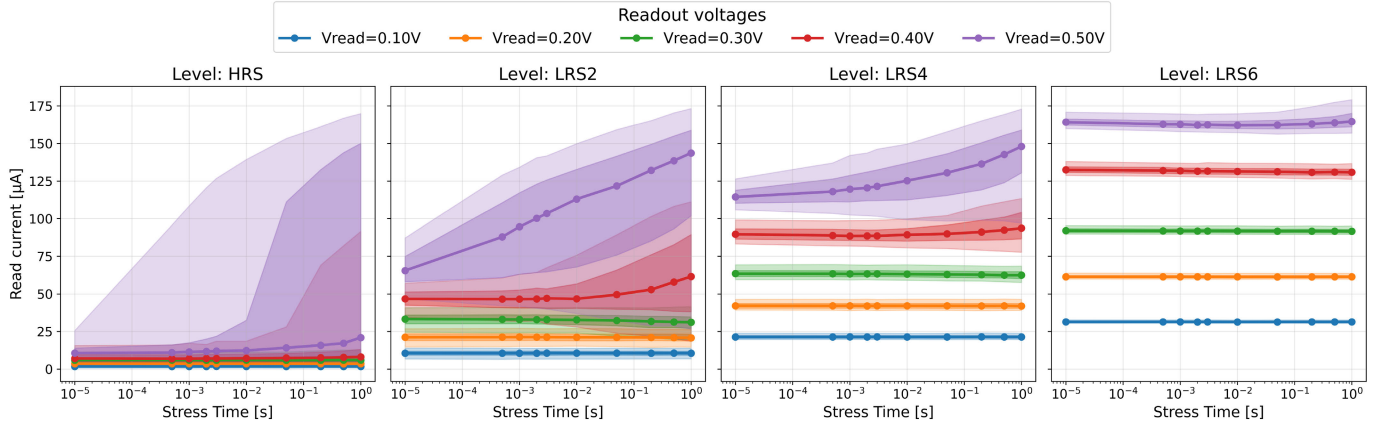


Fig. 4. Read-disturb summary (ISPVA): median, IQR, and 5th–95th percentiles vs. cumulative stress time for four representative levels and $V_{RD} = 0.1\text{--}0.5$ V.

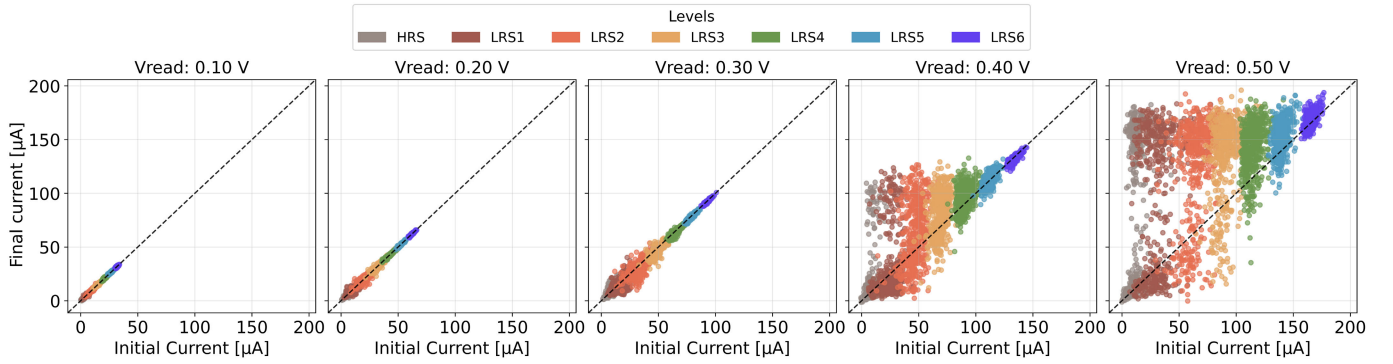


Fig. 5. ISPVA read-disturb: I_f vs. I_0 after 1 s cumulative stress for all levels and V_{RD} . Dashed line: $I_f = I_0$.

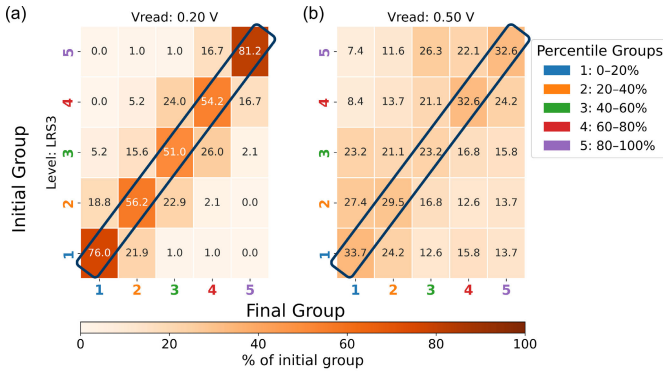


Fig. 6. Example of transition matrices for ISPVA-programmed cells (LRS3) under read-disturb. Each panel reports the bin-to-bin transition probability P_{ij} at nominal (0.2 V (a)) and highest (0.5 V (b)) read voltages V_{RD} .

becomes most pronounced in intermediate-to-high LRS levels (LRS4–LRS5); and at $V_{RD} = 0.5$ V, soft-programming is dominant for both schemes in most LRS levels, yet a noticeable residual excess for OSP persists in the highest level (LRS6). This comparison is made explicit in Fig. 8(b), which plots $\Delta = \text{SoftProg}_{\text{OSP}} - \text{SoftProg}_{\text{ISPVA}}$: $\Delta > 0$ indicates that OSP is more prone to read-induced soft-programming. Interestingly, HRS exhibits a different behavior, with Δ turning negative at higher V_{RD} , indicating a lower soft-programming incidence for OSP. This may suggest that the single pulse set/reset history

may yield a post-reset microstate that is less prone to partial reconnection under read stress (e.g., a more robust rupture/gap or a different defect/vacancy profile).

Overall, these observations show that, even under identical target currents and read-stress conditions, the programming algorithm shapes the susceptibility of each level to read-induced upward current shifts. This supports the hypothesis that the programming routine preconditions the distribution of filamentary microstates and thus provides a practical design knob to mitigate read-disturb.

V. RETENTION

In this section we analyze the evolution of the conductive levels under purely thermal stress. All experiments are performed on ISPVA-programmed cells.

A. Evolution of Multi-Level Distributions

Fig. 9 summarizes the evolution of the median, IQR and 5th–95th percentiles of the read current for seven levels at different temperatures and sampling times, from the end algorithm (EA) point up to 100 h. At RT, the median currents remain essentially constant over time for most LRS levels, while the IQR and the 5th–95th percentile bands show only a modest broadening, indicating limited drift within the time window explored.

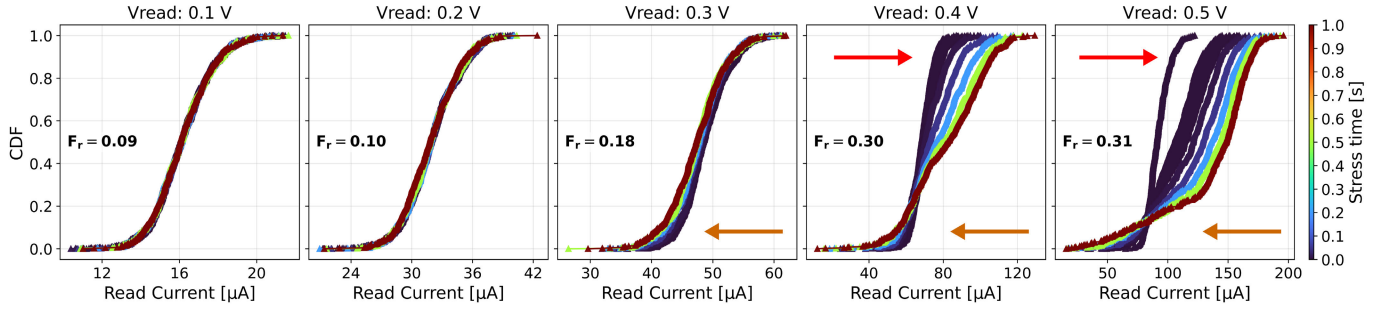


Fig. 7. Induced drift example (LRS3, ISPVA): CDFs at V_{RD} vs. cumulative stress time (0–1 s, 10^5 pulses). At 0.3 V the lower tail shifts downward; at higher V_{RD} both tails broaden (downward drift + soft-programming).

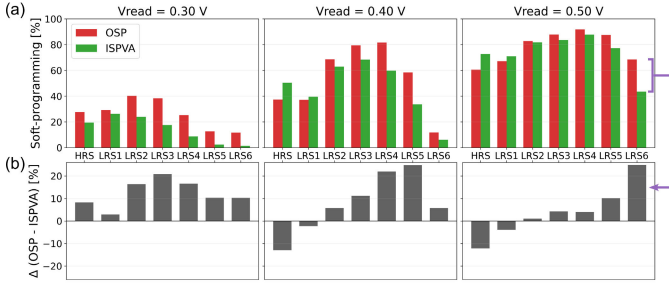


Fig. 8. Programming-algorithm impact on read-disturb directionality. (above) Soft-programming fraction (devices with $I_f > I_0$) for each level (HRS and LRS1–LRS6) after read-disturb at $V_{RD} = 0.30, 0.40$, and 0.50 V, comparing OSP (red) and ISPVA (green). (below) Corresponding difference $\Delta = \text{SoftProg}_{\text{OSP}} - \text{SoftProg}_{\text{ISPVA}}$ ($\Delta > 0$: higher soft-programming incidence for OSP).

As the temperature is increased to 85, 105, 125, and 150 °C, a clear temperature impact emerges. For each level, the median current decreases with time, with most of the drop occurring within the initial warm-up period of 30 min, during which the wafer reaches thermal equilibrium. Afterwards, the distributions evolve more gradually. The magnitude of this long-term drift between 0 h and 100 h depends on the programmed level: lower and intermediate LRS levels (LRS1–LRS3) exhibit a more pronounced decay of the median current, while the highest LRS levels (LRS5–LRS6) show a more robust behavior. At the same time, the 5th–95th percentile bands broaden with temperature, indicating that thermal stress also broadens the distributions.

To separate the fast temperature-induced change from the subsequent long-term drift, we quantify the initial distribution shift between EA at RT and the 0 h point (warm-up period) at each target temperature T . To estimate the warm-up shift, we use a percentile window that is empirically the least affected by tail deformation in our prior CDF-drift analysis [26]. Specifically, for each level L and temperature T , we average $\Delta I = I(0h, T) - I(EA, RT)$ over the 75th–100th percentile for LRS levels and over the 0th–25th percentile for HRS. This choice aims at capturing a predominantly rigid translation of the distribution while minimizing contamination from tail broadening. The resulting shift, reported in Fig. 10, becomes increasingly negative for LRS states as temperature rises, reflecting a fast reduction of the filament conductance as the device equilibrates at the new temperature. The absolute value

of this shift grows with both temperature and conductance level: more conductive LRS states suffer from a larger warm-up drop of current when the temperature is raised. Noticeably, HRS shows an opposite trend, with a small positive warm-up shift (higher read current) at elevated temperature.

Overall, these observations indicate that different LRS levels behave differently under thermal stress. Lower and intermediate levels (HRS–LRS4) experience a relatively modest initial distribution shift but a stronger subsequent drift over the 0–100 h window, whereas the highest LRS levels (LRS5–LRS6) are more affected by the fast distribution shift upon temperature increase but show a more limited long-term drift at fixed temperature.

B. Thermal Phases

To quantify how thermal stress impacts the internal ordering of devices, we evaluate the reshuffling factor F_r from the transition matrices between EA and the retention time points. Fig. 11 reports the average F_r versus time from EA to 100 h at each temperature, for high-resistance / low-LRS states (HRS and LRS1–LRS3) and for the most conductive states (LRS4–LRS6).

Two distinct phases emerge in both groups. First, a fast reshuffling takes place between EA and the 0 h point, i.e., during the 30 min warm-up required to reach thermal equilibrium at the target temperature. This initial increase of F_r is strongly temperature dependent: higher T systematically yields a larger reshuffling already at 0 h. This indicates that the temperature step does not only induce a rigid translation of the distributions (Fig. 10), but it also alters the internal rank ordering within each level, so that a non-negligible fraction of cells changes its percentile group immediately upon heating.

Second, once thermal equilibrium is reached, the further evolution from 0 h to 100 h proceeds much more gradually. For all temperatures, $F_r(t)$ continues to increase over time, yet with a much slower evolution rate, indicating that most of the temperature sensitivity of reshuffling is concentrated in the initial EA → 0 h transition.

Importantly, the reshuffling is not uniform across the programmed states. As highlighted by the two panels in Fig. 11, high-resistance / low-LRS states (HRS–LRS3) exhibit systematically larger reshuffling than the highest-conductance levels (LRS4–LRS6), especially during the warm-up phase at elevated temperature. This level dependence suggests that the

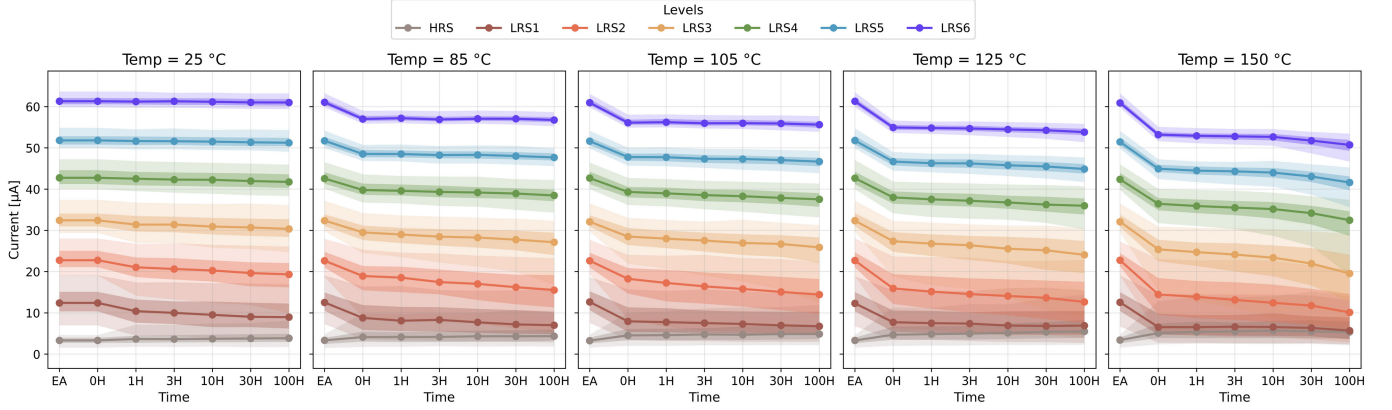


Fig. 9. Thermal readout current over time (ISPVA): median, IQR, and 5th–95th percentiles from EA to 100 h for seven levels at each temperature.

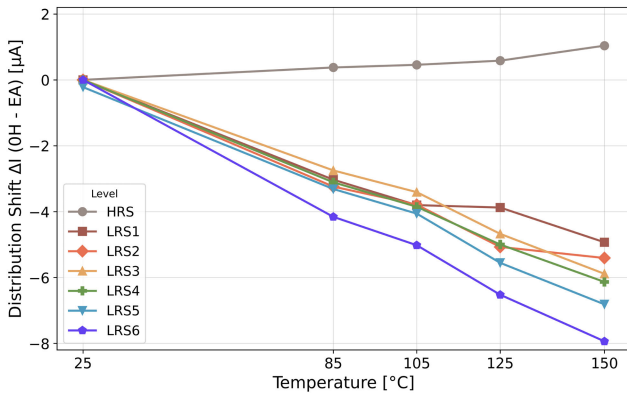


Fig. 10. Warm-up shift from EA (RT, ISPVA) to 0 h at T , extracted from drift-resistant regions (upper IQR for LRS, lower IQR for HRS).

stability of rank ordering under thermal stress is linked to filament robustness: weaker conductive paths are more susceptible to rapid thermally assisted rearrangements, whereas stronger conductive states better preserve their internal ordering.

Overall, these results show that thermal stress reshapes MLC levels through a dominant, temperature-activated warm-up reshuffling (EA $\rightarrow$ 0 h), followed by a weaker isothermal reshuffling (0 h $\rightarrow$ t).

VI. FIELD- VS THERMAL-INDUCED DRIFT

Read-disturb and thermal stress reshape MLC distributions through distinct redistribution patterns that are not fully captured by median and spread alone. Using intra-level transition matrices and the distance-weighted reshuffling factor F_r , we quantify how much the internal rank ordering of devices changes within each level, i.e., whether the same cells remain in the tails or whether tail membership is progressively reshuffled.

For read-disturb, the reshuffling increases strongly with the applied read voltage V_{RD} . As V_{RD} is raised from nominal conditions to the highest bias, the transition matrices progressively lose weight on the main diagonal and develop substantial off-diagonal probability, including long-range migrations between core and tail groups (Figs. 5 and 6). Consistently, the average F_r increases by about $\Delta F_r \approx 0.20$ between $V_{RD} = 0.2$ V and

0.5 V (Fig. 12), indicating a marked loss of rank correlation under repeated reads. This redistribution accompanies the coexistence of soft-programming events toward higher conductance and a concurrent tendency to drift toward lower currents starting at intermediate read voltages (induced drift), plausibly assisted by relaxation processes during read-pulse dissipation. Notably, even at low read voltages ($V_{RD} = 0.1$ – 0.2 V) F_r remains finite (average values around 0.1), indicating a measurably non-zero reshuffling even under nominal read voltages, although the resulting reshuffling is much smaller than at high V_{RD} .

For purely thermal stress, Fig. 11 shows that reshuffling has a dominant early contribution associated with the warm-up transient from EA to the 0 h point, which is strongly temperature dependent and more pronounced for low-/intermediate-conductance states (HRS and LRS1–LRS3) than for the most conductive states (LRS4–LRS6), indicating that large warm-up current shifts in high-LRS states (Fig. 10) do not necessarily translate into equally large rank reshuffling. After thermal equilibrium is reached, the subsequent evolution over the 0–100 h window is markedly slower. Accordingly, when retention reshuffling is evaluated over the steady isothermal window (from 0 h to 100 h), the average F_r shows only a mild temperature dependence up to 150 °C, with $\Delta F_r \approx 0.02$ between RT and 150 °C (Fig. 12). This suggests that once equilibrium is established, the internal ordering is largely preserved and the same devices tend to define the tails over time. At the same time, the retention data still exhibit significant temperature-dependent shifts and broadenings of the distributions (Figs. 9 and 10), indicating that limited reshuffling does not imply negligible drift of the absolute current values.

Overall, the comparison highlights two complementary reliability implications. Under read-disturb, reshuffling grows rapidly with V_{RD} and can involve long-range core-to-tail migration, calling for conservative read-voltage budgets and programming routines that improve filament robustness against both soft-programming and relaxation under repeated reads. Under thermal stress, the main temperature-driven reshuffling is concentrated in the warm-up transient (EA $\rightarrow$ 0 h), whereas the subsequent isothermal reshuffling is comparatively limited; therefore, improving long-term stability benefits from

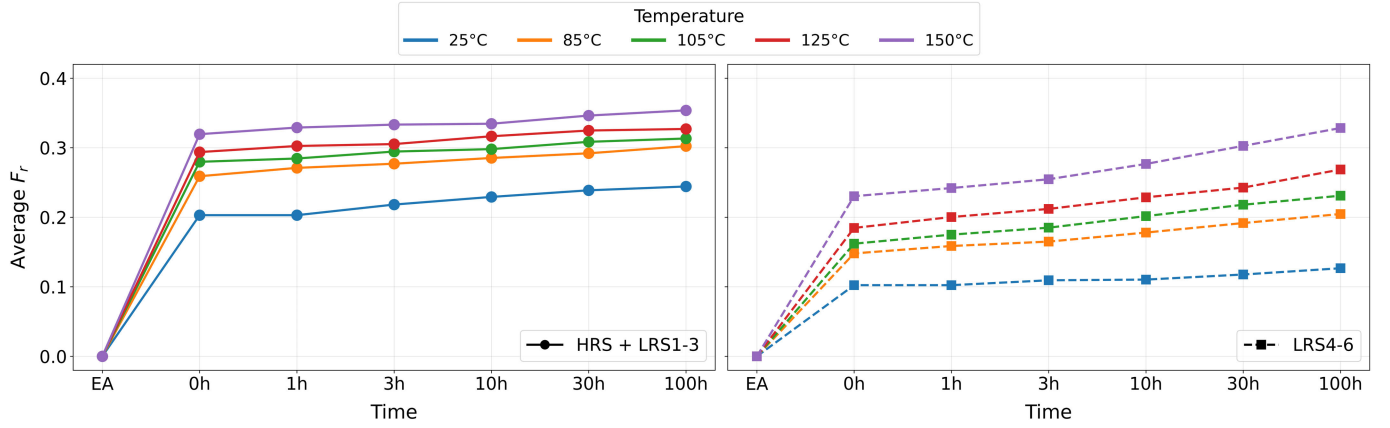


Fig. 11. Retention reshuffling dynamics under thermal stress (ISPVA). Average reshuffling factor F_r versus retention time from EA to 100 h for $T = 25, 85, 105, 125,$ and 150°C . Left: average F_r over the high-resistance / low-LRS states (HRS and LRS1–LRS3). Right: average F_r over the highest-conductance states (LRS4–LRS6). The 0 h point is measured after 30 min stabilization at the target temperature.

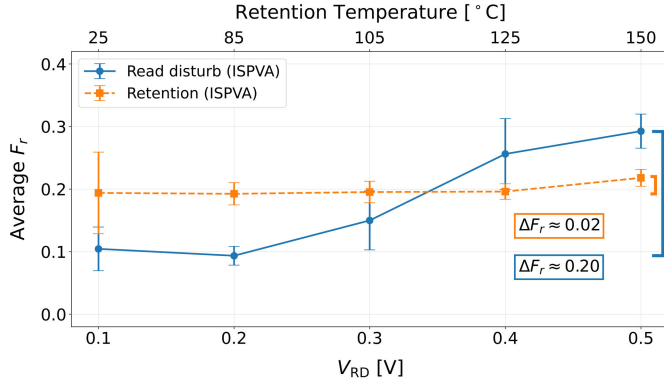


Fig. 12. Average reshuffling factor F_r for ISPVA-programmed cells under read-disturb (blue) and retention (orange) stress. The bottom axis reports the read voltage V_{RD} used during read-disturb, while the top axis indicates the retention temperatures over the isothermal window from thermal equilibrium (0 h) to 100 h. For each condition, F_r is averaged over all programmed levels and the error bars represent the standard deviation.

minimizing initial tails and accounting for warm-up at the system level (e.g., refresh or direct program at the operating temperature). In this framework, F_r and transition matrices provide a compact, application-relevant complement to conventional distribution metrics by explicitly tracking the evolution of tail membership in MLC RRAM arrays.

VII. CONCLUSION

In this work, we have experimentally investigated field- and temperature-induced drift in MLC Al5%:HfO₂ 1T1R RRAM devices. By introducing intra-level transition matrices and the distance-weighted reshuffling factor F_r , we distinguished distribution shifts from internal reshuffling and identified which devices populate the statistical tails under stress. Read-disturb experiments up to 10^5 reads (1 s cumulative pulse time) revealed a strong dependence of reshuffling on the read voltage V_{RD} . Transition matrices progressively develop off-diagonal probability and long-range core-to-tail migrations as V_{RD} increases, and the average reshuffling increases by about $\Delta F_r \approx 0.20$ between $V_{RD} = 0.2$ V and 0.5 V. These

trends accompany the coexistence of soft-programming events at high V_{RD} and a concurrent tendency toward downward redistribution at intermediate-to-high voltages (induced drift), highlighting the read-voltage budget as a primary reliability knob for MLC operation. The comparison between ISPVA and OSP showed that both algorithms experience a similar overall amount of read-induced redistribution, but with different drift directionality: ISPVA tends to favor more symmetric relaxation, while OSP exhibits a stronger bias toward soft-programming, with the most pronounced imbalance depending on the read voltage and LRS level. Multi-temperature retention experiments up to 150°C clarified that thermal reshuffling is dominated by an early warm-up contribution. The reshuffling factor exhibits a fast, temperature-dependent increase between EA and the 0 h point (thermal equilibrium), followed by a much slower evolution over the 0–100 h isothermal window. Once equilibrium is reached, the additional reshuffling remains comparatively limited, with only a mild temperature dependence up to 150°C (average $\Delta F_r \approx 0.02$ from RT to 150°C), despite significant temperature-dependent shifts and broadenings of the absolute current distributions. Overall, these results provide quantitative and physically interpretable guidelines to select safe read-voltage values, tailor programming schemes to improve filament robustness under repeated reads, and account for thermal warm-up transients at the system level to maximize the number of reliable levels in MLC RRAM arrays for storage and in-memory computing applications.

REFERENCES

- [1] G. W. Burr et al., “Experimental demonstration and tolerancing of a large-scale neural network (165 000 Synapses) using phase-change memory as the synaptic weight element,” *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3498–3507, Nov. 2015, doi: [10.1109/TED.2015.2439635](https://doi.org/10.1109/TED.2015.2439635).
- [2] M. A. Zidan, J. P. Strachan, and W. D. Lu, “The future of electronics based on memristive systems,” *Nature Electron.*, vol. 1, no. 1, pp. 22–29, Jan. 2018, doi: [10.1038/s41928-017-0006-8](https://doi.org/10.1038/s41928-017-0006-8).
- [3] W. He, J. Meng, S. K. Gonugondla, S. Yu, N. R. Shanbhag, and J.-S. Seo, “PRIVE: Efficient RRAM programming with chip verification for RRAM-based in-memory computing acceleration,” in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Apr. 2023, pp. 1–6, doi: [10.23919/DAT56975.2023.10137266](https://doi.org/10.23919/DAT56975.2023.10137266).

- [4] K. Nikiruy et al., "Blooming and pruning: Learning from mistakes with memristive synapses," *Sci. Rep.*, vol. 14, no. 1, p. 7802, Apr. 2024, doi: [10.1038/s41598-024-57660-4](https://doi.org/10.1038/s41598-024-57660-4).
- [5] M. Hu et al., "Memristor-based analog computation and neural network classification with a dot product engine," *Adv. Mater.*, vol. 30, no. 9, Mar. 2018, Art. no. 1705914, doi: [10.1002/adma.201705914](https://doi.org/10.1002/adma.201705914).
- [6] S. Yu, "Neuro-inspired computing with emerging nonvolatile memories," *Proc. IEEE*, vol. 106, no. 2, pp. 260–285, Feb. 2018, doi: [10.1109/JPROC.2018.2790840](https://doi.org/10.1109/JPROC.2018.2790840).
- [7] H. Bao et al., "Toward memristive in-memory computing: Principles and applications," *Frontiers Optoelectronics*, vol. 15, no. 1, p. 23, Dec. 2022, doi: [10.1007/s12200-022-00025-4](https://doi.org/10.1007/s12200-022-00025-4).
- [8] E. P. Quesada et al., "Experimental assessment of multilevel RRAM-based vector-matrix multiplication operations for in-memory computing," *IEEE Trans. Electron Devices*, vol. 70, no. 4, pp. 2009–2014, Apr. 2023, doi: [10.1109/TED.2023.3244509](https://doi.org/10.1109/TED.2023.3244509).
- [9] A. Fantini et al., "Intrinsic switching variability in HfO₂ RRAM," in *Proc. 5th IEEE Int. Memory Workshop*, May 2013, pp. 30–33, doi: [10.1109/IMW.2013.6582090](https://doi.org/10.1109/IMW.2013.6582090).
- [10] A. Grossi et al., "Fundamental variability limits of filament-based RRAM," in *IEDM Tech. Dig.*, Dec. 2016, pp. 4.7.1–4.7.4, doi: [10.1109/IEDM.2016.7838348](https://doi.org/10.1109/IEDM.2016.7838348).
- [11] E. Perez, M. K. Mahadevaiah, E. P.-B. Quesada, and C. Wenger, "Variability and energy consumption tradeoffs in multilevel programming of RRAM arrays," *IEEE Trans. Electron Devices*, vol. 68, no. 6, pp. 2693–2698, Jun. 2021, doi: [10.1109/TED.2021.3072868](https://doi.org/10.1109/TED.2021.3072868).
- [12] M. Zhao et al., "Investigation of statistical retention of filamentary analog RRAM for neuromorphic computing," in *IEDM Tech. Dig.*, Dec. 2017, pp. 39.4.1–39.4.4.
- [13] Y.-H. Lin et al., "Performance impacts of analog ReRAM non-ideality on neuromorphic computing," *IEEE Trans. Electron Devices*, vol. 66, no. 3, pp. 1289–1295, Mar. 2019, doi: [10.1109/TED.2019.2894273](https://doi.org/10.1109/TED.2019.2894273).
- [14] W. Shim, Y. Luo, J.-S. Seo, and S. Yu, "Investigation of read disturb and bipolar read scheme on multilevel RRAM-based deep learning inference engine," *IEEE Trans. Electron Devices*, vol. 67, no. 6, pp. 2318–2323, Jun. 2020.
- [15] J. Wen, A. Baroni, E. Perez, M. Ulbricht, C. Wenger, and M. Krstic, "Evaluating read disturb effect on RRAM based AI accelerator with multilevel states and input voltages," in *Proc. IEEE Int. Symp. Defect Fault Tolerance VLSI Nanotechnol. Syst. (DFT)*, Oct. 2022, pp. 1–6.
- [16] Z. Yu et al., "Early-stage fluctuation in low-power analog resistive memory: Impacts on neural network and mitigation approach," *IEEE Electron Device Lett.*, vol. 41, no. 6, pp. 940–943, Jun. 2020, doi: [10.1109/LED.2020.2986889](https://doi.org/10.1109/LED.2020.2986889).
- [17] V. Milo et al., "Accurate program/verify schemes of resistive switching memory (RRAM) for in-memory neural network circuits," *IEEE Trans. Electron Devices*, vol. 68, no. 8, pp. 3832–3837, Aug. 2021, doi: [10.1109/TED.2021.3089995](https://doi.org/10.1109/TED.2021.3089995).
- [18] Y. Huang, R. Hopkins, D. Janosky, Y.-C. Chen, Y.-F. Chang, and J. C. Lee, "Effect of temperature on analog memristor in neuromorphic computing," *IEEE Trans. Electron Devices*, vol. 69, no. 11, pp. 6102–6105, Nov. 2022, doi: [10.1109/TED.2022.3207710](https://doi.org/10.1109/TED.2022.3207710).
- [19] A. Baroni et al., "Low conductance state drift characterization and mitigation in resistive switching memories (RRAM) for artificial neural networks," *IEEE Trans. Device Mater. Rel.*, vol. 22, no. 3, pp. 340–347, Sep. 2022, doi: [10.1109/TDMR.2022.3182133](https://doi.org/10.1109/TDMR.2022.3182133).
- [20] A. Fantini et al., "Engineering of Hf_{1-x}Al_xO_y amorphous dielectrics for high-performance RRAM applications," in *Proc. IEEE 6th Int. Memory Workshop (IMW)*, May 2014, pp. 1–4, doi: [10.1109/IMW.2014.6849354](https://doi.org/10.1109/IMW.2014.6849354).
- [21] B. Traoré et al., "On the origin of low-resistance state retention failure in HfO₂-based RRAM and impact of doping/alloying," *IEEE Trans. Electron Devices*, vol. 62, no. 12, pp. 4029–4036, Dec. 2015, doi: [10.1109/TED.2015.2490545](https://doi.org/10.1109/TED.2015.2490545).
- [22] K. Lee, Y. Kim, H. Na, and H. Sohn, "Effect of aliovalent impurities on the resistance switching characteristics of sputtered hafnium oxide films," *J. Vac. Sci. Technol. B, Nanotechnol. Microelectron., Mater., Process., Meas., Phenomena*, vol. 33, no. 3, May 2015, Art. no. 032204, doi: [10.1116/1.4917495](https://doi.org/10.1116/1.4917495).
- [23] E. Pérez, A. Grossi, C. Zambelli, P. Olivo, R. Roelofs, and C. Wenger, "Reduction of the cell-to-cell variability in Hf_{1-x}Al_xO_y based RRAM arrays by using program algorithms," *IEEE Electron Device Lett.*, vol. 38, no. 2, pp. 175–178, Feb. 2017, doi: [10.1109/LED.2016.2646758](https://doi.org/10.1109/LED.2016.2646758).
- [24] S. Roy et al., "Toward a reliable synaptic simulation using Al-doped HfO₂ RRAM," *ACS Appl. Mater. Interface*, vol. 12, no. 9, pp. 10648–10656, Mar. 2020, doi: [10.1021/acsami.9b21530](https://doi.org/10.1021/acsami.9b21530).
- [25] H. He, X. Yuan, W. Wu, C. Lee, Y. Zhao, and Z. Liu, "Impact of Al alloying/doping on the performance optimization of HfO₂-based RRAM," *Electronics*, vol. 13, no. 12, p. 2384, Jun. 2024, doi: [10.3390/electronics13122384](https://doi.org/10.3390/electronics13122384).
- [26] A. Baroni et al., "Enhancing RRAM reliability: Exploring the effects of Al doping on HfO₂-based devices," *IEEE Trans. Device Mater. Rel.*, vol. 25, no. 3, pp. 379–387, Sep. 2025, doi: [10.1109/TDMR.2025.3581061](https://doi.org/10.1109/TDMR.2025.3581061).
- [27] C. Bengel et al., "Reliability aspects of binary vector-matrix-multiplications using ReRAM devices," *Neuromorphic Comput. Eng.*, vol. 2, no. 3, Jun. 2022, Art. no. 034001.
- [28] K. Dorai Swamy Reddy et al., "Optimization of technology processes for enhanced CMOS-integrated 1T-1R RRAM device performance," *Eur. Phys. J. B*, vol. 97, no. 11, p. 181, Nov. 2024, doi: [10.1140/epjb/s10051-024-00821-1](https://doi.org/10.1140/epjb/s10051-024-00821-1).
- [29] A. Baroni et al., "Comparing short and long-term reliability of HfO₂ and Al:HfO₂ RRAM devices," in *Proc. IEEE Int. Integr. Rel. Workshop (IIRW)*, Oct. 2024, pp. 1–6.
- [30] E. Pérez, C. Zambelli, M. K. Mahadevaiah, P. Olivo, and C. Wenger, "Toward reliable multi-level operation in RRAM arrays: Improving post-algorithm stability and assessing endurance/data retention," *IEEE J. Electron Devices Soc.*, vol. 7, pp. 740–747, 2019, doi: [10.1109/JEDS.2019.2931769](https://doi.org/10.1109/JEDS.2019.2931769).
- [31] E. Pérez, M. K. Mahadevaiah, C. Zambelli, P. Olivo, and C. Wenger, "Characterization of the interface-driven 1st reset operation in HfO₂-based 1T1R RRAM devices," *Solid-State Electron.*, vol. 159, pp. 51–56, Sep. 2019, doi: [10.1016/j.sse.2019.03.054](https://doi.org/10.1016/j.sse.2019.03.054).