Abstract—Stochastic behaviors of resistive random access memory (RRAM) play an important role in the design of cross-point memory arrays. A Monte Carlo compact model of oxide RRAM is developed and calibrated with experiments on various device stack configurations. With Monte Carlo SPICE simulations, we show that an increase in array size and interconnect wire resistance will statistically deteriorate write functionality. Write failure probability (WFP) has an exponential dependency on device uniformity and supply voltage ($V_{DD}$), and the array bias scheme is a key knob. Lowering array $V_{DD}$ leads to higher effective energy consumption (EEC) due to the increase in WFP when the variation statistics are included in the analysis. Finally, we show that a pseudo-sub-array topology with uniformly distributed pre-forming cells in the pristine high resistance state is able to reduce both WFP and EEC, enabling higher net capacity for memory circuits due to improved variation tolerance.

Index Terms—Resistive random access memory (RRAM, ReRAM), non-volatile memory, variability, statistical analysis, Monte Carlo, compact model, cross-point array, SPICE.

I. INTRODUCTION

RESISTIVE random access memory (RRAM) is a promising building block of nonvolatile information storage systems for data-centric applications [1]-[4], leveraging its simple structure, high performance, and good scalability [5]-[12]. Variability of RRAM characteristics imposes constraints on memory cell/array design, but also inspires new applications such as stochastic neuromorphic systems and physical unclonable function [13]-[15]. Hence, a deeper understanding of the device and circuit variability is essential to architect RRAM for memory and logic applications. Previous research efforts were mainly focused on the intrinsic device-level variability and corresponding physical mechanisms, which were correlated with the choice for the materials, stack configuration, as well as operation conditions [16]-[21]. Array analysis of the variability impact on read performance was performed using simplified resistor models [22], [23]. However, cross-layer study that links the physical picture of intrinsic device variability with circuit-level stochastic behaviors is still lacking. In this work, we scrutinize the stochastic behaviors of cross-point RRAM arrays from the perspective of device and circuit interaction. Large-scale SPICE simulations are performed using a Monte Carlo (MC) compact model of RRAM, which is built upon a stochastic conductive filament (CF) evolution model and calibrated by device measurements. Through a suite of variation-aware circuit analysis, we probe into the array write functionality, reliability, and performance in a statistical manner. The ‘translation’ of variations from device level into circuit level is neither linear additive nor analytical, and larger arrays tend to ‘amplify’ the circuit variations translated from device tails, especially for low-power embedded applications. This key observation renders the methodology in this work essential for variation-aware RRAM array optimization.

II. MONTE CARLO COMPACT MODEL

The variability of CF geometry leads to the widely observed random variations of high resistance states (HRS) and low resistance states (LRS) [16]. Specifically, it has been shown that the resistance distributions of LRS ($R_{LRS}$) and HRS ($R_{HRS}$) result from the fluctuations in the CF radius and the tunneling...
The complexity of the device is determined by physical variability, which can be incorporated into a compact model to efficiently model the variation in a compact model. Developing a method to incorporate the variations in a compact model is even more critical given the complexity of some of the in-memory computing circuits being designed [24]-[26]. Fig. 1 shows the model hierarchy, from physics of variability to Verilog-A sub-circuit implementation of the compact model. Physical variability is described by the stochastic location of oxygen vacancies around the CF and the gap region. The variations of electrical characteristics of RRAM are then described by the intrinsic variability of CF geometry in terms of length and width. In the Verilog-A implementation, the median CF evolution is modeled by the hopping/ohmic paths as a deterministic sub-circuit [27], while the stochastic CF geometry is modeled by a Monte Carlo (MC) resistor as a variation sub-circuit. For LRS, the MC resistor and deterministic CF part are in parallel, modeling the CF width variations. For HRS, the MC resistor is in series with the deterministic CF part, modeling the CF length or gap distance variations. The dominant equations that describe the deterministic CF evolution behaviors follow the same set as those in an experimentally calibrated compact model [27]-[29], and the variation sub-circuit is calibrated using the statistical data from measurements. The MC compact model captures the essential variations in an efficient manner, while retaining the physics and accuracy of the deterministic part. The generality of MC variation sub-circuit representation is confirmed by the measurements on a variety of RRAM devices, as shown in Fig. 2. The measured devices include (a) HfO2 RRAM, (b) HfO2/TiO2 bi-layer RRAM, (c) HfO2/TiO2/HfO2/TiO2 multi-layer RRAM [30], (d) Ta2O5/Ta2O5, RRAM, (e) HfO2, 3D vertical RRAM [31], and (f) HfO2/AlOx, 3D vertical RRAM. The fitting procedure for each type of device is as follows: statistical distributions of $R_{LRS}$ and $R_{HRS}$ are first obtained by 100-cycle SET/RESET measurements, with median and standard deviation (SD) extracted assuming Gaussian distributions. The deterministic sub-circuit in the compact model captures the median values of measured $R_{LRS}$ and $R_{HRS}$ following the principle in [28]. To add the variations, the variation sub-circuit, i.e. the MC resistor, takes the extracted SD values as an input parameter, and generates normal distributions through statistical MC simulations in HSPICE [33]. Note that the distributions in (a)-(f) present different curvatures and tails, in both experimental and modeled data. For experiments, the statistical distribution results from the intrinsic randomness of CF position, size, and even morphology. The measured distributions may keep varying with more cycles involved (if insufficient samples are measured initially, or if some irreversible changes in the RRAM occur, such as endurance cycling degradation). For modeling, HSPICE uses non-identical random seeds to generate specified distributions from device to device, which mirrors such dynamics of tails and curvatures observed in our measurements. This model formulation enables the compact model to keep its analytical nature for modeling both deterministic and stochastic behaviors efficiently, which further enables us to conduct the array-level circuit/device interaction analysis. A case study using the device characteristics in Fig. 2(c) is performed and discussed in

![Fig. 3. Schematic of a cross-point RRAM array and the corresponding framework for device and circuit interaction analysis (SWL: selected word line; UWL: unselected word line; SBL: selected bit line; UBL: unselected bit line; WAV: write access voltage; WFP: write failure probability).](image-url)
III. STOCHASTIC BEHAVIORS IN RRAM ARRAYS

To illustrate how RRAM variations are ‘translated’ into the circuit-level stochastic behaviors, non-cross-point array structure without selection devices is used as an example. Fig. 3 illustrates the analysis framework incorporating the measured/modelled RRAM variations. In the following simulations, the technology node is assumed to be 22 nm for cross-point arrays. Metal wires have a 44-nm pitch, an aspect ratio (AR) of ~2, and a sheet resistance of 1.405 Ω/square. The capacitance of the wires is 1.045 fF/μm. Therefore, in the array model, wire resistance (Rwire) is 2.81 Ω/cell and wire capacitance (Cwire) is 0.046 fF/cell [34]. RRAM MC model is calibrated to experimental measurements of the device in Fig. 2(c). A worst-case scenario is adopted, where the farthest-corner selected cell is in HRS and all the unselected cells are in LRS [35]. V/2 scheme and V/3 scheme, the two major bias schemes for cross-point RRAM arrays [36], are used for write operations:

V/2:  \[ V_{\text{V/2}} = V_{\text{V/2 low}} = V_{\text{V/2 high}} = \frac{V_{\text{DD}}}{2} \]  

V/3:  \[ V_{\text{V/3}} = V_{\text{V/3 low}} = V_{\text{V/3 high}} = \frac{V_{\text{DD}}}{3}, V_{\text{V/3 mid}} = 2 \left( \frac{V_{\text{DD}}}{3} \right). \]

With 1000-cycle MC simulations in HSPICE for each device/circuit configuration, three aspects of stochastic behaviors in cross-point RRAM arrays are investigated: write functionality, reliability, and energy consumption.

A. Write Functionality

Write access voltage (WAV) is defined as the actual voltage drop on the selected cell in a memory array, which is lower than the array voltage supply VDD due to interconnect IR drop [29], [35], [36]. Fig. 4 shows the variation-aware assessment of cross-point RRAM arrays. Statistical distributions of WAV for various array sizes are obtained by simulating write operations under the aforementioned V/2 scheme (VDD = 2.6 V), as shown in Fig. 4(a). The WAV distributions result directly from the device-level resistance variations. Correspondingly, the difference in specific resistance patterns due to device resistance variations are ‘translated’ to differences in sneak path configurations and IR drop along interconnect wires. Thereby, WAV diverges from its nominal value. Fig. 4(a) shows that WAV for 4-kb array has a tight distribution and a higher median value, whereas the distribution tends to spread out in larger arrays. This is mainly because the possible number of configurations of either sneak paths or data patterns grows exponentially with array size. This observation, on the other hand, implies that the ‘translation’ of variations from device measurables (e.g., R_LRS and R_HRS) to circuit outputs (e.g., WAV) is nonlinear. Therefore, MC simulation framework is necessary to help visualize and quantify such translation and interaction.

Fig. 4(b) shows that using V/3 scheme results in relatively higher median WAV but worse distributions compared with V/2 scheme. This illustrates the dominated role of unselected cells (cross-points of all the unselected WLs and BLs) in WAV distributions for the in the V/3 scheme write operation. In V/3 scheme, all the unselected cells have biases near VDD/3. In comparison, the unselected cells in V/2 scheme have near-zero biases, which reduces the impact of stochastic configurations of sneak paths on WAV distributions. Such difference between V/3 and V/2 schemes becomes more significant in larger arrays, as indicated by Fig. 4(c). In addition to the scaling up effect of array sizes, the impact of interconnect scaling down is quantified as well. Fig. 5(a) shows the statistical distributions of WAV as interconnect wire resistance increases due to scaling-induced scattering and skin effects [34], [37]. The median shift with the increase in wire resistivity is attributed to the degradation of IR drop along the selected signal path. SD of the WAV is extracted and normalized, as shown in Fig. 5(b). The largest Rwire does not necessary correspond to the widest distributions. This trend is in line with the median’s shift towards lower write margin shown in Fig. 5(a). Specifically, the variations of RRAM cells become less important as Rwire increases with larger IR drop, and thus the WAV distributions tend to become tighter (as Rwire variation is not included in the model).
Write Failure Probability (%)

RLRS Variation Coefficient ($\sigma/\mu$)

Array Size (bit)

V/2 scheme
V/3 scheme

-10
10
0.1
1
10

Effective Energy (J)

Array Size (bit)

2.5×
V/3 scheme
V/2 scheme

1-kb array
4-kb array
16-kb array

Cumulative Probability (%)

Resistance (Ω)

-1 k
4 k
16 k

CDF (%)

Energy Consumption (J)

10
-11
10
-10
10
-9
10
-8

50
90
98
99.5

Average energy
Effective energy

Effective energy consumption as a function of array size and bias scheme.

V/3 scheme
V/2 scheme

10.4×

Effective energy consumption increases exponentially under V/2 and V/3 schemes as array $V_{DD}$ decreases.

Besides the write functionality discussed above, array reliability in terms of statistical write failure behaviors is also highly dependent on the interaction between device variations and circuit conditions. Write failure probability (WFP) is quantified by simulating SET operations to analyze such dependency in cross-point RRAM arrays. For other data patterns, WFP analysis of RESET operations follows the same principles as SET operations. A write failure event occurs when the access voltage is not sufficient to switch (i.e., SET/RESET) the selected RRAM cell. And WFP is defined as the ratio of failed write events to total write events. As shown in Fig. 6(a), the WFP induced by device-level variations generally increases with larger array sizes. V/3 scheme has lower WFP compared with V/2 scheme. This is mainly due to higher median WAV in the V/3 scheme even though the V/3 scheme has relatively higher degree of variance. Such higher variance also leads to a rapid increase in WFP as array size scales up. High median voltage stress in V/3 scheme might cause earlier endurance failure [38], whereas the low-median low-variance V/2 scheme could lead to better endurance performance. There are tradeoffs between different reliability metrics such as WFP and endurance.

C. Energy Consumption

Energy efficiency is one of the major design objectives for architecting non-volatile memory sub-systems [1]. A deeper analysis from the statistical perspective is essential but rarely provided. Using MC simulations in HSPICE, the total energy consumption during write operations is obtained and analyzed. As shown in Fig. 7(a), device variations also result in a significant spread of energy consumption at the memory array level, which can lead to substantial degradation in energy.
efficiency. For instance, in 1-kb arrays biased under V/2 scheme, a 38% deviation below nominal $R_{LRS}$ of unselected cells results in a 45× deviation above nominal array energy consumption. Compared with the V/2 scheme, the V/3 scheme shows an order of magnitude higher energy consumption on average, mainly due to more sneak paths among unselected cells. The leakage current through these unselected cells with near V$_{DD}$/3 bias contributes substantial static energy consumption. Since the major variation source for these MC simulations is the worst-case array resistance pattern with cell resistance variability, the tails in energy consumption distributions correspond to the outliers of static energy due to unselected cells with abnormally low resistances. A missing piece in the analysis above is the write failure statistics. Statistically speaking, the energy consumption during failed write events is wasted. WFP should be taken into account to provide a metric for an effective energy consumption (EEC), which can be defined as:

$$ EEC = \frac{Average\ Eenergy}{1 - WFP} \cdot (J/\text{op}) \cdot (3) $$

Fig. 7(b) compares the average energy and EEC under V/2 scheme for a range of array V$_{DD}$. The average energy consumption decreases with V$_{DD}$ as expected, whereas the EEC first drops and then rises. This is attributed to the low-V$_{DD}$-induced write failure. EEC can be roughly thought of the actual ‘work’ required for each successful write operation. Even though the specific values of such ‘work’ may change if error correction, adaptive write, or other techniques are used in the peripheral circuits, the key point of the analysis here still holds: as long as there exist failure events induced by intrinsic variability, there will be extra energy overhead. Fig. 7(c) shows the comparison between V/2 scheme and V/3 scheme in terms of EEC. It is shown that V/2 scheme is more energy efficient with an order of magnitude lower EEC than V/3 scheme. This is a race condition where V/2 scheme has relatively higher WFP but much lower leakage-induced static energy. The net effect of such interaction leads to lower EEC under V/2 scheme, again indicating the predominant role of unselected cells in overall energy efficiency. As array size increases from 1 kb to 16 kb, the leakage paths under V/3 scheme result in even larger spread of energy consumption and the rapid increase in WFP, which is reflected by the >10x increase in EEC.

Fig. 8 Case study of a variation-tolerant pseudo-sub-array (PSA) topology for cross-point RRAM arrays. (a) Schematic of PSA topology where pre-forming cells are uniformly distributed. (b) Comparison of PSA topology and full-size cross-point array in terms of WFP (array size is 64 kb and V$_{DD}$ is 4.0 V). PSA topology significantly lowers the WFP by restraining array-level variations induced by sneak paths. (c) Comparison of PSA topology and full-size cross-point array in terms of EEC. PSA topology enables higher energy efficiency due to improved write reliability and reduction of array leakage current. (d) Allowed memory net capacity under conventional full-size and PSA topologies given the fixed 1% WFP.

IV. OPTIMIZATION: CASE STUDY & GUIDELINES

Based on the understanding of the stochastic behaviors in cross-point RRAM arrays, a case study for array optimization is performed. It was shown that uniformly distributed insulating (i.e., pre-forming high resistance) cells in RRAM arrays are able to help restrain sneak paths and thus reduce leakage current [41]. Here, we consider such design’s tolerance to variations. Fig. 8(a) shows the schematic of the pseudo-sub-array (PSA) topology, which is inspired by the partition of sub-arrays in a memory bank [42]. Since the working cell distribution is uniform in both WL and BL directions, memory sub-system design does not require fundamental changes except for the address encoding/decoding. Write operations on 64-kb arrays with 4.0-V V$_{DD}$ are simulated. As shown in Fig. 8(b), 4.0 V is insufficient for the write operations on full-size 64-kb arrays, with very high WFP. In contrast, the optimized PSA topology has significantly reduced WFP under the same V$_{DD}$ and bias schemes for 64-kb arrays. This is mainly because the uniform pattern of insulating cells effectively constrains the possible number of sneak path configurations, and thereby reduces the variations that are related to leakage current. Energy efficiency is also analyzed. As shown in Fig. 8(c), PSA topology reduces the EEC of 64-kb arrays by ~30× under V/2 scheme and ~33× under V/3 scheme. Improvement in EEC is a collaborative effect of lowered WFP and reduced static energy consumption. For cross-point RRAM arrays, the maximum allowed array size is usually limited by...
the write/read margin and write reliability of accessing individual cells [35]. A large array means large IR drop and reduced margin for WAV, leading to higher WFP. Variation-tolerant PSA topology is able to unlock the key limit factors of cross-point RRAM arrays, and therefore, may enable higher net capacity even after considering the non-programmable (unused) cells. Fig. 8(c) shows the allowed net capacity under conventional full-size and PSA topologies. VDD is fixed as 4.0 V and the array sizes are screened to hit the target 1% WFP for each topology. The net capacity for PSA topology includes the capacity loss due to non-programmable cells, yet a higher net capacity is obtained over conventional design due to the significantly improved variation tolerance.

Optimization of RRAM-based memory systems relies on the understanding of device and circuit interaction with an emphasis on variations. Fig. 9 summarizes the cross-layer design efforts discussed throughout this paper, which can also serve as guidelines for variation-aware RRAM optimization. Due to a collaborative effect of device properties, circuit topologies, and operation conditions, device-level variations are nonlinearly 'translated' into circuit-level stochastic behaviors. These circuit behaviors are characterized by the statistics related to functionality, reliability, and energy. A full understanding of how device variability would affect these metrics will provide insights to guide device engineering. Holistically, the optimization space lies in the knowledge-based device tuning, architecture optimization, and tradeoffs between different metrics by circuit design knobs.

V. CONCLUSION

Using a MC compact model of the RRAM, the device and circuit interaction analysis reveals the trend, dependency, and main contributors of stochastic behaviors in cross-point RRAM arrays. Scaling-up of array size, scaling-down of interconnect and VDD, and choice of bias schemes, play dominant roles in array functionality, reliability, and energy consumption. V/3 scheme has lower WFP, but is also more sensitive to increased device variations. V/2 scheme is more energy efficient from a statistical perspective. More generally, for V/n bias schemes, a larger n gives a lower WFP at the cost of worse energy efficiency and sensitivity to device variations. Design implications inspired by the circuit behaviors and metrics tradeoffs are discussed throughout the paper and exemplified in the case study of array topology optimization. This work provides a new optimization methodology for RRAM memory systems that require awareness of variations across all design layers.

ACKNOWLEDGEMENT

The authors would like to thank Z. Fang from IME, A*STAR, Singapore, H.-Y. Chen from Stanford University, and Z. Chen from Peking University, China, for providing RRAM samples used in this work.

REFERENCES

[1] H.-S. P. Wong and S. Salahuddin, “Memory leads the way to better computing,” Nature Nanotech., vol. 10, no. 3, pp. 191–194, March 2015.

[2] M. M. S. Aly et al., “Energy-efficient abundant-data computing: the N3XT 1,000 x...,” IEEE Computer, vol. 48, no. 12, pp. 24–33, Dec. 2015.

[3] J. Yang, D. B. Strakov, and D. R. Stewart, “Memristive devices for computing,” Nature Nanotech., vol. 8, no. 1, pp. 13–24, Jan. 2013.

[4] H. Li et al., “A learnable parallel processing architecture towards unity of memory and computing,” Sci. Rep., vol. 5, Art. ID 13330, Aug. 2015.

[5] R. Waser et al., “Redox-based resistive switching memories-nanoionic mechanisms, prospects, and challenges,” Adv. Mater., vol. 21, no. 25-26, pp. 2632-2663, July 2009.

[6] I. G. Baek et al., “Realization of vertical resistive memory (VRRAM) using cost effective 3D process,” in Proc. IEEE IEDM, Dec. 2011, pp. 737-740.

[7] M.-J. Lee et al., “A fast, high-endurance and scalable non-volatile memory device made from asymmetric TaOx/TaOx bilayer structures,” Nature Mater., vol. 10, pp. 625–630, Aug. 2011.

[8] A. Kawahara et al., “An 8Mb multi-layered cross-point ReRAM macro with 443MB/s write throughput,” in IEEE ISSTC Dig. Papers, Feb. 2012, pp. 432-434.

[9] T.-Y. Liu et al., “A 130.7 mm² 2-layer 32 Gb ReRAM memory device in 24 nm technology,” in IEEE ISSTC Dig. Papers, Feb. 2013, pp. 210–211.

[10] S. Yu et al., “Stochastic learning in oxide binary synaptic device for neuromorphic computing,” Frontiers in Neuroscience, vol. 7, no. 186, pp. 1–9, Oct. 2013.

[11] M. Saiti et al., “Bio-inspired stochastic computing using binary CBRAM synapses,” IEEE Trans. Electron Devices, vol. 60, no. 7, pp. 2402-2409, July 2013.

[12] A. Chen, “Utilizing the Variability of resistive random access memory to implement reconfigurable physical unclonable functions,” IEEE Electron Device Lett., vol. 36, no. 2, pp. 138-140, Feb. 2015.

[13] S. Dong et al., “Cell variability impact on the one-shot data set learning of machine learning algorithms,” IEEE Trans. Electron Devices, vol. 61, no. 8, pp. 2912-2919, Aug. 2014.

[14] S. Long et al., “Cycle-to-cycle intrinsic reset statistics in HIO-based unipolar RRAM devices,” IEEE Electron Device Lett., vol. 34, no. 5, pp. 643-645, May 2013.
[28] P. Huang et al., “A physics-based compact model of metal-oxide-based RRAM DC and AC operations,” IEEE Trans. Electron Devices, vol. 60, no. 12, pp. 4090–4097, Dec. 2013.

[29] H. Li et al., “Variation-aware, reliability-emphasized design and optimization of RRAM using SPICE model,” in Proc. DATE, March 2015, pp. 1425-1430.

[30] Z. Fang et al., “HfO2/TiO2/HfO2/TiO2 multilayer-based forming-free RRAM devices with excellent uniformity,” IEEE Electron Device Lett., vol. 32, no. 4, pp. 566–568, Apr. 2011.

[31] H.-Y. Chen et al., “HfO2 based vertical RRAM for cost-effective 3D cross-point architecture without cell selector,” in Proc. IEEE IEDM, Dec. 2012, pp. 497–500.

[32] B. Gao et al., “3-D cross-point array operation on AlO/HfO2-based vertical resistive switching memory,” IEEE Trans. Electron Devices, vol. 61, no. 5, pp. 1377-1381, May 2014.

[33] HSPICE Reference Manual, Synopsys Inc., Mountain View, CA, USA, 2013.

[34] International Technology Roadmap for Semiconductors, ITRS, Denver, CO, USA, 2013.

[35] J. L. Liang and H.-S. P. Wong, “Cross-point memory array without cell selectors-device characteristics and data storage pattern dependencies,” IEEE Trans. Electron Devices, vol. 57, no. 10, pp. 2531–2538, Oct. 2010.

[36] A. Chen, “A comprehensive crossbar array model with solutions for line resistance and nonlinear device characteristics,” IEEE Trans. Electron Devices, vol. 60, no. 4, pp. 1318-1326, April. 2013.

[37] J. Liang, S. Yeh, S. S. Wong, and H.-S. P. Wong, “Scaling challenges for the cross-point resistive memory array to sub-10nm node-an interconnect perspective,” in Proc. IEEE ICMW, May 2012, pp. 1-4.

[38] Y.-Y. Chen et al., “Balancing SET/RESET pulse for > 1010 endurance in HfO2/Hf1T1R bipolar RRAM,” IEEE Trans. Electron Devices, vol. 59, no. 12, pp. 3243-3249, Dec. 2012.

[39] B. Gao et al., “Oxide-based RRAM: Uniformity improvement using a new material-oriented methodology,” in Proc. Symp. VLSI Tech., June 2009, pp. 30-31.

[40] J. Park et al., “Improved switching uniformity and speed in filament-type RRAM using lightning rod effect,” IEEE Electron Device Lett., vol. 32, no. 1, pp. 63-65, Jan. 2011.

[41] H. Li et al., “3-D resistive memory arrays: from intrinsic switching behaviors to optimization guidelines,” IEEE Trans. Electron Devices, vol. 62, no. 10, pp. 3160-3167, Oct. 2015.

[42] X. Dong, C. Xu, Y. Xie, and N. Jouppi, “Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory,” IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, vol. 31, no. 7, pp. 994-1007, July 2012.