Improving Multilevel Writes on Vertical 3D Cross-Point Resistive Memory

Chengning Wang, Dan Feng, *Member, IEEE,* Wei Tong, Yu Hua, *Senior Member, IEEE,* Jingning Liu, Bing Wu, Wei Zhao, Linghao Song, Yang Zhang, Jie Xu, Xueliang Wei, and Yiran Chen, *Fellow, IEEE*

***Abstract*—Resistive memory is promising to be constructed as high-density storage-class memory. Multilevel cell, access-transistor-free cross-point array structure, and three-dimensional array integration are three approaches to scale up the density of resistive memory. However, composing the three approaches together strengthens the interactions between array-level and cell-level nonidealities (interconnect resistance-induced IR drop, sneak current, and device variability) of resistive memory arrays during write operations and significantly degrades write performance and reliability. In this article, we analyze the dynamic voltage-dividing effect along a selected write current path in 3D cross-point memory arrays. We propose a nonideality-tolerant high-density resistive memory (HD-RRAM) architecture, that can weaken the interactions between nonidealities and mitigate their degradation effects on the performance and reliability of array multilevel write operations. HD-RRAM is equipped with a double-transistor array architecture with two-transistor-n-resistor (2TnR) cell organization along pillars to reduce the current driving requirement and the large undesired voltage drop across each vertical pillar access transistor. Moreover, multiside asymmetric bias improves the resistive switching velocity by leveraging current-dividing effects. Variability-aware multilevel state partition reduces the worst-case write error rate by leveraging target state dependency of variability. Proportional-control multilevel state tuning reduces the average number of required write-and-verify iterations by leveraging pulse amplitude dependency of variability. Multilevel cell parallel writing improves the cell-level parallelism by leveraging the pass-through feature of intermediate resistance states. The evaluations show that HD-RRAM reduces both memory access latency and energy consumption over an aggressive baseline.**

***Index Terms*—high-density memory devices, crossbar, nonideal device characteristics, voltage drop analysis, memory array operation scheme, performance optimization.**

I. Introduction

In the big-data era, the ever-growing data-intensive applications raise higher requirements for the capacity of memory systems. Resistive memory is promising to break through the memory scalability bottleneck faced by one-transistor-one-capacitor (1T1C) DRAM [1], [2]. Metal-oxide resistive random-access memory (RRAM) has small cell area, 3D integration potential, long data-retention time ($\geqslant$1 year), and low read latency ($\leqslant$20 ns) [2]. Therefore, resistive memory can be built as high-density storage-class memory to reduce the slow I/O data movement when a page fault occurs.

As bit-line parasitic capacitance and memory capacitor charge leakage are two important device nonidealities of 1T1C DRAM arrays for read operations, IR drop, sneak current, and device variability are three important nonidealities of high-density resistive memory arrays for write operations [2], [3]. These array-level and cell-level nonidealities of resistive memory arrays become worse when the feature size scales down and significantly degrade the performance and reliability of write operations [3]–[5]. Interconnect resistance-induced IR drop lowers the effective write voltage and causes its cell-to-cell non-uniformity, which significantly enlarges the write latency [2], [6]–[10]. Moreover, IR drop and sneak current depend on interconnect geometry [2], [11], which shows a tradeoff between density and write reliability. Device variability becomes significant in nanoscale [12]. RRAM variability depends on the resistance state [13], [14], which may enlarge the write error during multilevel write operations.

To scale up the density of resistive memory, there are three common approaches: (1) multilevel memory cell (MLC), (2) access-transistor-free cross-point memory array structure, and (3) three-dimensional (3D) memory array integration. However, the interactions between array-level and cell-level non-ideal effects during write operations are more complex when the three high-density approaches are composed together. First, the sneak currents through half-selected cells enlarge the IR drop along the selected plane or lines, while the IR drop along the selected plane or lines decays the sneak currents through half-selected cells [2]. If conventional single-side bias scheme is applied on cross-point memory arrays, the effective write voltage is highly non-uniform among cells in the array and depends on multiple dynamic write operation parameters [2]. Moreover, the write latency setting of multilevel cell cross-point memory arrays is more sensitive to the IR drop-induced cell-to-cell non-uniformity of effective write voltage than that of single-level cell cross-point arrays. Second, multilevel write latencies of multilevel cells are more sensitive to the switching variability than those of single-level cells. Device variability makes the IR drop fluctuate, exacerbates the tail write latency problem, and causes write error, while IR drop complicates the pulse amplitude-dependent switching variability. Third, 3D

C. Wang, D. Feng, W. Tong, Y. Hua, J. Liu, B. Wu, W. Zhao, Y. Zhang, J. Xu, and X. Wei are with the Wuhan National Laboratory for Optoelectronics, Key Laboratory of Information Storage System, Engineering Research Center of data storage systems and Technology, Ministry of Education of China (School of Computer Science and Technology, Huazhong University of Science and Technology), Wuhan, Hubei, 430074, China (e-mail: dfeng@hust.edu.cn).

L. Song and Y. Chen are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA.

This work was supported in part by the National Natural Science Foundation of China No. 61832007, No. 61821003, No. 61772222, No. U1705261, No. 61772212, the National Science and Technology Major Project No. 2017ZX01032-101, and the Fundamental Research Funds for the Central Universities No. 2019kfyXMBZ037. (Corresponding author: Dan Feng.)

cross-point memory arrays have more sneak current branches than the 2D cross-point array with the same layer size and thus have larger IR drop [2]. Therefore, the interactions between coupled nonidealities of memory cells and interconnects in the memory arrays and their impact on write performance need to be better quantification and understood, in order to provide guidelines for the design of memory array operation scheme.

In this article, we implement multilevel memory based on 3D cross-point memory arrays, since multilevel memory is achieved by applying voltage pulses on the memory array to tune the cell resistance into the target state [15]. We choose the vertical 3D cross-point array architecture because of its lower cost-per-bit than the horizontal counterpart [12]. We first build a dynamic memory array model and analyze the interactions between cells and interconnects and the dynamic voltage-dividing effect in multilevel cell 3D cross-point memory arrays during write operations. We observe that in the conventional single-transistor array architecture, the small-size vertical pillar access transistor (VPAT) divides up a large portion of array bias voltage during write operations, causing switching speed degradation. Also, the cell-to-cell non-uniformity of effective write voltage in 3D cross-point arrays is significantly larger and is more complex than that in 2D cross-point arrays with the same layer size. Since the switching speed is an exponential function of the cell effective voltage [16], the cell-to-cell non-uniformity of effective write voltage makes it difficult to share a group of multilevel write latencies among all the multilevel cells in an array [2]. In this scenario, the high-performance cells near to voltage drivers will suffer from significant write performance loss if the memory controller only adopts the multilevel write latencies of the lowest-performance cell in the arrays. Therefore, we carry out static parameter and dynamic operation co-optimization from the viewpoint of the interactions between nonidealities of devices and interconnects in memory arrays, to improve the performance and reliability of multilevel write operations on ultra-high-density resistive memory while keeping its natural advantage of energy efficiency [2]. The contributions of this article are summarized as follows:

- We propose a double-transistor array architecture with two-transistor-n-resistor (2TnR) cell organization along pillar electrodes for 3D cross-point memory, to reduce the current driving requirement and the large undesired voltage drop across the small-size VPATs without extra area cost. We further propose multiside asymmetric bias to improve the resistive switching velocity of the selected cells by leveraging the current-dividing effects.
- We provide variability-aware multilevel state partition based on logarithmic cell resistance metric to reduce the write error rate by exploiting the target-state dependency of switching variability. We design proportional-control multilevel state tuning to reduce the average number of required write-and-verify iterations in a voltage pulse ramp by leveraging the pulse-amplitude dependency of switching variability.
- We devise multilevel cell parallel writing in cross-point memory arrays by leveraging the pass-through feature of intermediate resistance states.
- We evaluate the proposed high-density resistive memory architecture with comprehensive performance metrics both at the array level and full-system level.

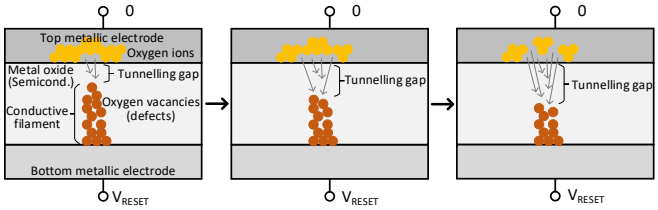

Fig. 1. Multilevel RESET process of a resistive memory cell achieved by tuning the voltage pulse duration.

II. High-Density Memory Approaches

Multilevel resistive memory cells use multilevel resistance states to store information, as shown in Fig. 1. The electron tunneling gap length can be taken as an internal physical state variable to represent the stored data pattern, and the cell resistance is an exponential function of the gap length [16]. The continuously-tunable dynamic range is subdivided into four states to store two bits in a cell. The switching to decrease cell resistance is called SET, whereas the switching to increase cell resistance is called RESET. The write latency is much longer than the read latency [16]. For the switching dynamics, the write velocity is an exponential function of the applied voltage amplitude across the cell [16]. Therefore, *the effective write voltage across the selected cell* is a key metric for analyzing multilevel write operations and characterizing multilevel write latencies of the selected cell.

Access-transistor-free cross-point memory arrays are constructed by two groups of interconnects (word-lines and bit-lines) perpendicular to each other with memory cells sandwiched in between [18], [19]. The removal of access transistors can reduce the equivalent cell area to $4/n\ F^2$, where n is the number of memory layers, and F is the feature size. Interconnects include 1D interconnect lines and 2D interconnect planes. For the same material, the latter has lower unit resistance than the former. 1/2 voltage bias scheme can be used to write the target cells in a cross-point memory array. The selected word-line and selected bit-lines are biased at 0 and V respectively, while unselected lines are biased at $V/2$ [2]. For read operations, we adopt the unselected word-line grounding with unselected bit-line floating (WG-BF) read scheme that outperforms with large read margin and low array power consumption among conventional read bias schemes [20]. Usually, a small group of neighboring cells on the selected word-line in a cross-point memory array is written at a time [6], due to the word-line IR drop criterion and the limited current drivability of the peripheral CMOS word-line driver [21], [22]. The read latency of the selected multilevel cell in a cross-point memory array is independent to the cell effective read voltage and array nonidealities, and it is calculated by the interconnect *RC* delay plus the peripheral transimpedance amplifier delay [21].

3D cross-point memory arrays can be classified into the wafer-like horizontal type [2], [23] and the multiholed vertical

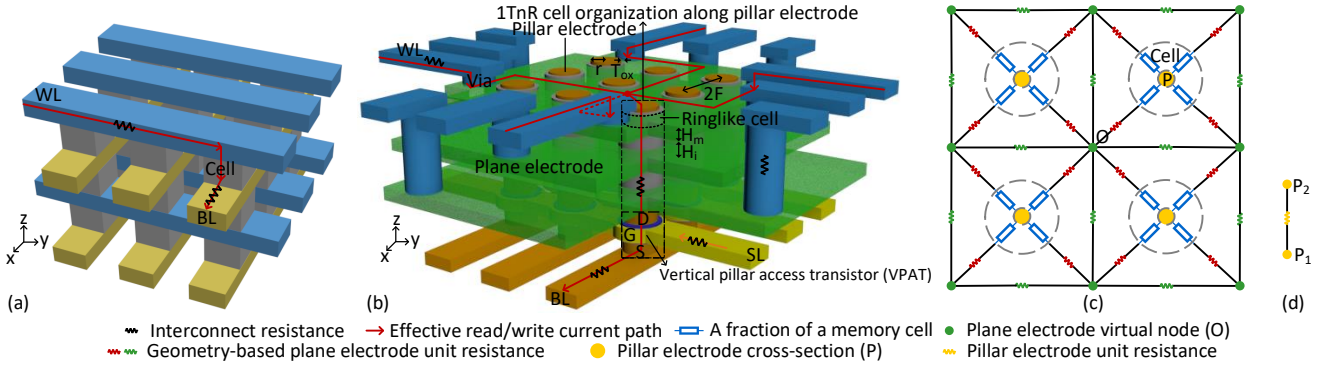

Fig. 2. 3D cross-point memory arrays. (a) horizontal line-type. (b) vertical plane-type [17]. (c) Resistor network model for the multiholed plane core and ringlike cells, where OP is described as a subcircuit. (d) Interplane connection. WL: word-line, BL: bit-line, SL: select-line.

type [12], [17], [24]–[26]. The array architectures are shown in Fig. 2(a) and Fig. 2(b) respectively. The vertical type has lower cost-per-bit than the horizontal counterpart [12]. Also, the vertical type uses the perforated plane electrodes as 2D interconnects, which reduces the unit interconnect resistance and thus the IR drop [12], [27], [28]. Thus, we mainly focus on the vertical type in this article. In the vertical type, ringlike cells are sandwiched between mutually perpendicular plane electrodes and pillar electrodes. A VPAT is serially connected with n parallel-connected memory cells to form a one-transistor-n-resistor (1TnR) structure along a pillar, and a select-line selects a vertical 2D cross-point array.

III. Analysis of Multilevel Write Operations on 3D Cross-Point Memory Arrays

To better understand the detailed dynamic interactions between multilevel memory cells, interconnects, and access transistors during multilevel write operations and better evaluate new techniques, we build a dynamic memory array model that can simulate write operations and analyze write process in multilevel cell 3D cross-point memory arrays. Nonlinear *I*–*V* characteristics and nonlinear switching dynamics of the bipolar-switching memory cell are described by the generic Verilog-A device model [16]. The C2C variability is introduced into the memory cell Verilog-A module [16] by incorporating the variability as a random fluctuation into the mean gap length change in the cell ($\mathrm{d}g/\mathrm{d}t$) [16] that expresses the resistance state change in a time step, where the random variable is realized by generating the random number following Gaussian distribution with the given deviation [29]–[32]. BSIM4 NMOS transistor model [16] in the predictive technology model [33] is incorporated into the array netlist for HSPICE simulation. The bias voltage that the transistor can endure is up to 5 V. We use the same interconnect materials as previous work [14], [25], [34]. The interconnect resistance and coupling capacitance are calculated based on geometry. The cell-level and array-level parameters in the model are shown in Table I. The feature size of memory cell is calculated as $F = r + T_{ox} + G_P$, and $2F$ is typically larger than the vertical transistor width (W) [24].

The ringlike memory cells in vertical 3D cross-point arrays are different from the cuboid memory cells in the horizontal-type, and the multiholed continuous 2D interconnect planes are also different from the 1D interconnect lines in the horizontal-type. To generate the array netlist for HSPICE analysis, the ringlike cells and plane electrodes should be discretized. The resistor network model for multiholed interconnect planes and ringlike cells is shown in Fig. 2(c). Every ringlike cell is discretized into 4 fractions along the current flow directions. The 4 fractions of a cell can be approximately viewed as parallel connected, thus the resistance of each fraction is 4 times of the whole cell resistance. The plane margin has the same mesh structure as the plane core. Inter-plane unit coupling capacitance is connected at the center point (O) of four neighboring pillar electrodes and between two neighboring plane electrodes.

However, the HSPICE system solving with operation simulation time and memory usage of simulating the dynamic array model of vertical 3D cross-point memory arrays are more than one-order-of-magnitude larger than those of simulating the static array model of horizontal-type 3D arrays or 2D arrays with the same number of memory cells. First, the grid discretization of plane electrodes and ringlike cells significantly increases the number of nodes in the network which equals to the number of KCL equations (i.e. system scale) and also

TABLE I
Parameters in the 3D Cross-Point Dynamic Memory Array Model

Parameter description	Range (Typical value)
Memory cell properties	HfO_2-based [16]
Gap length in the memory cell (l_{gap})	0.6∼1.23 nm
Variation factor of the memory cell (σ_f)	0∼5×10^4 (2.5×10^4) Ω/V
TiN pillar electrode top radius (r)	9∼33 (25) nm
Pillar etching slope (AR)	1280
HfO_2-based cell thickness (T_{ox})	5 nm
Inter-pillar half-gap at the top layer (G_P)	7∼31 (15) nm
Pt Plane electrode thickness (H_m)	5∼65 (20) nm
Insulation layer thickness (H_i)	5∼65 (14) nm
Cu word-line or bit-line thickness (H_l)	65 nm
Width of vertical NMOS transistors (W)	64 nm
Length of vertical NMOS transistors (L)	32 nm
Number of select-lines (n_s)	4∼128 (32)
Number of plane electrodes (n_l)	4∼64 (16)
Number of bit-lines (n_b)	4∼128 (32)
Maximum number of selected multilevel cells in an array during write op. (n_{max})	4
Inter-plane unit coupling capacitance (C_c)	0.3 fF
Array write bias voltage (V_{RESET})	2.55 V
Select-line voltage during RESET (V_{gRESET})	3.55 V
Unselected select-line voltage	0 V
Array read bias voltage (V_{read})	0.2 V
Select-line voltage during read (V_{gread})	1 V

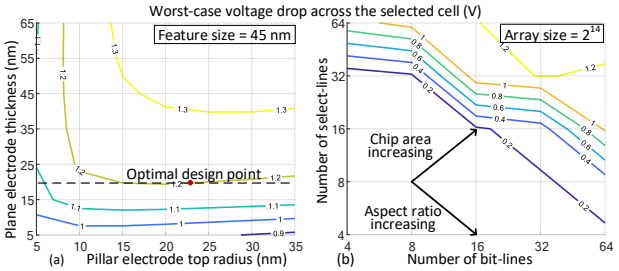

Fig. 3. Effective voltage as a function of geometric parameters: (a) plane thickness and pillar radius, and (b) array aspect ratio and chip area.

the number of branches (include behavioral cell fractions). Second, the network topology of vertical 3D cross-point arrays is denser than that of horizontal-type 3D arrays or 2D arrays. The node degree of post-discretized vertical 3D cross-point arrays is up to 8, whereas the node degree of horizontal-type 3D arrays and 2D arrays is no larger than 4 and 3, respectively. The node degree plus 1 equals to the number of nonzero entries on the row corresponding to that node in the system coefficient matrix (i.e. reflect sparsity). Third, as the dynamic Verilog-A memory device array and two transistor arrays are incorporated into the array netlist, they significantly increase the computational complexity. To reduce the simulation time and memory space usage of operating 3D cross-point memory arrays with plane electrodes and ringlike cells, we suggest that bias-operation-adaptive current-density-aware resistor network construction with only the principal components of the array may reduce the number of nodes and the degree of nodes in the network.

There are two array geometry tradeoffs between the mutually perpendicular plane electrodes and pillar electrodes for reducing the IR drop: pillar radius and plane thickness [3]. If the pillar electrode radius is too small, the pillar resistance is high. However, with the same feature size, if the pillar electrode radius is too large, the inter-pillar half-gap is small, thus the plane electrode resistance is high. Besides, if the plane electrode is too thin, the plane resistance is high. However, if the plane electrode is too thick, the pillar electrode is tall, thus the pillar electrode resistance is high. Therefore, we sweep the geometric parameters to identify the optimal design point with the highest effective voltage. The results are plotted in Fig. 3. In Fig. 3(a), we can see that the effective voltage is sensitive to pillar radius and plane thickness. Fig. 3(b) indicates that for the same array size, the effective voltage is higher when the plane size is larger, since the plane resistivity is lower than the pillar resistivity. As the cost-per-bit increases with the plane thickness increasing, we choose the array aspect ratio (length:width:height) as 2:2:1, and the selected design point is shown in Table I.

Since the switching velocity is an exponential function of the effective write voltage across the selected cell [16], we analyze the voltage drop breakdown along the write current path in the conventional single-transistor array (SITA) architecture with single-side bias (SSB) scheme. We observe that (1) In SITA architecture, the small-size VPAT divides up more than 30% of the array bias voltage, and only a small portion of voltage falls on the selected cell during write operations. (2) The effective voltage and the corresponding multilevel write latencies are highly non-uniform with SSB scheme (i.e., single-point voltage delivery), due to the interaction between interconnect IR drop and cell sneak currents in the memory arrays. The cell-to-cell latency non-uniformity ($>10\times$) makes it hard to share multilevel write latencies among cells in an array. (3) In the 3D cross-point memory arrays, the effective write voltage across the selected cell significantly increases with time during a RESET process, due to the time-varying resistive switching process of the selected cells and the dynamic voltage-dividing effect among the VPAT, the interconnects (plane and line), and the memory cell along the selected write current path. Thus, the multilevel write latencies of multilevel cell cross-point memory arrays are quite different from the multilevel write latencies of a standalone multilevel cell.

IV. HD-RRAM Architecture

By weakening the interactions between IR drop and the other two nonidealities and conquering multiple nonideality issues in each part of our design respectively, we can cope with each nonideality of resistive memory. After the interactions between nonidealities are weakened, the design complexity of the nonideality-tolerant memory array operation scheme is reduced by dividing the design of multilevel write scheme into four independent design stages to form a complete set of solutions: ❶ array voltage bias (DOTA-based MAB), ❷ multilevel cell state partition (VASP), ❸ multilevel cell state tuning (POST), and ❹ multilevel cell parallel writing (MPW). The bank-level architecture of high-density RRAM (HD-RRAM) design is illustrated in Fig. 4. Motivated by (1) the current-dividing effect along the vertical pillars, (2) the pulse-amplitude-dependent feature of switching variability, and (3) the pass-through feature of intermediate resistance states, HD-RRAM improves the performance of write operations on multilevel cell 3D cross-point memory arrays from three dimensions respectively: (1) improving the resistive-switching velocity under an array write pulse, (2) reducing the average number of write-and-verify iterations in a voltage pulse ramp, and (3) improving the cell-level write parallelism. The performance improvement of multilevel write operations can be decomposed into the product of the three dimensions.

A. Double-Transistor Array Architecture

The write operating current of metal-oxide resistive memory devices typically ranges from 10 μA to 100 μA, and it does not significantly decrease with the downscaling of device cross-sectional area due to the filamentary conduction mechanism [12]. 3D vertical cross-point arrays utilize a vertical pillar access transistor (VPAT) array underneath the memory array to provide 2D address decoding and drive sufficient write current. Since the VPATs have to be aligned to the pillar electrodes to retain high memory density, the channel width of the VPAT is horizontally limited by the array feature size, which limits the current drivability of the VPATs. Although the driving current of the VPAT can be increased by tuning the select-line and boosting its gate-source voltage by up to 5 V, it is still limited yet [24], [35]. The small-size VPAT (i.e. pillar current driver)

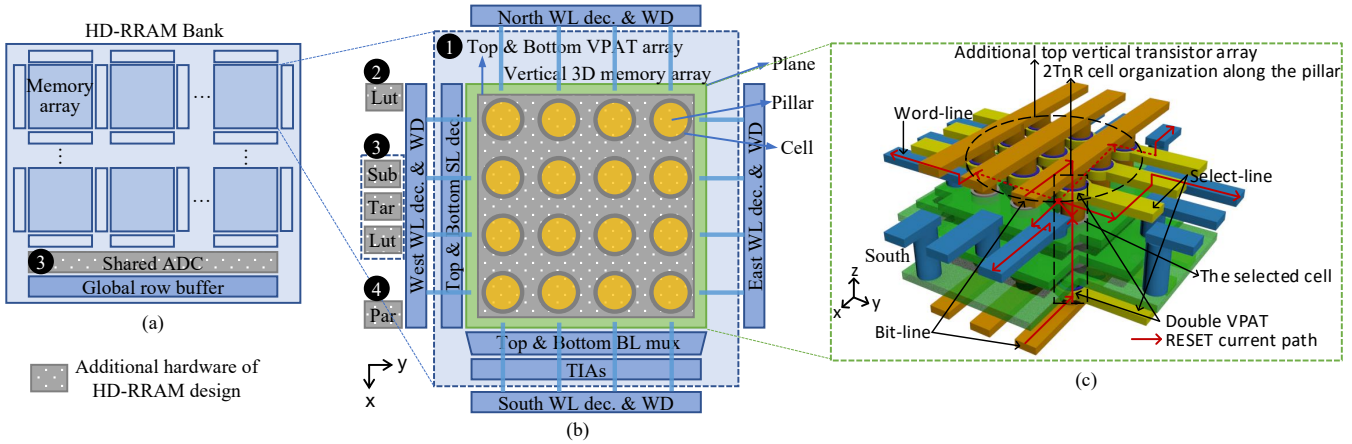

Fig. 4. HD-RRAM architectural overview. (a) array organization in a memory bank. (b) memory array and periphery. (c) double-transistor array architecture with distributed voltage delivery on the selected pillars and the selected plane. WL: word-line, BL: bit-line, SL: select-line, VPAT: vertical pillar access transistor, TIA: transimpedance amplifier, WD: write driver, dec.: decoder, Sub: substractor, Tar: target register, Lut: Lookup table, Par: parallelizing logic.

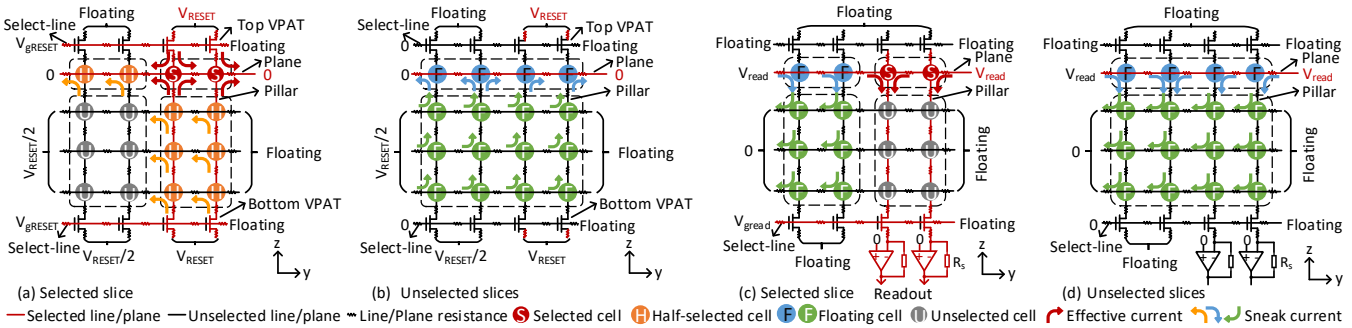

Fig. 5. Multiside asymmetric bias scheme. (a) and (b) are for 1/2 bias *write operations* on the selected vertical slice and unselected slices respectively. (c) and (d) are for WG-BF *read operations* on the selected vertical slice and unselected slices respectively. The two (bottom and top) groups of bit-lines connected to the two VPAT arrays are along the x-axis and not shown here.

has to drive *all the* $n_l - 1$ *sneak currents through* $n_l - 1$ *half-selected cells and only one effective write current through a selected cell along the selected pillar electrode* during 1/2 write bias operations, where n_l is the number of memory layers. We observe that the VPAT divides up more than 30% of the array bias voltage in the conventional single-transistor array (SITA) architecture with one-transistor-n-resistor (1TnR) cell organization along pillar electrodes where currents are only driven from the bottom end of the selected pillars (Fig. 2(b)) [17], [36]. The write current driving requirement of the selected pillars further increases as n_l increasing, since there are more sneak current branches connected to and thus more sneak currents flowing out of the selected pillar electrodes. We further observe that the enhancement which only connects four sides of the selected plane electrode to the voltage source simultaneously (multidirectional write driver (MWD) scheme [17]) *further enlarges the undesired voltage drop that the VPAT divides up* at the same array bias voltage. This is because in this case, the four parts of the selected plane electrode can be viewed as parallel connected, and the plane equivalent resistance and word-line equivalent resistance are reduced. Therefore, the VPAT has to drive larger current. The large voltage drop across the VPAT lowers the effective write voltage across the selected cell due to the voltage-dividing effect along the selected write current path, ultimately enlarging the write latency and array write power consumption.

To reduce the current driving requirement of the small-size VPATs, we propose *double-transistor array (DOTA) architecture* for 3D cross-point memory arrays, as shown in Fig. 4(c). The sandwiched array structure is successively composed of the bottom-layer VPAT array, the intermediate 3D memory array, and the top-layer VPAT array. The main enhancement is placing another VPAT array on top of the vertical 3D cross-point memory array to work with the bottom-layer VPAT array and form a two-transistor-n-resistor (2TnR) cell organization along a pillar electrode, where n is the number of memory layers (n_l). Two parallel VPATs are serially connected with n parallel-connected memory cells to divide and drive all the $n_l - 1$ sneak currents and the effective write current along a selected pillar electrode from both the top and the bottom ends of the pillar simultaneously, leveraging the current-dividing effect along the selected pillar electrodes. As the top-layer and bottom-layer VPATs can be viewed as parallel connected, the equivalent channel width of VPAT connected to a pillar is equivalently doubled. Thus, the equivalent drain-source resistance of VPAT is reduced, and the current drivability of VPAT is improved. As the additional top-layer VPAT array is aligned to the top end of the pillar electrode array, the top-layer VPAT array only increases the height of the whole array and does not incur extra area cost.

The top-layer VPAT array is typically fabricated separately from the 3D memory array with pillar electrode array [34] and

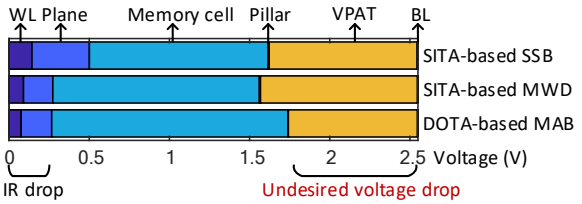

Fig. 6. Voltage-dividing effects on a *write* current path at the beginning of RESET operations for the conventional single-side bias [25], multidirectional write driver [17], and multiside asymmetric bias under their worst-case cell location respectively. VPAT refers to each VPAT.

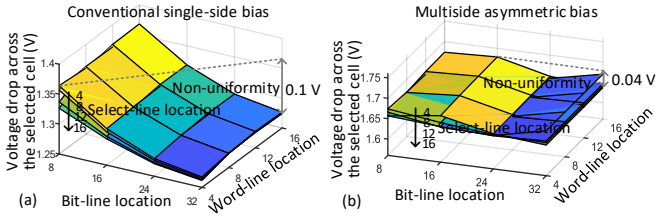

Fig. 7. Cell-to-cell non-uniformity of effective write voltage under (a) single-side bias and (b) multiside asymmetric bias, due to the interaction between interconnect IR drop and cell sneak currents in the 3D cross-point array.

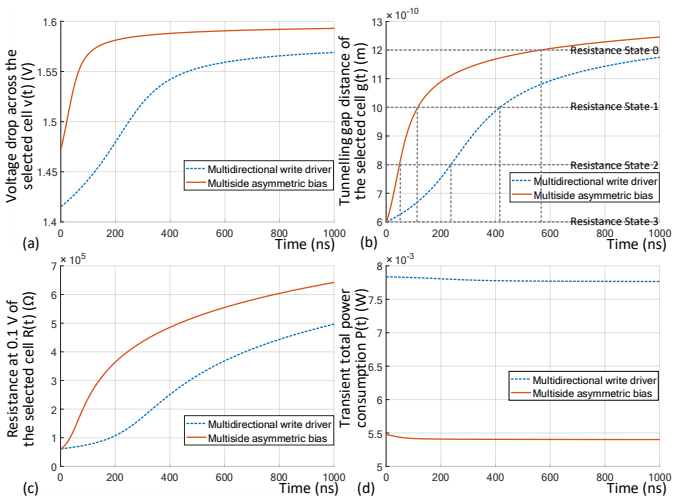

Fig. 8. Array-level time-domain dynamic write process of the worst-case location cell with all other cells in the lowest resistance state. Time evolution of (a) effective write voltage, (b) tunnelling gap length, (c) cell resistance, and (d) transient array total power consumption with MWD and MAB schemes. Array bias voltage for MWD and MAB is 2.9 V and 2.55 V respectively.

tacked onto the pillar electrode array [37]. Since vertical 3D cross-point resistive memory is compatible with the back-end-of-line, we can put the standalone-fabricated VPAT array on the memory array without incurring fabrication issues [37], [38]. The top-layer VPAT arrays and their decoders incur 16.9% extra cost-per-bit [24]. In our configuration, the feature size is determined by the bottom-most memory layer. Considering the limited pillar etching aspect ratio, as the top-layer VPAT array is aligned and tacked onto the pillar electrode array, the channel width of the top-layer VPATs is slightly larger than that of the bottom-layer VPATs. This scenario slightly enlarges the current drivability of the top-layer VPATs and improves the effect of multiside asymmetric bias scheme. Also, the nonideal etching effect may partially compensate the lower current drivability of the top-layer VPATs caused by possible larger leakage current of top-layer VPATs.

B. Multiside Asymmetric Bias

To reduce the undesired voltage drop along the selected multidirectional write current paths in multilevel cell 3D cross-point memory arrays, we propose multiside asymmetric bias (MAB) scheme. The side view of MAB scheme for the vertical-sliced 2D cross-point components of the vertical 3D cross-point array is shown in Fig. 5. MAB scheme is essentially a line connection scheme [3]. For write operations, MAB scheme connects two ends of the selected pillar electrodes and four sides of the selected plane electrode to the voltage sources simultaneously. Meanwhile, MAB only connects one end of unselected pillar electrodes and one side of unselected plane electrodes to the voltage source. For 1/2 bias method, full write bias voltage is applied to the selected electrodes while half write bias voltage is applied to unselected electrodes. The bias voltage of the top-layer select-line is set as the same as that of the bottom-layer select-line. The distributed voltage delivery of MAB makes the two terminals of a selected memory cell have multidirectional effective current write paths, thanks to the current-dividing effects along the selected pillars and the selected plane. For DOTA-based MAB, the effective write voltage across the selected cell is 1.47 V, 1.33 V, and 1.17 V at the beginning of RESET operations when the number of memory layers (n_l) is 16, 32, and 64 with 32×32 layer size respectively. To guarantee the cell effective voltage is larger than the RESET threshold voltage, up to 64 RRAM cells along a pillar can be controlled by the 2 VPATs, i.e., 2T-64R structures. The limiting factor of n_l is the increasing driving requirement of $n_l - 1$ sneak currents through $n_l - 1$ half-selected cells connected to a selected pillar electrode. 1/2-based MAB write scheme can be similarly extended to 1/3-based MAB write scheme.

For read operations, MAB scheme only connects one end of a pillar to the TIA for virtual grounding, and the top-layer VPAT array is not used, as shown in Fig. 5(c). This is because for WG-BF read method, other than a selected cell, there are only unselected cells and thus no sneak current along a selected pillar. The readout current of a selected bit-line equals to a single effective read current, thus the IR drop along a selected pillar can be neglected [21] during WG-BF read operations. WG-BF MAB read scheme does not require additional TIAs, compared with conventional single-side read bias schemes [20]. However, there are floating cells on the selected plane and sneak currents along the selected plane, thus the IR drop along the selected plane is still large, especially when multiple cells on the plane are selected to read at a time. Therefore, MAB read scheme still connects four sides of the selected plane to the voltage source.

Owing to the current-dividing effects and the distributed voltage delivery, DOTA-based MAB can reduce the write latency using even lower array bias voltage. Compared with SITA-based MWD [17], DOTA-based MAB can improve the worst-case effective write voltage from 1.28 V to 1.47 V when the array bias voltage is 2.55 V, as shown in Fig. 6. MAB can also lower the array bias voltage from 3 V to 2.55 V with similar write performance. Besides, MAB reduces the worst-case cell-to-cell non-uniformity of effective write voltage in the array by 60%, as shown in Fig. 7. MAB makes the

effective write voltage almost independent to the selected cell location by weakening the interactions between IR drop and the other two nonidealities, thus all the multilevel cells in the array can share the same group of multilevel write latencies. The comparisons of performance and energy consumption between SITA-based MWD [17] and DOTA-based MAB are demonstrated in Fig. 8. The figure reveals the dynamic voltage-dividing effect along the selected write current path in the memory array during write operations. DOTA-based MAB improves the transient RESET velocity by around 6.3 times. The instantaneous array power consumption during a dynamic write process includes the power consumption of all components in the memory array, and the results are shown in Fig. 8(d). The overall array power consumption only has a small decline during the RESET process, since the total current decreases while the effective write voltage increases. For DOTA-based MAB, the current-dividing effect along the selected pillar electrodes reduces the power consumption of the VPATs. Consequently, the total RESET power is reduced by 30.8%. The array energy consumption during a dynamic write process is reduced by 6.9 times, which is calculated as the time integral of the instantaneous array power consumption.

C. Variability-Aware Multilevel State Partition

The variability of resistive memory cells has two dimensions: extrinsic spatial device-to-device (D2D) variability and intrinsic temporal cycle-to-cycle (C2C) variability [4], [5]. Metal-oxide resistive memory suffers from much larger C2C switching variability than D2D variability due to its filament conduction mechanism [4], and the large variability significantly degrades write reliability. Generally, there are two influence factors of C2C variability: the target resistance state and the applied pulse amplitude [14]. C2C variability determines the maximum number of state levels that the dynamic range can be subdivided into, since two adjacent states are indistinguishable if they are overlapped with each other. C2C variability of a resistive memory cell makes the target state resistance after a write operation follow lognormal distribution [39]–[41]. Moreover, C2C variability is non-uniform across different states (i.e. interstate non-uniformity in a cell), and the variability exhibits some statistical rules. C2C variability increases with resistance state increasing [13], [39], [42], [43]. This is because when the memory device works in HRS, the conductive filament is thin, thus a small migration of the oxygen vacancies from the filament may significantly change the tunneling gap length and thus cell resistance [14]. The relative variation (σ/μ) can be approximated as a linear function of the mean value of the logarithm of cell resistance $\log_{10}(R_{cell})$: $\sigma/\mu = k\mu + b$, where σ and μ are the standard deviation and the mean value of the logarithm of cell resistance respectively, and $k = 1.9\times10^{-3}$ and $b = 1\times10^{-4}$ are constants [14], [44]. Here, σ is the deviation under the condition of a single long write pulse to switch one state to the adjacent state.

Since the cell resistance is an exponential function of the tunneling gap length in the cell [16], the conventional uniform multilevel state partition scheme [15] using the logarithmic

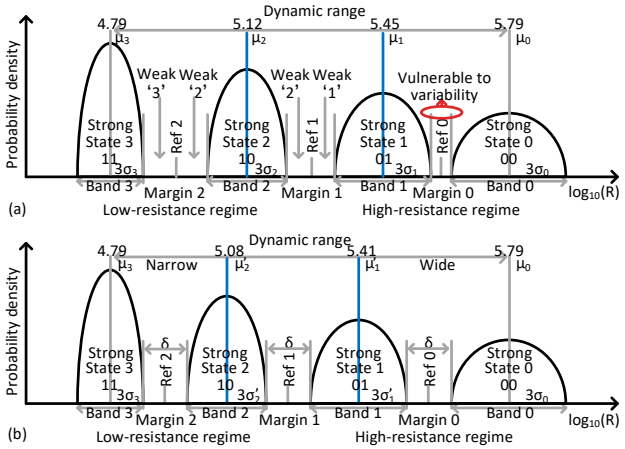

Fig. 9. (a) Conventional equal-difference state partition [15], [39] and (b) Variability-aware state partition with equal state margin based on the logarithm of cell resistance for resistive memory. Proper read margins between neighboring states can be further used to tolerate noises during read operations [21].

metric of cell resistance $\log_{10}(R_{cell})$ (resulting in linear increment of tunneling gap length) to equally divide the dynamic range into multilevel resistance states does not fully consider the state-to-state non-uniformity of C2C variability. If using conventional tunneling gap-equidistant state partition, the state margin in high-resistance regime is narrower than that in low-resistance regime, and two adjacent resistance state bands may even overlap with each other, as shown in Fig. 9(a). Also, in high-resistance regime, when a voltage pulse is applied on the cell in a multilevel write process, the cell resistance may overshoot the target state due to C2C variability and may easily fall into the next state, causing write error. In this scenario, a new voltage pulse is needed to tune the cell resistance back to within the target state and thus to recover the write error [39]. However, the extra recovering-write operation incurs extra energy and latency cost [45].

To reduce the write error rate by reducing the state overshoot probability in a multilevel write process, we design variability-aware multilevel state partition (VASP) scheme based on the $\log_{10}(R_{cell})$ cell state metric, shown in Fig. 9(b). The scheme makes high-resistance states wider whereas low-resistance states narrower by fixing the two boundary resistance states and appropriately moving the position of intermediate states to low resistance regime. To determine the mean position of intermediate states, we further formulate the state quantization rule as: $\mu_i - \mu_{i-1} = 3\sigma_{i-1} + \delta + 3\sigma_i$, where μ and σ are the mean value and the standard deviation of the lognormal resistance distribution of a state respectively, and the constant δ is the common state margin between physical states to tolerate the variability. As the positions of the two boundary states are known, μ_i for every intermediate states and the state margin δ can be determined by solving the system of quadratic equations. VASP makes every resistance states have the same tolerance to variability. Since VASP does not produce new resistance state levels, it does not affect the total number of intermediate resistance states. For the multilevel cell vertical 3D cross-point arrays, we use Monte Carlo simulation [8] for emulating the stochastic write process in our model. We generate the time-domain trace file including the evolution of the effective write voltage and cell resistance of each

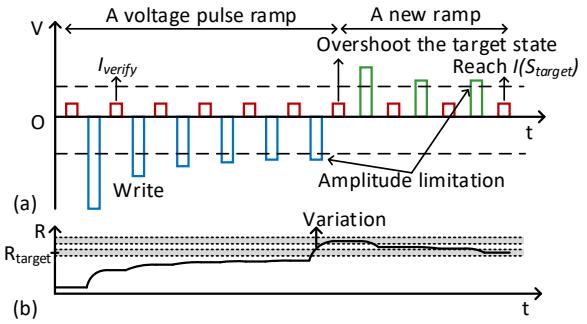

Fig. 10. (a) Proportional-control multilevel state tuning. (b) Illustration of cell resistance evolution.

time step (1 ns) in HSPICE transient analysis mode for each dynamic multilevel write operation process. We take each trace file corresponding to a stochastic write process as a sample. We run 500 times of the transient analysis process to get the resistance statistics. The result shows that VASP design can reduce the write error rate from around 2.6×10^{-3} to 8×10^{-4} compared with conventional gap-length equal-difference partition scheme.

D. Proportional-Control Multilevel State Tuning

Iterative write-and-verify resistance-state tuning methods are commonly used to overcome C2C variability during multilevel write operations that switches the current state to the target state [39], [45]. C2C variability has a feature that the absolute variation of the resistance change is proportional to the applied pulse amplitude: $\sigma = \sigma_f V_d$, where σ_f is the variation factor, and V_d is the effective voltage across the cell [14], [30], [42], [46]. To overcome C2C variability, iterative write-and-verify process with variable steps are used for precise multilevel write operations [45] by subdividing a single long write pulse into a series of short write pulses that each pulse incurs small absolute variation of resistance change. The verify pulses are used to determine the termination of the write process and make the write operation final resistance tolerant to C2C variability. However, to cover the worst-case tail latency caused by C2C variability, much higher latency value should be set for the memory controller. Conventional schemes [45], [46] using incremental voltage pulse ramps guarantee the reachability of the target state but may frequently overshoot much from the target state. Also, these schemes do not fully utilize the feedback information obtained from the verify pulses. The large number of write-and-verify iterations makes the verify pulses take a significant portion of time [39], [45], [46], enlarging the latency of a multilevel write operation.

To reduce the average number of write-and-verify iterations while tolerating C2C variability during multilevel write operations, we propose a proportional-control variability-adaptive multilevel state tuning (POST) algorithm at the array level. POST adjusts the pulse amplitude in a voltage pulse ramp to tune the selected cell into the target state with fast convergence by leveraging the pulse amplitude-dependent feature of variability, as illustrated in Fig. 10. The key idea is to apply large write pulse when the current state resistance is far away from the target state resistance whereas to apply small write pulse when the current state resistance is close to the target state resistance. This is because if the current state resistance is far away from the target state resistance, the memory cell is tolerable to large resistance variation, and we can apply large pulses but with large absolute variation to make the current resistance go faster to the target resistance range. However, if the current state resistance reaches close to the target state resistance range, the memory cell is vulnerable to variability, and we use small write pulses with small absolute resistance variation to refine the resistance change.

At the array level, since the RESET latency is slightly longer than the SET latency [16], we take the RESET operation as an example to illustrate the detailed design of POST algorithm. We fix word-line and select-line voltage amplitude while varying bit-line voltage pulse amplitude during a multilevel write operation. We further formulate the pulse amplitude V_{RESET} in a voltage pulse ramp applied on the selected bit-lines as

$$V_{RESET}(I^d_{verify}) = k_{RESET}\cdot(I^d_{verify} - I^d(S_{target})) + b_{RESET}$$

where k_{RESET} is the proportional coefficient, I^d_{verify} is the 4-bit digitalized value of the readout current measured from the 4-bit ADC by the verify pulse, $I^d(S_{target})$ is the 4-bit digitalized value of the readout current at the target state prestored in the target register, and b_{RESET} is the voltage pulse amplitude limitation that guarantees the target state can be timely reached. The value of b_{RESET} is the array basic RESET bias voltage in our configuration (2.55 V). For 1/2 bias scheme, unselected bit-lines and unselected word-lines are biased with $V_{RESET}(I^d_{verify})/2$ at the same time. To implement POST, an additional customized 4-bit ADC [47] instead of the original 2-bit ADC for MLC reading is connected to the output terminal of the transimpedance amplifier (TIA) [21], [46] at the end of each selected bit-line, which is to quantize and verify the readout current corresponding to the current cell resistance. Then the digital value is substracted with the digital readout value of the target state that is prestored in the target register by a 4-bit subtracter, as shown in Fig. 11. The result is input to a 16-entry lookup table to determine the V_{RESET} value, where the format of each entry is 2-bit integer part with 6-bit decimal part. The V_{RESET} value is taken as the input of the voltage pulse generator [48] for the next write-and-verify iteration. The schematic diagram of POST implementation is shown in Fig. 11. The additional components are globally shared among subbanks [2] in a memory bank. The latency of the V_{RESET} lookup and the write pulse generation [48] is estimated to be less than 2 ns. In a write-and-verify iteration, the write pulse width is fixed as 10 ns, and the verify pulse width is 6 ns, to cover the latency of peripheral circuitry. During SET operations, the bias voltage of the selected select-line can control the gate voltage of the VPAT, limit the SET current through the selected cell, and settle the over-SET problem. Since the gate of VPAT is connected to the select-line, for both RESET and SET operations, we use gate-first scheme, i.e., biasing the select-lines 1-ns before biasing the word-lines and bit-lines, and floating the select-lines 1-ns after floating the word-lines and bit-lines [16].

Proportional coefficient k_{RESET} influences the resistance change in a write-and-verify iteration step and the state

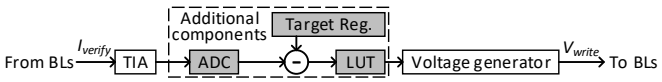

Fig. 11. Illustration of the POST algorithm implementation.

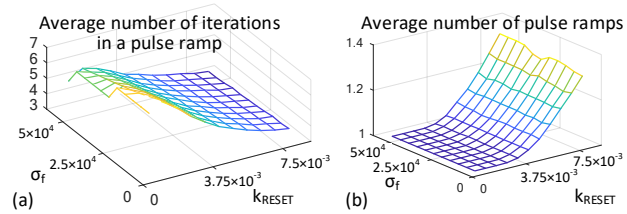

Fig. 12. The tradeoff between (a) the average number of iterations in a voltage pulse ramp and (b) the average number of required pulse ramps. σ_f is the variation factor, and k_{RESET} is the proportional coefficient.

TABLE II
THE MINIMUM AVERAGE NUMBER OF WRITE-AND-VERIFY ITERATIONS, LATENCY, AND ENERGY CONSUMPTION OF ADJACENT STATE TRANSITIONS USING POST (NUMBERS BEFORE BRACKETS) AND RISING PULSE AMPLITUDE STATE TUNING (NUMBERS IN BRACKETS) [45] AT THE ARRAY LEVEL.

State transit	3 to 2	2 to 1	1 to 0
Iterations	3.2 (4.5)	2.8 (3.7)	4.4 (6.7)
Latency (ns)	51.2 (72)	44.8 (59.2)	70.4 (107.2)
Energy (nJ)	0.246 (0.43)	0.38 (0.61)	0.55 (1.18)

overshoot rate in a voltage pulse ramp. If $k_{RESET} = 0$, it corresponds to the conventional equal-amplitude pulse write scheme. The variation factor σ_f that describes the pulse amplitude-dependent C2C variability of multilevel cells is chosen based on 10 ns write pulse width [44]. As shown in Fig. 12, as k_{RESET} increases, the average number of iterations decreases, but the average number of required pulse ramps increases. Therefore, the value of k_{RESET} shows a tradeoff between the average number of iterations in a voltage pulse ramp and the average number of required pulse ramps. As the overall write latency is proportional to the multiplication of these two metrics, we further use our model to statistically estimate the minimum number of required write-and-verify iterations for adjacent state transitions. The minimum number is achieved when $k_{RESET} = 7.5 \times 10^{-3}$, and the results are shown in Table II.

E. Multilevel Cell Parallel Writing

Finally, we provide a device-architecture level optimization to improve the cell-level write parallelism of multilevel cell resistive cross-point memory arrays. Parallel write operations on a cross-point array are achieved by selecting multiple bit-lines at a time. SET operations and RESET operations have to be performed separately on the same word-line in a cross-point array [7]. For multilevel cell cross-point arrays, the challenge of parallelizing the write operations is that a group of cells starts from and stops with different resistance states. Based on the observation that intermediate resistance states are frequently passed through, we propose an array-level multilevel cell parallel writing (MPW) method for resistive cross-point memory arrays, as shown in Algorithm 1. The key idea of MPW is to extract the common write-and-verify iterations among the to-be-written multilevel cells to improve cell-level write parallelism. Once the write process of a selected cell is terminated which is detected by the verify pulse, the corresponding bit-line is biased at the half of the array write bias voltage to mask the cell. To synchronize between cells, the overall write latency is set as the maximum latency among the selected cells. For vertical 3D cross-point arrays, parallel write operations are performed on the selected plane electrode in our configuration.

Algorithm 1: Multilevel cell parallel writing in the cross-point memory array

Input: Final state vector of the cell group with n_w cells: $\overrightarrow{s_f}$
Readout current state vector of the cell group: $\overrightarrow{s_c}$;
Assign total state changes $\overrightarrow{d_{RESET}}$ and $\overrightarrow{d_{SET}}$, where
$d_{RESET}(i) := s_f(i) < s_c(i)\ ?\ (s_c(i) - s_f(i)) : 0$ and
$d_{SET}(i) := s_f(i) > s_c(i)\ ?\ (s_f(i) - s_c(i)) : 0,\ i \in [1, n_w]$;
while $\overrightarrow{d_{RESET}} \neq \overrightarrow{0}$ **do**
  Common state change $\delta := \min\{d_{RESET}(i)\}, i \in [1, n_w]$;
  Parallel RESET the cells with $d_{RESET}(i) \neq 0$ using POST,
  where $I^d_{verify} - I^d(S_{target}) :=$
  $\min\{I^d_{verify}(i) - I^d(s_f(i) + d_{RESET}(i) - \delta)\}, i \in [1, n_w]$.
  Time cost of this step: $t = \max\{t_{cell}(i)\}, i \in [1, n_w]$;
  $\overrightarrow{d_{RESET}} := \overrightarrow{d_{RESET}} - \overrightarrow{\Delta}$, where $\Delta(i) = \delta$ if
  $d_{RESET}(i) \neq 0$ and $\Delta(i) = 0$ if $d_{RESET}(i) = 0$;
end
The SET operation process is similar to the above RESET process and is not shown here.

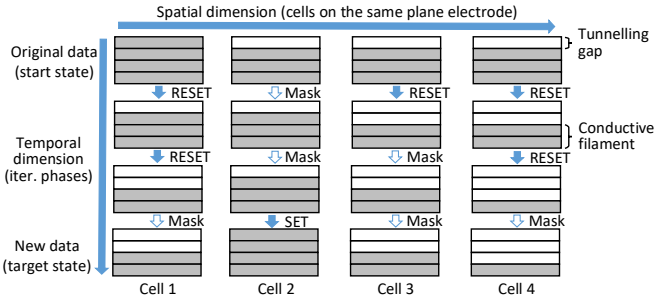

Fig. 13. Spatiotemporal diagram of the multilevel cell parallel writing process.

The multilevel write latencies of multilevel cells are much more sensitive to the cell effective voltage than those of single-level cells. If we select too much neighboring multilevel cells to write in parallel in an array, the non-uniformity of effective write voltage between these selected cells is significant. In this scenario, the write processes of these selected cells are not identical in a write operation and hard to synchronize. To reduce the non-uniformity of effective write voltage among the selected multilevel cells in a multilevel write operation and precisely control the state of each selected multilevel cell, we choose a relatively smaller number of neighboring multilevel cells to write in parallel in an array (4 in our configuration) than that for single-level cell parallel writing. A detailed illustration of the change process of cell resistance state of a cell group is shown in Fig. 13. Each write-and-verify iteration phase is synchronized. In this example, the parallelism of these phases are 3, 2, and 1 respectively. Sequential multilevel write scheme requires 6 iteration phases, whereas MPW only needs 3 phases, thanks to the improved cell-level write parallelism.

V. SYSTEM-LEVEL EVALUATION

A. Setup

We use gem5 simulator [49] integrated with NVMain [50] to evaluate the HD-RRAM design. Table III shows the con-

TABLE III
EVALUATED SYSTEM CONFIGURATION

Processor core	3 GHz, 4 cores, x86, out-of-order
Level-1 I&D cache	16 KB per core, 2 ways, 2-cycle access, private
Level-2 cache	4 MB, 64-byte block size, 16 ways, 20-cycle access, shared
Memory controller	Open-page policy, FR-FCFS scheduler
Memory organization	Stand-alone, 4 GB, 4 channels, 1 rank per channel, 8 banks per rank, global row buffer size: 64 Bytes
Resistive memory interface & bank timing parameters	DDR3-1066 with zero activate latency and zero restoration latency for memory arrays. The variable write latency (t_{WR}) is set according to the state transition in Table II. Bank read latency: 20 ns.

Rank	Bank	Subbank	Select-line	Word-line	Bit-line	Array
2	3	9	5	4	3	6

Fig. 14. The address-bit format of the 32-bit address.

figuration of the system processor and memory. The physical address-bit format in byte granularity is shown in Fig. 14. We configure the storage-class resistive memory as a DDR-compatible manner. For vertical 3D cross-point resistive memory, the precharge latency (t_{RP}) is modified as the worst-case charging time of the parasitic capacitance of word-line with plane electrode and bit-line with pillar electrode, to keep steady potential of the interconnects at the beginning of a voltage bias write or read operation [2], [51]. We run gem5 with the SPEC CPU2006 benchmarks for 500 million instructions to evaluate our four optimization methods: double-transistor array (DOTA) architecture with multiside asymmetric bias (MAB), variability-aware state partition (VASP), proportional-control state tuning (POST), and multilevel-cell parallel writing (MPW). The RRAM baseline is composed of the state-of-the-art approaches: single-transistor array (SITA) architecture with multidirectional write driver (MWD) [17], $\log_{10}(R)$ equal-difference state partition (ESP) [15], and increasing pulse-amplitude state tuning (IPST) [45]. We implement the basic schemes in the simulator as follows:

- Baseline: SITA-based MWD + ESP + IPST.
- Design 1: DOTA-based MAB + ESP + IPST.
- Design 2: DOTA-based MAB + VASP + IPST.
- Design 3: DOTA-based MAB + VASP + POST.
- Design 4: DOTA-based MAB + VASP + POST + MPW.

B. Performance

Fig. 15 shows the results of average write latency. Generally, the benchmarks that have larger writes-per-kilo-instructions (WPKI) have more significant write latency reduction, e.g., perlbench, mcf, leslie3d, and lbm. DOTA-based MAB (Design 1) and MPW (Design 4) have the most contribution to write latency reduction, because of the significant write velocity improvement and higher cell-level parallelism respectively. The MPW improvement varies with the average write parallelism of the benchmarks. POST (Design 3) also has a remarkable contribution to write latency reduction, benefiting from the reduction of write-and-verify iterations compared with IPST. Compared with the baseline RRAM, HD-RRAM reduces the average write latency by 34.1% across the benchmarks.

As write operations are accelerated, the read latency is also slightly reduced by 5%, since the request queuing latency is

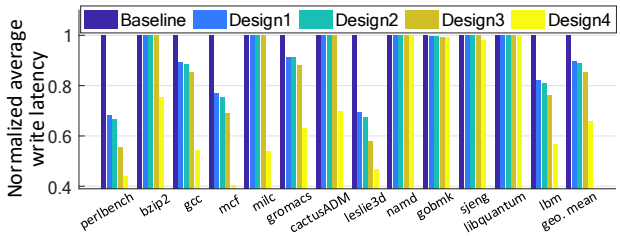

Fig. 15. Average write latency of different schemes normalized to the baseline.

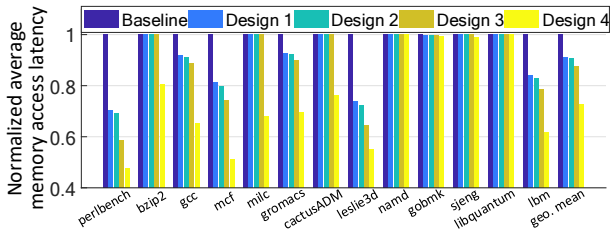

Fig. 16. Memory access latency of different schemes normalized to baseline.

reduced. The normalized average memory access latency is only a little higher than the normalized average write latency, as shown in Fig. 16. This is because the multilevel write latency is much longer than the multilevel read latency and the write latency dominates the memory access latency. HD-RRAM reduces the average memory access latency by 27.5% on average across the benchmarks.

The instructions per cycle (IPC) result is shown in Fig. 17. The IPC improvement relates to the write latency reduction and also the intensity of write access. Some benchmarks have significant reduction of memory access latency but do not have significant IPC improvement, due to small portion of memory write access, e.g., gcc and gromacs. The average IPC improvement of HD-RRAM system is 19.4% across the benchmarks, compared with the baseline.

C. Energy Consumption

The results of the energy consumption and energy-delay product (EDP) of HD-RRAM system are shown in Fig. 18. Here, EDP is used to reveal the energy efficiency of memory access operations. The energy consumption is mainly reduced by the DOTA-based MAB (Design 1), thanks to the reduction of array bias voltage and the distributed voltage delivery. Compared with the baseline, the energy reduction and EDP reduction of HD-RRAM system are 37.2% and 54.4% on average across the benchmarks, respectively.

D. Memory Density

The memory density of HD-RRAM can achieve 3.68 Gb·mm^{-2} at 45 nm feature size, and the density will further improve as the feature size scaling down. Moreover, compared with the designs that only adopt one or two high-density approaches, such as multilevel cell 2D one-transistor-one-resistor (1T1R) resistive memory [46], single-level cell 2D cross-point resistive memory [6], and single-level cell 3D cross-point resistive memory [17] at the 45 nm feature size, HD-RRAM improves the memory density by 24×, 32×, and 2×, respectively.

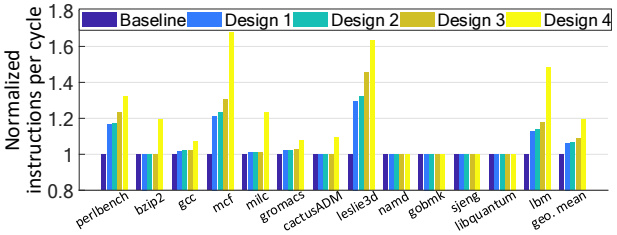

Fig. 17. Instructions per cycle of different schemes normalized to the baseline.

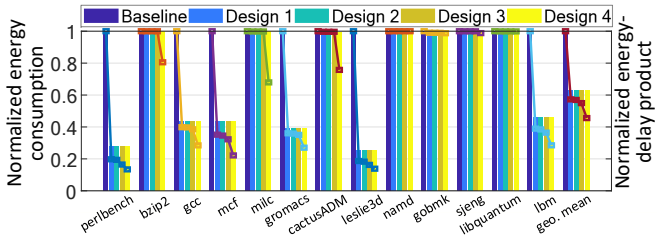

Fig. 18. Average energy consumption per memory access (bar plot) and energy-delay product (line plot) of different schemes normalized to baseline.

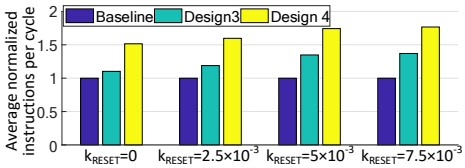

Fig. 19. Performance sensitivity to the proportional coefficient in POST over four write intensive benchmarks (perlbench, mcf, leslie3d, and lbm).

TABLE IV
EXTRA AREA/POWER/LATENCY OVERHEAD OF THE BANK-LEVEL PERIPHERAL CIRCUITRY OF HD-RRAM SYSTEM

	DOTA-based MAB	VASP	POST	MPW
Area (%)	-6.5	+0.059	+3.9	+1.4
Power (mW)	-147.2	+0.001	+70.2	+12.2
Latency (ns)	0.0	+0.1	+5.63	+0.20

E. Sensitivity Analysis

The proportional coefficient k_{RESET} in POST trades off between the number of pulses in a voltage ramp and the number of required voltage ramps. As we boost the array bias voltage at the beginning of a RESET process by enlarging k_{RESET}, the switching velocity improves but the variability also increases. So there is a latency tradeoff between the average number of write-and-verify iterations in a voltage pulse ramp and the average number of required pulse ramps. Fig. 19 shows the sensitivity of performance improvement to different k_{RESET} values in the POST scheme from 0 to 7.5×10^{-3}. Boosting k_{RESET} in POST achieves fast convergence to the target state without significantly increasing the state overshoot rate. The performance sensitivity analyses on other parameters (e.g. the array size and the number of cells to write in parallel) are similar to those in prior works [6], [15].

F. Hardware Implementation Cost

We implement the hardware components of HD-RRAM system in Synopsys Design Compiler by Verilog HDL to evaluate the extra hardware cost compared with the baseline, with respect to the area, power consumption, and latency at bank level. The memory array peripheral circuitry is evaluated based on TSMC 130 nm cell library and scaled to 22 nm technology size. The results are summarized in Table IV. First, to implement DOTA-based MAB, the extra components for each memory array include a top-layer VPAT array, a top-layer select-line decoder to control the top-layer VPAT array, and a top-layer bit-line multiplexer. These components are the same as the bottom part underneath the memory array, and the top and the bottom components are working in parallel. Thus, the top peripheral components do not incur extra area and latency cost. The extra peripheral circuitry can be hidden underneath the memory arrays by sharing ADCs among sub-banks. Moreover, benefiting from the current-dividing effects and distributed voltage delivery, the size of each voltage driver which is proportional to the line current [6], [12], [22] is even reduced by 48.3%. Owing to that, the total area of peripheral circuitry is reduced by 6.5%, and the power consumption of the bank-level peripheral circuitry is reduced by 147.2 mW. Second, to implement VASP, we only need to shift the reference input of the read reference comparators [6], [21] that are connected to the transimpedance amplifiers at the end of the selected bit-lines. The readout current values of the two intermediate states are input to the reference comparator [6], [15] to determine the termination of the write-and-verify iteration process when the verify pulse is applied. Therefore, for each memory array, VASP only requires a 2-entry 32-bit lookup table for recording the readout current values in unit of nA of the two intermediate resistance states. Third, additional components to implement POST for writing a cache-line in a 1Gb bank include 256 4-bit ADCs, 256 4-bit target registers, 256 4-bit substracters, and 64 16-entry 8-bit lookup tables. A customized 4-bit ADC [47] in a bank consumes extra 0.0137% chip area, 0.254 mW power, and 4 ns latency. In total, the peripheral circuitry including the ADCs to implement POST in a bank takes extra 3.9% chip area, 70.2 mW power, and 5.63 ns latency. Finally, the peripheral logic to support MPW in a bank consumes extra 1.4% chip area, 12.2 mW power, and 0.2 ns latency. These extra hardware costs are considered in the system-level evaluation.

G. Comparison with DRAM

1) DDR3 DRAM: We add the simulation of DDR3 DRAM in the NVMain to compare with HD-RRAM. The normalized IPC result is shown in Fig. 20. For most of the benchmarks, the IPC for HD-RRAM is comparable to that for DRAM. For some read-intensive benchmarks, the IPC for HD-RRAM is even higher than that for DRAM. Overall, the IPC for HD-RRAM system is 1.23 times of that for the DDR3 DRAM system on average across different benchmarks. Besides, the normalized energy consumption and energy-delay product results are shown in Fig. 21. Generally, HD-RRAM has great energy reduction over DRAM. For some benchmarks, the normalized EDP reduction of RRAM is less than the normalized energy reduction, since RRAM multilevel write latency is longer than DRAM write latency. On average, the EDP of HD-RRAM system is 16.7% of that of the DDR3 DRAM system across different benchmarks. Thus, HD-RRAM system has superior energy efficiency and performance advantage over DDR3 DRAM system.

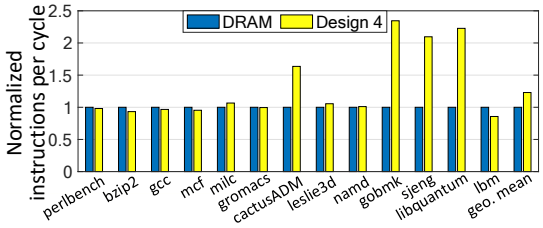

Fig. 20. Instructions per cycle for HD-RRAM system across different benchmarks normalized to the DDR3 DRAM system.

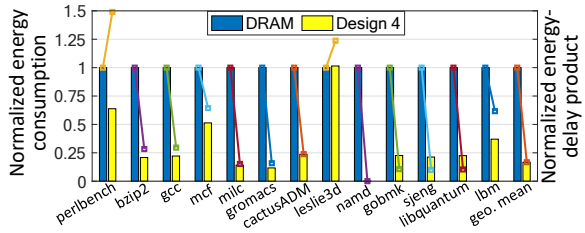

Fig. 21. Average energy consumption per memory access (bar plot) and energy-delay product (line plot) of HD-RRAM system across different benchmarks normalized to the DDR3 DRAM system.

2) HBM DRAM: We further add the gem5 simulation with 3D through-silicon via (TSV)-based high-bandwidth memory (HBM) DRAM [52]. Here the TSV parasitic capacitance of HBM is not considered in the simulation. The evaluation results show that the IPC for HD-RRAM is 52.6% of that for HBM DRAM system on average across the benchmarks. The HBM DRAM is integrated on the processor in our simulation, whereas HD-RRAM is configured as standalone and placed off-chip. HBM utilizes a 2D interface at die level to improve memory access parallelism and thus enlarges the bandwidth, whereas HD-RRAM does not use the wide interface. The design objective of HD-RRAM is different from that of HBM DRAM. The design objective of HBM DRAM is to improve memory access bandwidth by improving interface parallelism at die level to expand the conventional DDR DRAM and provide parallel memory access for bandwidth-bounded processors, while not aiming at reducing latency. On contrary, the design objective of HD-RRAM is to reduce write latency of high-density resistive memory at the memory array level, which is orthogonal to improve memory access parallelism by extending interface.

3) Write Endurance Concern: The write endurance of RRAM devices is up to 10^{10} write cycles [6], [13], which is lower than that of the storing-charge-based DRAM capacitors that can endure around 10^{16} writes [12]. This suggests that RRAM should be mainly applied in storage-class memory systems [2]. But the limited write endurance problem of RRAM might become a reliability challenge when RRAM is used for main memory. Table-based wear-leveling scheme can be used to improve the lifetime of cross-point memory arrays by exploiting the non-uniformity of write wear-out as a function of cell location [53]. Our proposed scheme aims to reduce the write latency of high-density resistive memory arrays at the physical-device level and is orthogonal to the system-level wear-leveling schemes.

VI. Related Work and Discussion

Multilevel resistive memory cells improve memory density by subdivides up intermediate resistance states from the two boundary states. Xu et al. [15] proposed equal-difference state partition (ESP) scheme for multilevel cells. Alibart et al. [45] proposed an increasing pulse-amplitude state tuning (IPST) scheme for the iterative write-and-verify process of a standalone multilevel cell. Hu et al. [46] increased the word-line pulse voltage and bit-line pulse voltage to realize multilevel states during SET operations in 1T1R memory arrays. Yao et al. [32] achieved precise multilevel write operations on eight 1T1R memory arrays. Zhang et al. [54] proposed to lower the gate voltage and enlarge the width/length ratio of the access transistors in the single-level cell 1T1R memory array to reduce the variability during SET operations.

Access-transistor-free cross-point memory arrays can achieve higher memory density than 1T1R memory arrays. 2D cross-point memory arrays with size of 17×17 [55], 20×20 [56], 24×24 [57], 32×32 [58]–[60], 48×48 [61], 54×108 [62], and 64×64 [63] were demonstrated. 1TnR memory array with multilevel cells was also demonstrated [36]. 3D memory array integration further improves memory areal density by stacking memory layers. 3D cross-point memory arrays with size of 2×8×8 [64] and 2×10×10 [65] were demonstrated. Yu et al. [27] compared three types of 3D cross-point arrays with respect to effective voltage, read margin, and write energy. Deng et al. [28] studied design guidelines for vertical 3D resistive memory. Chen et al. [66] profiled the voltage drop distribution on the selected slice of vertical 3D cross-point arrays. Xu et al. [24] built a static model for vertical 3D cross-point arrays. Xu et al. [17] proposed a multidirectional write driver (MWD) scheme based on the single-transistor array (SITA) architecture to reduce the IR drop along the selected plane electrode by connecting four sides of the distributed word-line to the voltage source at a time. Gao et al. [34] demonstrated cell-grouping write and read array bias schemes for the single-level cell vertical 3D cross-point memory array integrated with a single VPAT array at the bottom, where the write bias scheme is derived from unselected word-line 1/2 biasing with unselected bit-line floating scheme, and the read bias scheme is derived from unselected word-line grounding read-in-a-row scheme. Both write and read bias schemes are to the single side of interconnects. Hexagonal pillar layout, center landing of compact staircase contacts, and interarray global-shared-word-line memory array architecture were designed to improve areal density [26], [38], [67], [68]. The hexagonal pillar layout complicates the alignment of select-lines and bit-lines to the pillar electrodes and increases the array modeling complexity and operation simulation time.

Different from existing work, our proposed memory array architecture features double vertical-pillar-access-transistor arrays, and our distributed array operating bias scheme features multisided and asymmetric voltage sources connecting to different ends or sides of a interconnect line or plane. Besides, our solution is to reduce the latency of write operations on multilevel cell 3D cross-point memory arrays by decoupling

the array-level and cell-level nonideality issues and coping with these nonidealities in each part of the design stages respectively by dynamic operation optimization.

VII. Conclusion

From the perspective of the interactions between interconnects and devices in memory arrays, we analyze the impacts of the interactions between array-level and cell-level nonideal device properties (IR drop, sneak current, and device variability) on the performance and reliability of write operations and the design challenges of memory array operation scheme, when we compose multilevel memory cell, cross-point memory array structure, and 3D memory array integration approaches together. We propose HD-RRAM, a nonideality-tolerant ultra-high-density resistive memory architecture, to improve the performance and reliability of multilevel write operations by static parameter and dynamic operation co-optimization. Evaluations show that HD-RRAM significantly reduces both memory access latency and energy consumption on average across the benchmarks compared with the aggressive baseline.

References

[1] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nat. Nanotechnol.*, pp. 1–16, 2020.

[2] C. Wang, D. Feng, W. Tong, J. Liu, B. Wu, W. Zhao, Y. Zhang, and Y. Chen, "Improving write performance on cross-point RRAM arrays by leveraging multidimensional non-uniformity of cell effective voltage," *IEEE Trans. Comput.*, pp. 1–15, 2020.

[3] C. Wang, D. Feng, W. Tong, J. Liu, Z. Li, J. Chang, Y. Zhang, B. Wu, J. Xu, W. Zhao, Y. Li, and R. Ren, "Cross-point resistive memory: Nonideal properties and solutions," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 24, no. 4, pp. 46:1–46:37, 2019.

[4] P.-Y. Chen, B. Lin, I.-T. Wang, T.-H. Hou, J. Ye, S. Vrudhula, J.-s. Seo, Y. Cao, and S. Yu, "Mitigating effects of non-ideal synaptic device characteristics for on-chip learning," in *Proc. 2015 IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, 2015, pp. 194–199.

[5] S. Yu, P.-Y. Chen, Y. Cao, L. Xia, Y. Wang, and H. Wu, "Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect," in *Proc. 2015 IEEE Int. Electron Devices Meet. (IEDM)*, 2015, pp. 1–4.

[6] C. Xu, D. Niu, N. Muralimanohar, R. Balasubramonian, T. Zhang, S. Yu, and Y. Xie, "Overcoming the challenges of crossbar resistive memory architectures," in *Proc. 2015 IEEE 21st Int. Symp. High Performance Comput. Archit. (HPCA)*, 2015, pp. 476–488.

[7] C. Wang, D. Feng, J. Liu, W. Tong, B. Wu, and Y. Zhang, "DAWS: Exploiting crossbar characteristics for improving write performance of high density resistive memory," in *Proc. 2017 IEEE Int. Conf. Comput. Design (ICCD)*, 2017, pp. 281–288.

[8] Y. Zhang, D. Feng, W. Tong, Y. Hua, J. Liu, Z. Tan, C. Wang, B. Wu, Z. Li, and G. Xu, "CACF: A novel circuit architecture co-optimization framework for improving performance, reliability and energy of ReRAM-based main memory system," *ACM Trans. Archit. Code Optim.*, vol. 15, no. 2, pp. 1–26, 2018.

[9] Y. Zhang, D. Feng, W. Tong, J. Liu, C. Wang, and J. Xu, "Tiered-ReRAM: A low latency and energy efficient TLC crossbar ReRAM architecture," in *Proc. 2019 35th Symp. Massive Storage Syst. Technol. (MSST)*, 2019, pp. 92–102.

[10] B. Wu, D. Feng, W. Tong, J. Liu, C. Wang, W. Zhao, and Y. Zhang, "A low power reconfigurable memory architecture for complementary resistive switches," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, pp. 1–14, 2019.

[11] W. Steinhögl, G. Schindler, G. Steinlesberger, M. Traving, and M. Engelhardt, "Comprehensive study of the resistivity of copper wires with lateral dimensions of 100 nm and smaller," *J. Appl. Phys.*, vol. 97, no. 2, pp. 1–7, 2005.

[12] S. Yu and P.-Y. Chen, "Emerging memory technologies: Recent trends and prospects," *IEEE Solid-State Circuits Mag.*, vol. 8, no. 2, pp. 43–56, 2016.

[13] H. Lee, Y. Chen, P. Chen, P. Gu, Y. Hsu, S. Wang, W. Liu, C. Tsai, S. Sheu, P. Chiang, W. P. Lin, C. H. Lin, W. S. Chen, F. T. Chen, C. H. Lien, and M.-J. Tsai, "Evidence and solution of over-RESET problem for HfOx based resistive memory with sub-ns switching speed and high endurance," in *Proc. 2010 Int. Electron Devices Meet. (IEDM)*, 2010, pp. 1–4.

[14] B. Gao, Y. Bi, H.-Y. Chen, R. Liu, P. Huang, B. Chen, L. Liu, X. Liu, S. Yu, H.-S. P. Wong, and J. Kang, "Ultra-low-energy three-dimensional oxide-based electronic synapses for implementation of robust high-accuracy neuromorphic computation systems," *ACS nano*, vol. 8, no. 7, pp. 6998–7004, 2014.

[15] C. Xu, D. Niu, N. Muralimanohar, N. P. Jouppi, and Y. Xie, "Understanding the trade-offs in multi-level cell ReRAM memory design," in *Proc. 2013 50th ACM/EDAC/IEEE Des. Autom. Conf. (DAC)*, 2013, pp. 1–6.

[16] P.-Y. Chen and S. Yu, "Compact modeling of RRAM devices and its applications in 1T1R and 1S1R array design," *IEEE Trans. Electron Devices*, vol. 62, no. 12, pp. 4022–4028, 2015.

[17] C. Xu, P.-Y. Chen, D. Niu, Y. Zheng, S. Yu, and Y. Xie, "Architecting 3D vertical resistive memory for next-generation storage systems," in *Proc. 2014 IEEE/ACM Int. Conf. Comput.-Aided Des. (ICCAD)*, 2014, pp. 55–62.

[18] B. Wu, D. Feng, W. Tong, J. Liu, S. Li, M. Yang, C. Wang, and Y. Zhang, "Aliens: A novel hybrid architecture for resistive random-access memory," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, 2018, pp. 1–8.

[19] B. Wu, D. Feng, W. Tong, J. Liu, C. Wang, W. Zhao, and M. Peng, "ReRAM crossbar-based analog computing architecture for naive bayesian engine," in *Proc. 2019 IEEE 37th Int. Conf. Comput. Design (ICCD)*, 2019, pp. 147–155.

[20] J. Zhou, K.-H. Kim, and W. Lu, "Crossbar RRAM arrays: Selector device requirements during read operation," *IEEE Trans. Electron Devices*, vol. 61, no. 5, pp. 1369–1376, 2014.

[21] C. Wang, D. Feng, W. Tong, J. Liu, B. Wu, W. Zhao, and Y. Zhang, "Design and analysis of address-adaptive read reference settings for multilevel cell cross-point memory arrays," *IEEE Trans. Electron Devices*, vol. 66, no. 12, pp. 5347–5352, 2019.

[22] R. Berdan, T. Marukame, K. Ota, M. Yamaguchi, M. Saitoh, S. Fujii, J. Deguchi, and Y. Nishi, "Low-power linear computation using nonlinear ferroelectric tunnel junction memristors," *Nat. Electron.*, pp. 1–8, 2020.

[23] D. Kau, S. Tang, I. V. Karpov, R. Dodge, B. Klehn, J. A. Kalb, J. Strand, A. Diaz, N. Leung, J. Wu, S. Lee, T. Langtry, K.-w. Chang, C. Papagianni, J. Lee, J. Hirst, S. Erra, E. Flores, N. Righos, H. Castro, and G. Spadini, "A stackable cross point phase change memory," in *Proc. 2009 IEEE Int. Electron Devices Meet. (IEDM)*, 2009, pp. 1–4.

[24] C. Xu, D. Niu, S. Yu, and Y. Xie, "Modeling and design analysis of 3D vertical resistive memory–A low cost cross-point architecture," in *Proc. 2014 19th Asia & South Pacific Des. Autom. Conf. (ASP-DAC)*, 2014, pp. 825–830.

[25] H.-Y. Chen, S. Yu, B. Gao, P. Huang, J. Kang, and H.-S. P. Wong, "HfOx based vertical resistive random access memory for cost-effective 3D cross-point architecture without cell selector," in *Proc. 2012 Int. Electron Devices Meet.*, 2012, pp. 1–4.

[26] Z. Jiang, S. Qin, H. Li, S. Fujii, D. Lee, S. Wong, and H.-S. P. Wong, "Selector requirements for tera-bit ultra-high-density 3D vertical RRAM," in *Proc. 2018 IEEE Symp. VLSI Technol.*, 2018, pp. 107–108.

[27] S. Yu, Y. Deng, B. Gao, P. Huang, B. Chen, X. Liu, J. Kang, H.-Y. Chen, Z. Jiang, and H.-S. P. Wong, "Design guidelines for 3D RRAM cross-point architecture," in *Proc. 2014 IEEE Int. Symp. Circuits & Syst. (ISCAS)*, 2014, pp. 421–424.

[28] Y. Deng, H.-Y. Chen, B. Gao, S. Yu, S.-C. Wu, L. Zhao, B. Chen, Z. Jiang, X. Liu, T.-H. Hou, Y. Nishi, J. Kang, and H.-S. P. Wong, "Design and optimization methodology for 3D RRAM arrays," in *Proc. 2013 IEEE Int. Electron Devices Meet. (IEDM)*, 2013.

[29] X. Guan, S. Yu, and H.-S. P. Wong, "A SPICE compact model of metal oxide resistive switching memory with variations," *IEEE Electron Device Lett.*, vol. 33, no. 10, pp. 1405–1407, 2012.

[30] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Statistical fluctuations in HfOx resistive-switching memory: Part I-set/reset variability," *IEEE Trans. Electron Devices*, vol. 61, no. 8, pp. 2912–2919, 2014.

[31] W. Sun, B. Gao, M. Chi, Q. Xia, J. J. Yang, H. Qian, and H. Wu, "Understanding memristive switching via in situ characterization and device modeling," *Nat. commun.*, vol. 10, no. 1, pp. 1–13, 2019.

[32] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, "Fully hardware-implemented memristor convolutional neural network," *Nature*, vol. 577, no. 7792, pp. 641–646, 2020.

[33] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45 nm early design exploration," *IEEE Trans. Electron Devices*, vol. 53, no. 11, pp. 2816–2823, 2006.

[34] B. Gao, B. Chen, R. Liu, F. Zhang, P. Huang, L. Liu, X. Liu, J. Kang, H.-Y. Chen, S. Yu *et al.*, "3-D cross-point array operation on AlOy/HfOx-based vertical resistive switching memory," *IEEE Trans. Electron Devices*, vol. 61, no. 5, pp. 1377–1381, 2014.

[35] J. Zahurak, K. Miyata, M. Fischer, M. Balakrishnan, S. Chhajed, D. Wells, H. Li, A. Torsi, J. Lim, M. Korber, K. Nakazawa, S. Mayuzumi, M. Honda, S. Sills, S. Yasuda, A. Calderoni, B. Cook, G. Damarla, H. Tran, B. Wang, C. Cardon, K. Karda, J. Okuno, A. Johnson, T. Kunihiro, J. Sumino, M. Tsukamoto, K. Aratani, N. Ramaswamy, W. Otsuka, and K. Prall, "Process integration of a 27nm, 16Gb Cu ReRAM," in *Proc. 2014 IEEE Int. Electron Devices Meet. (IEDM)*, 2014, pp. 140–143.

[36] E. Hsieh, M. Giordano, B. Hodson, A. Levy, S. Osekowsky, R. Radway, Y. Shih, W. Wan, T. Wu, X. Zheng *et al.*, "High-density multiple bits-per-cell 1T4R RRAM array with gradual SET/RESET and its effectiveness for deep learning," in *Proc. 2019 IEEE Int. Electron Devices Meet. (IEDM)*, 2019, pp. 1–4.

[37] R. Micheloni, S. Aritome, and L. Crippa, "Array architectures for 3-D NAND flash memories," *Proc. IEEE*, vol. 105, no. 9, pp. 1634–1649, 2017.

[38] Z. Jiang, S. Qin, H. Li, S. Fujii, D. Lee, S. Wong, and H.-S. P. Wong, "Next-generation ultrahigh-density 3-D vertical resistive switching memory (VRSM)—Part II: Design guidelines for device, array, and architecture," *IEEE Trans. Electron Devices*, vol. 66, no. 12, pp. 5147–5154, 2019.

[39] B. Li, P. Gu, Y. Shan, Y. Wang, Y. Chen, and H. Yang, "RRAM-based analog approximate computing," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 12, pp. 1905–1917, 2015.

[40] G. Medeiros-Ribeiro, F. Perner, R. Carter, H. Abdalla, M. D. Pickett, and R. S. Williams, "Lognormal switching times for titanium dioxide bipolar memristors: origin and resolution," *Nanotechnology*, vol. 22, no. 9, 2011.

[41] B. Liu, H. Li, Y. Chen, X. Li, Q. Wu, and T. Huang, "Vortex: variation-aware training for memristor x-bar," in *Proc. 52nd Annual Des. Autom. Conf. (DAC)*, 2015, pp. 1–6.

[42] E. J. Merced-Grafals, N. Dávila, N. Ge, R. S. Williams, and J. P. Strachan, "Repeatable, accurate, and high speed multi-level programming of memristor 1T1R arrays for power efficient analog computing applications," *Nanotechnology*, vol. 27, no. 36, 2016.

[43] S. Cosemans, B. Verhoef, J. Doevenspeck, I. Papistas, F. Catthoor, P. Debacker, A. Mallik, and D. Verkest, "Towards 10000TOPS/W DNN inference with analog in-memory computing–A circuit blueprint, device options and requirements," in *Proc. 2019 IEEE Int. Electron Devices Meet. (IEDM)*, 2019, pp. 1–4.

[44] H. Jiang, L. Han, P. Lin, Z. Wang, M. H. Jang, Q. Wu, M. Barnell, J. J. Yang, H. L. Xin, and Q. Xia, "Sub-10 nm Ta channel responsible for superior performance of a HfO2 memristor," *Sci. rep.*, vol. 6, 2016.

[45] F. Alibart, L. Gao, B. D. Hoskins, and D. B. Strukov, "High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm," *Nanotechnology*, vol. 23, no. 7, 2012.

[46] M. Hu, C. E. Graves, C. Li, Y. Li, N. Ge, E. Montgomery, N. Davila, H. Jiang, R. S. Williams, J. J. Yang, Q. Xia, and J. P. Strachan, "Memristor-based analog computation and neural network classification with a dot product engine," *Adv. Mater.*, vol. 30, no. 9, 2018.

[47] K. D. Choo, J. Bell, and M. P. Flynn, "27.3 Area-efficient 1GS/s 6b SAR ADC with charge-injection-cell-based DAC," in *Proc. 2016 IEEE Int. Solid-State Circuits Conf.*, 2016, pp. 460–461.

[48] N. Papandreou, H. Pozidis, A. Pantazi, A. Sebastian, M. Breitwisch, C. Lam, and E. Eleftheriou, "Programming algorithms for multilevel phase-change memory," in *Proc. 2011 IEEE Int. Symp. Circuits & Syst. (ISCAS)*, 2011, pp. 329–332.

[49] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood, "The gem5 simulator," *ACM SIGARCH Comput. Archit. News*, vol. 39, no. 2, pp. 1–7, 2011.

[50] M. Poremba and Y. Xie, "NVMain: An architectural-level main memory simulator for emerging non-volatile memories," in *Proc. 2012 IEEE Comput. Soc. Annual Symp. VLSI*, 2012, pp. 392–397.

[51] H. Li, B. Gao, H.-Y. H. Chen, Z. Chen, P. Huang, R. Liu, L. Zhao, Z. J. Jiang, L. Liu, X. Liu *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, 2015.

[52] H. Jun, J. Cho, K. Lee, H.-Y. Son, K. Kim, H. Jin, and K. Kim, "HBM (high bandwidth memory) DRAM technology and architecture," in *Proc. 2017 IEEE Int. Memory Workshop (IMW)*, 2017, pp. 1–4.

[53] W. Wen, Y. Zhang, and J. Yang, "Wear leveling for crossbar resistive memory," in *Proc. 2018 55th ACM/ESDA/IEEE Des. Autom. Conf. (DAC)*, 2018, pp. 1–6.

[54] Y. Zhang, Z. Zhou, P. Huang, M. Fan, R. Han, W. Shen, L. Liu, X. Liu, B. Gao, H. Wu *et al.*, "An improved RRAM-based binarized neural network with high variation-tolerated forward/backward propagation module," *IEEE Trans. Electron Devices*, vol. 67, no. 2, pp. 469–473, 2020.

[55] Z. Li, M. D. Pickett, D. Stewart, D. A. Ohlberg, X. Li, W. Wu, W. Robinett, and R. S. Williams, "Experimental demonstration of a defect-tolerant nanocrossbar demultiplexer," *Nanotechnology*, vol. 19, no. 16, 2008.

[56] F. M. Bayat, M. Prezioso, B. Chakrabarti, H. Nili, I. Kataeva, and D. Strukov, "Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits," *Nat. commun.*, vol. 9, no. 1, 2018.

[57] H. Li, Z. Chen, W. Ma, B. Gao, P. Huang, L. Liu, X. Liu, and J. Kang, "Nonvolatile logic and in situ data transfer demonstrated in crossbar resistive RAM array," *IEEE Electron Device Lett.*, vol. 36, no. 11, pp. 1142–1145, 2015.

[58] P. M. Sheridan, F. Cai, C. Du, W. Ma, Z. Zhang, and W. D. Lu, "Sparse coding with memristor networks," *Nat. nanotechnol.*, vol. 12, no. 8, pp. 784–789, 2017.

[59] S. H. Jo, K.-H. Kim, and W. Lu, "High-density crossbar arrays based on a Si memristive system," *Nano lett.*, vol. 9, no. 2, pp. 870–874, 2009.

[60] H. Yeon, P. Lin, C. Choi, S. H. Tan, Y. Park, D. Lee, J. Lee, F. Xu, B. Gao, H. Wu *et al.*, "Alloying conducting channels for reliable neuromorphic computing," *Nat. Nanotechnol.*, pp. 1–6, 2020.

[61] I. Kataeva, S. Ohtsuka, H. Nili, H. Kim, Y. Isobe, K. Yako, and D. Strukov, "Towards the development of analog neuromorphic chip prototype with 2.4 M integrated memristors," in *Proc. 2019 IEEE Int. Symp. Circuits & Syst. (ISCAS)*, 2019, pp. 1–5.

[62] F. Cai, J. M. Correll, S. H. Lee, Y. Lim, V. Bothra, Z. Zhang, M. P. Flynn, and W. D. Lu, "A fully integrated reprogrammable memristor-CMOS system for efficient multiply-accumulate operations," *Nat. Electron.*, vol. 2, no. 7, 2019.

[63] M. R. Mahmoodi, H. Kim, Z. Fahimi, H. Nili, L. Sedov, V. Polishchuk, and D. B. Strukov, "An analog neuro-optimizer with adaptable annealing based on 64×64 0T1R crossbar circuit," in *Proc. 2019 IEEE Int. Electron Devices Meet. (IEDM)*, 2019, pp. 14.7.1–14.7.4.

[64] C. Li, L. Han, H. Jiang, M.-H. Jang, P. Lin, Q. Wu, M. Barnell, J. J. Yang, H. L. Xin, and Q. Xia, "Three-dimensional crossbar arrays of self-rectifying Si/SiO2/Si memristors," *Nat. Commun.*, vol. 8, 2017.

[65] G. C. Adam, B. D. Hoskins, M. Prezioso, F. Merrikh-Bayat, B. Chakrabarti, and D. B. Strukov, "3-D memristor crossbars for analog and neuromorphic computing applications," *IEEE Trans. Electron Devices*, vol. 64, no. 1, pp. 312–318, 2016.

[66] H.-Y. Chen, B. Gao, H. Li, R. Liu, P. Huang, Z. Chen, B. Chen, F. Zhang, L. Zhao, Z. Jiang, L. Liu, X. Liu, J. Kang, S. Yu, Y. Nishi, and H.-S. P. Wong, "Towards high-speed, write-disturb tolerant 3D vertical RRAM arrays," in *Proc. 2014 Symp. VLSI Technol.*, 2014, pp. 1–2.

[67] S. Qin, Z. Jiang, H. Li, S. Fujii, D. Lee, S. S. Wong, and H.-S. P. Wong, "Next-generation ultrahigh-density 3-D vertical resistive switching memory (VRSM)—Part I: Accurate and computationally efficient modeling," *IEEE Trans. Electron Devices*, vol. 66, no. 12, pp. 5139–5146, 2019.

[68] S.-H. Chen, H.-T. Lue, Y.-H. Shih, C.-F. Chen, T.-H. Hsu, Y.-R. Chen, Y.-H. Hsiao, S.-C. Huang, K.-P. Chang, C.-C. Hsieh, G.-R. Lee, A.-T.-H. Chuang, C.-W. Hu, C.-J. Chiu, L. Y. Lin, H.-J. Lee, F.-N. Tsai, C.-C. Yang, T. Yang, and C.-Y. Lu, "A highly scalable 8-layer vertical gate 3D NAND with split-page bit line layout and efficient binary-sum MiLC (minimal incremental layer cost) staircase contacts," in *Proc. 2012 Int. Electron Devices Meet. (IEDM)*, 2012, pp. 1–4.

Authors' photographs and biographies not available at the time of publication.