Resistive RAM-Centric Computing: Design and Modeling Methodology

Haitong Li, *Student Member, IEEE*, Tony F. Wu,
Subhasish Mitra, *Fellow, IEEE*, and H.-S. Philip Wong, *Fellow, IEEE*

***Abstract*—Memory-centric computing with on-chip non-volatile memories provides unique opportunities for native and local information processing in an energy-efficient manner. Design and modeling methodology based on resistive random access memory (RRAM) is presented in this work. A hierarchical RRAM SPICE model having different levels of physics realism is described, where the incorporated stochasticity provides a more accurate representation of RRAM operations. Three in-memory operation schemes are developed and experimentally verified for reconfigurable in-memory logic, using RRAM built in 3D vertical structure (i.e., 3D RRAM). As a case study for RRAM-centric computing systems, we evaluate the use of 3D RRAMs for a language recognition system using the hyperdimensional (HD) computing model. Utilizing the inherent properties of 3D RRAM, we demonstrate, using fabricated 3D RRAM integrated with FinFET, the essential kernels for HD operations: multiplication, addition, and permutation (MAP). RRAM-centric HD systems exhibit strong resilience to hard errors induced by RRAM endurance failures, making a promising case for using various types of RRAM for memory-centric HD systems.**

***Index Terms*—Resistive random access memory (RRAM, ReRAM), memory-centric computing, processing-in-memory, hyperdimensional computing, SPICE model, Internet of Things (IoT).**

I. Introduction

THE Internet of Things (IoT) aims at distributed sensing, processing, and exchange of data in a world that is more interconnected than ever. The rising demand for real-time processing capabilities at the IoT nodes poses new challenges for the semiconductor industry across multiple technology layers, from device and manufacturing to circuit and architecture [1]. Embedded memories in IoT nodes are important as they hold great potential of locally and natively processing these data to overcome the 'memory wall', which arises due to the off-chip traffic to pin-limited DRAMs in conventional systems [2]-[4]. Among various emerging non-volatile memory technologies that may help to address these issues, 3D RRAM technology offers a high-capacity, high-bandwidth on-chip storage solution, as well as the capability towards fine-grained monolithic 3D integration with logic [5]-[7]. Additionally, memory subsystems can be designed in a smart way by equipping logic functionalities therein, leading towards memory-centric computing systems. Previous examples in such context include performing or accelerating certain computations within SRAM and TCAM [8]-[10], as well as various emerging non-volatile memories [11]-[17]. In this work, to facilitate the design and modeling of RRAM-centric computing systems, first an RRAM SPICE model having three hierarchical levels of physics realism is described. This hierarchical model is more compatible with circuit and system analysis compared with several RRAM physical models [18]-[20], and meanwhile still maintains different levels of depth of device physics such as sub-threshold stochastic switching, which makes it more accurate with better generalization than analytical models [21], [22]. Next, as an example of building blocks for RRAM-centric computing systems, three in-memory operation schemes are developed for flexible and reconfigurable in-memory logic operations within 3D RRAM. The operation schemes are experimentally verified with electrical measurements on one transistor-four resistor (1T-4R) 3D RRAM devices. Finally, as a case study, we design and implement multiplication-addition-permutation (MAP) kernels using 1T-4R 3D RRAMs for hyperdimensional (HD) computing, a neural-inspired cognitive computation model capable of learning from few examples [23]-[25]. Taking the device variations and endurance limitations into account, system-level simulations are performed to evaluate an RRAM-based HD language recognition system, showing strong resilience to hard errors induced by RRAM endurance failures.

The rest of the paper is organized as follows. In Section II, a brief overview of RRAM technologies is given, emphasizing the use of RRAM for memory-centric computing systems. Section III discusses the stochastic switching behaviors of RRAM and the corresponding SPICE model. Section IV introduces three in-memory operation schemes and experimental demonstrations of logic operations with 1T-4R 3D RRAMs. Section V further presents the evaluation of a RRAM-centric HD language recognition system. Finally, we conclude the paper in Section VI.

Manuscript received 2017. This work was supported in part by NSF-SRC E2CDA, NSF NCN NEEDS, Stanford SystemX Alliance, STARnet SONIC, and by the Member Companies through the Stanford Non-Volatile Memory Technology Research Initiative (NMTRI).

H. Li[*], T. F. Wu[*], S. Mitra[*, #] and H.-S. P. Wong[*] are with [*]Department of Electrical Engineering, [#]Department of Computer Science, and Stanford SystemX Alliance, Stanford University, Stanford, CA 94305, USA (haitongl@stanford.edu, tonyfwu@stanford.edu, subh@stanford.edu, hspwong@stanford.edu).

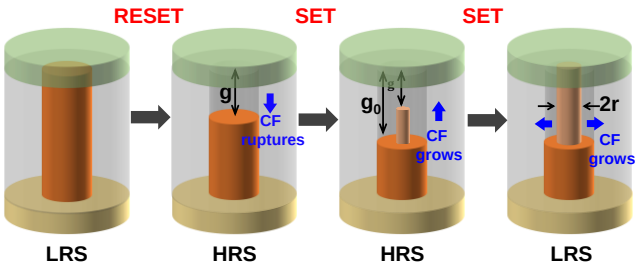

Fig. 1. Illustration of conductive filament (CF) evolution process during SET and RESET operations of RRAM. *g* is the gap distance between the top electrode and the CF tip. *r* is the radius of the formed CF, modeling the lateral filament growth during SET operation.

II. Overview of RRAM Technologies

RRAM employs resistive switching phenomenon in metal-oxide materials to store information in a non-volatile manner [26], [27]. An RRAM cell consists of a top electrode (TE), a bottom electrode (BE), and a metal-oxide layer in between, forming a simple metal-insulator-metal (MIM) structure. Resistive switching can be induced by applying voltage across an RRAM cell. The applied electrical field drives the generation, motion, and recombination of oxygen vacancies (V_O), which leads to the formation and rupture of conductive filaments (CFs) in the oxide layer. As illustrated in Fig. 1, the formation of CFs connects the top and bottom electrodes, and the two-terminal device changes from a high resistance state (HRS) to a low resistance state (LRS) once the applied voltage exceeds the threshold voltage for this specific write trial. This is referred to as the SET process and the threshold voltages are called the SET voltage (V_{SET}), which typically follows a statistical distribution due to device variability. The rupture of CFs causes the RRAM to switch from LRS back to HRS once the applied voltage exceeds the threshold value for this write trial. This is referred to as the RESET process and the threshold voltages are called the RESET voltage (V_{RESET}) following a statistical distribution. RRAM can be fabricated in the back-end-of-line (BEOL) interconnect wiring layers using CMOS-compatible materials [28], and further can be built into a high-density 3D structure using low BEOL-compatible fabrication temperatures [5], [6], enabling monolithic 3D ICs [7]. Promising characteristics have been reported on various types of RRAM, including 50-nA low-current AlO_x RRAM [29], 300-ps fast-switching HfO_x RRAM [30], 10 nm scaled RRAM cells [31], TaO_x bilayer RRAM with 10^{12} endurance cycles [32], as well as Gb-level-capacity functional chip demonstrations [33], [34]. Recent advancements have also been reported regarding closer integration with state-of-the-art CMOS platform, monolithic 3D chip demonstration, and reliability improvement at cell level and array level [35]-[38].

These demonstrated characteristics along with inherent device properties enables unique logic designs and applications using RRAM. Various RRAM-based designs have been reported for material implication (*IMP*) logic [17], [39]-[42], sequential logic [43]-[45], and majority-inverter graphs [46]. For memory-centric computing systems, RRAM endurance might become the bottleneck if write operations are frequent on certain physical addresses. At the device level, several strategies for improving the endurance have been reported. RRAM devices dominated by interfacial switching physics have improved endurance characteristics, trading-off the retention time [47]. Use of a via-hole device structure can lead to two orders of magnitude improvement in endurance over conventional cross-point structure, due to confined CF paths [48]. A plasma-oxidized bilayer oxide structure exhibits endurance up to 10^{12} cycles [32]. From a circuit operation perspective, a large optimization space exists that includes the bias scheme, pulse amplitude, pulse width, rise/fall time, and pulse shape. For instance, it is reported that optimized SET/RESET pulses or shorter RESET pulse width is beneficial for HfO_x-based and TaO_x-based RRAM [49]. Furthermore, recovery scheme can be employed in the controller circuitry to recover failed bits [50]. Many endurance failures are not due to hard breakdown; therefore, a failure bit can be recovered by a one-shot DC sweep [49] or AC pulse operation [50]. Besides device-level and circuit-level strategies, it may be even more important to co-design the algorithms and the memory-centric computing systems that can mitigate the endurance limitation in the first place. We will elucidate on this point with a case study in Section V.

III. SPICE Model of RRAM with Stochasticity

Developing RRAM SPICE models that capture and generalize the key device characteristics is important for RRAM circuit design [18]. Moreover, in the context of designing RRAM-centric computing systems, having a hierarchy of model levels in terms of complexity and depth of RRAM device physics can be greatly beneficial for a variety of design objectives as well as reducing simulation time. Here, we extend our previous Verilog-A coded SPICE models [51], [52], and describe three model levels (*Level 1*, *Level 2*, and *Level 3*).

Resistive switching characteristics of RRAM devices have a strong correlation with the CF evolution processes, where the electric field and Joule heating effects play critical roles. Based on this physical picture, our previous SPICE model captures the two-dimensional CF evolution process with tunneling gap distance (g) and CF radius (r) as the key variables, and uses a set of differential equations, dependent on the electrical field and temperature, to describe the evolution of g and r during SET/RESET operations [51]. The I-V characteristics are determined by the conduction mechanisms consisting of hopping current and metallic current, as described by:

$$I_{hop} = I_0 \left(\pi r^2 / 4\right) \exp\left(-g / g_T\right) \sinh\left(V_{gap} / V_T\right) \quad (1)$$

$$I_{CF} = \pi r^2 V_{CF} / 4\rho\left(g_0 - g\right) \quad (2)$$

where I_0 is the hopping current density, and g_T and V_T are fitting parameters. The *Level 1* model reproduces the core resistive switching behaviors during SET/RESET operations, as indicated by Fig. 2 that compares the measured and modeled DC I-V characteristics of TiN/HfO_x/TiO_x/Pt RRAM. The

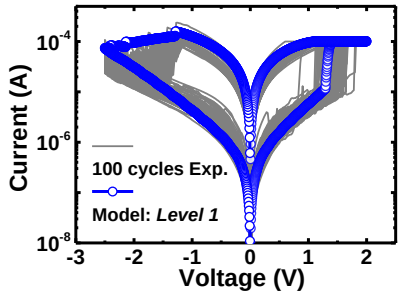

Fig. 2. Measured and modeled DC I-V characteristics of HfO_x/TiO_x RRAM using *Level 1* model.

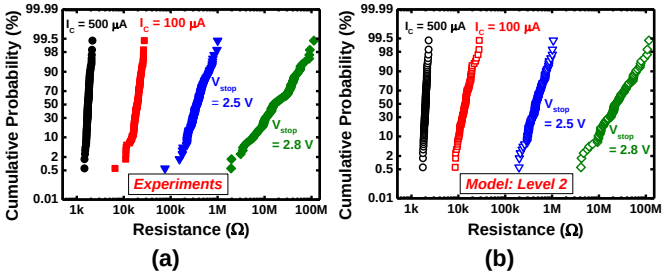

Fig. 3. Measured (a) and modeled (b) four-level resistance distributions of RRAM using *Level 2* model. I_C is compliance current controlling different R_{LRS} levels, while V_{stop} is the maximum voltage during RESET controlling R_{HRS} levels.

variability of the gap distance, CF length, and CF radius due to the stochastic nature of the processes of formation/migration/dissolution of the CF results in a statistical distribution of HRS resistance (R_{HRS}) and LRS resistance (R_{LRS}) after the write operations [53]-[55]. To model such characteristics, randomness is added to the CF geometry during resistive switching process in the *Level 2* model. The CF variables g and r are determined by:

$$g = \int (dg/dt + \delta g \times \chi(t))dt \quad (3)$$

$$r = \int (dr/dt + \delta r \times \chi(t))dt \quad (4)$$

where $\chi(t)$ is a zero-mean Gaussian sequence with a root mean square of unity. δ_g and δ_r are amplitude fitting parameters. Fig. 3(a) and Fig. 3(b) show the measured and modeled resistance distributions, respectively, obtained under various programming conditions. The compliance current (I_C) used for SET impacts the CF lateral growth, leading to lower R_{LRS} levels for larger compliance currents. The maximum voltage during RESET (V_{stop}) determines the final gap distance or the CF length, resulting in a higher R_{HRS} for a larger maximum voltage. Such variations are reproduced using the *Level 2* model, which dynamically captures two-dimensional CF evolution in a formulation that includes CF geometry variability.

For a SET operation of a filamentary RRAM device, when the applied voltage is lower than the nominal V_{SET} threshold (median threshold voltage extracted from DC I-V curves), switching from HRS to LRS becomes a stochastic process, which can be characterized by a SET probability (P_{SET}). Such sub-threshold stochastic behaviors of RRAM from cycle to cycle and from device to device are important for the circuit design in the low-voltage regime, where lowering voltage

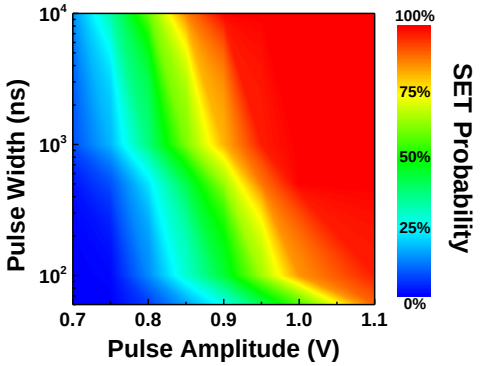

Fig. 4. Measured SET probability (P_{SET}) map as a function of applied pulse amplitude and pulse width on a typical device. Pulse rise/fall times are 10% of pulse width.

reduces both static/resistive (~$t \times V^2/R$) and dynamic (~CV^2) energy consumption. They may also inspire unconventional applications of RRAM in certain cases such as randomness generator, which can be useful for various scenarios. We characterize the stochastic SET behaviors on HfO_x RRAM [56] by applying 100 cycles of SET-read-RESET operations on a typical device and calculating a P_{SET} value. In this way, a complete map of P_{SET} is obtained as a function of applied pulse amplitude and pulse width, as shown in Fig. 4. It is observed that for certain P_{SET} values obtained from an individual RRAM cell, the pulse amplitude and pulse width follows a nonlinear voltage-time relationship in the sub-threshold regime. This suggests that a large design and optimization space exists in the low-voltage regime. To capture the characteristics in a model (*Level 3*), we incorporate the cycle-to-cycle (C2C) and device-to-device (D2D) stochasticity into the V_O activation energy (E_a) and the oxygen ion (O^{2-}) hopping barrier (E_h). Previous experimental evidence has implied that C2C and D2D variations are mathematically equivalent [54]. Therefore, such add-on stochasticity for the key energy barriers for cycle-to-cycle operations is also valid for variations across different devices. On top of *Level 1* and *Level 2*, *Level 3* can be described by a Monte Carlo approach in SPICE:

$$\begin{cases} g = \int \left(dg/dt(E_a, E_h)\right) dt \\ E_a(n) = E_{a0} + \delta E_a \\ E_h(n) = E_{h0} + \delta E_h \end{cases} \quad (5)$$

where E_a is associated with the SET process and E_h is associated with the RESET process. Since the condition of the filamentary switching region varies from cycle to cycle, the energy barriers are effective values that change from cycle to cycle, instead of fixed ones. E_a and E_h are sampled from normal distributions during the Monte Carlo simulations of CF evolution. To reproduce stochastic SET behaviors, C2C and D2D measurements are performed on RRAM devices using three pulse conditions selected from Fig. 4 that result in 50% P_{SET}, while the stochasticity of E_a is turned on in the *Level 3* model. Using the stochasticity incorporated into the energy barriers, the Monte Carlo simulations using *Level 3* model are able to reproduce the experimental observations under different pulse conditions, as indicated in Fig. 5. Although the three pulse conditions are chosen to achieve 50% P_{SET}, there exists

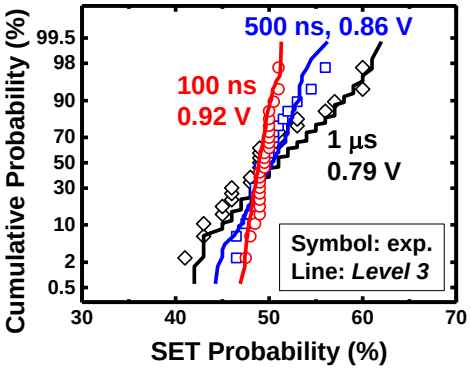

Fig. 5. Measured and modeled P_{SET} values obtained from multiple RRAM devices under stochastic SET mode, using three different pulse conditions extracted from Fig. 5. Each symbol represents one series of C2C measurements to obtain a P_{SET} value. *Level 3* model well reproduces the stochastic behaviors.

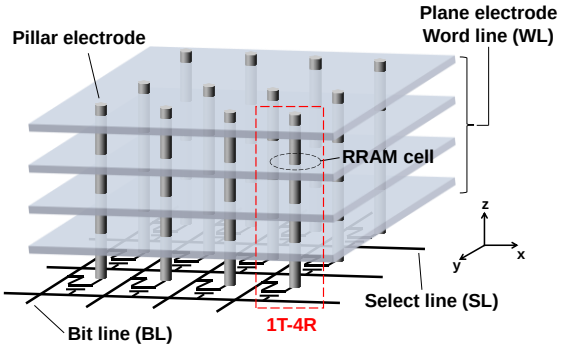

Fig. 6. Schematic of a 3D vertical RRAM array with select transistors for in-memory computing.

D2D variations of P_{SET} values collected from different RRAM devices. As exhibited by both experimental and modeling data, shorter pulses produce tighter distributions of P_{SET} centering around 50% across multiple RRAM devices. The P_{SET} values measured from multiple devices (D2D measurements) show relatively tight distributions as compared to C2C variability; C2C variability is large under weak programming conditions. This hierarchical RRAM SPICE model with a more accurate representation of resistive device properties can facilitate RRAM circuit designs that may capitalize on the inherent RRAM physics.

IV. Logic Operations within 3D RRAM

In this section, we describe the design principles and experimental demonstrations of in-memory logic operations with 1T-4R 3D vertical RRAM. Compared with previous studies [17], [39]-[46], our design methodology features the following aspects:

(1) 3D RRAM device/circuit structure provides cost-effective, high-density data storage/manipulation capabilities over conventional 2D RRAM structures;

(2) Three in-memory operation schemes are provided addressing different computation needs: half-V_{DD} scheme, V_{DD}/GND scheme, and 3D-LUT scheme;

(3) operations are non-volatile, cascadable, and free from destructive read;

(4) reconfigurable LUTs are formed in 3D memories during computation to help mitigate endurance limitations.

A. Three In-Memory Operation Schemes

Fig. 6 shows the schematic of a 3D vertical RRAM array with select transistors, where RRAM cells are located between vertical pillar electrodes and horizontal plane electrodes and can be individually randomly accessed by addressing the corresponding word line (WL), bit line (BL), and select line (SL). Three in-memory operation schemes for implementing logic functions are developed for the 3D vertical RRAM. The four-layer HfO_x 3D RRAM/FinFET devices reported in [56] are used for early-stage experimental verification. Logic variables are initialized and stored in 3D RRAM in the non-volatile fashion, where HRS corresponds to bit '0' and LRS corresponds to bit '1'. Executing Boolean logic functions is essentially a mapping from inputs to an output. To achieve the initialization and mapping operations, three operation schemes are designed, as illustrated in Fig. 7 for a 2-input logic function. In the 4-layer 3D RRAM structure, *A* and *B* store logic inputs, and *C* and *D* are logic outputs. The input values (*A* and *B*) are loaded in via conventional SET/RESET operations on RRAM cells [57].

In the half-V_{DD} scheme, the integrated transistor is turned on and biased in the linear region. Meanwhile, a pair of V_{DD} and $V_{DD}/2$ biases are applied on the WL plane electrodes of two RRAM cells, e.g., *A* and *C*, which share the same vertical pillar electrode. The output cell *C* is initialized to '1' via a SET operation. As shown in Fig. 7(a), the biased RRAM cells and the linear-region transistor together form a voltage dividing structure. As a result, the voltage across cell *C* (V_C) is dependent on the logic input or the resistance state of cell *A*. If bit '1' is written into *A*, voltage on the pillar electrode (V_P) will be pulled up due to the voltage dividing with the linear-region transistor. Since $V_C = V_{DD} - V_P$, it leaves insufficient voltage for a write operation on *C*. In contrast, if bit '0' is written into *A*, V_P will be pulled down as the high resistance of *A* will 'cut off' the current path from $V_{DD}/2$. Thus, when *A* is a "1", a sufficient voltage is generated across *C* to change the logic state. Additionally, a variant of half-V_{DD} scheme can be used, as shown in Fig. 7(b). In this case, V_{DD} is applied to the drain of the transistor, $V_{DD}/2$ is applied to *A*, and *C* is grounded. In this variant, the *INV* function and *IMP* logic can be realized more efficiently. It is worth noting that *C* can also serve as an input either for a 3-input logic function, or for another cascading function that takes the *C* output as the new input.

In the V_{DD}/GND scheme, a different bias condition is employed on the same structure, with the same purpose of triggering the logic mapping from inputs to an output. As shown in Fig. 7(c), the select transistor is turned off, and a pair of V_{DD} and *GND* are applied to the WL plane electrodes of two RRAM cells, e.g., *B* and *D*. The two biased RRAM cells form an in-series structure due to the common vertical pillar. This operation mode with two RRAM cells in series is different from the complementary switches [43], as it does not suffer from the destructive read issue that complementary switches have. In this scheme, all the logic outputs such as *D* are initialized to '0'

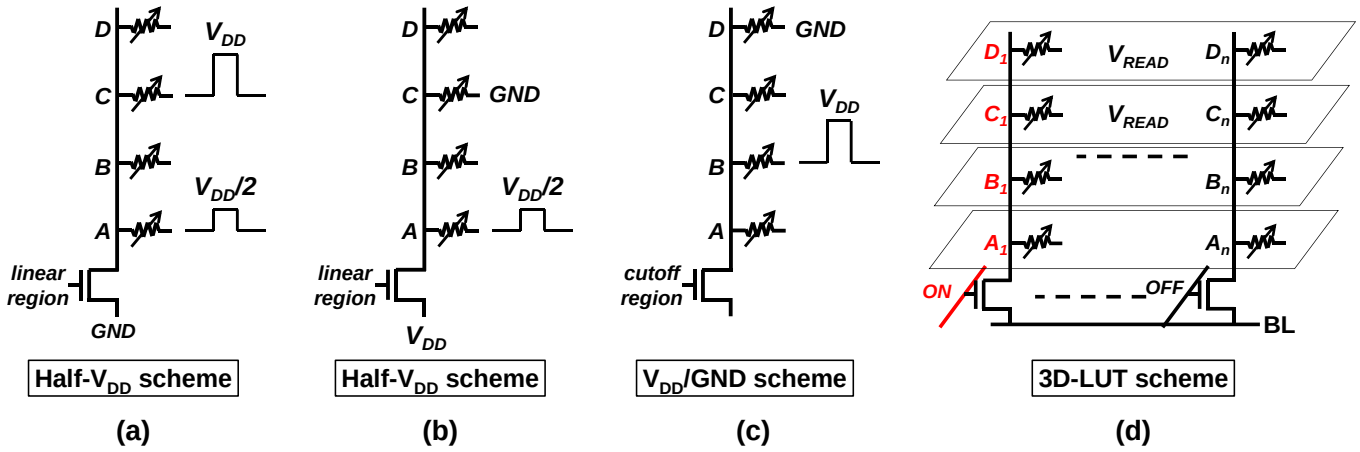

Fig. 7. Illustration of three in-memory operation schemes on the 3D vertical RRAM structure, where logic inputs (*A*, *B*) and outputs (*C*, *D*) are initialized and stored in RRAM cells. (a) In the half-V_{DD} scheme, a pair of V_{DD} and $V_{DD}/2$ biases are applied to two target RRAM cells (e.g., *A*, *C*), and the select transistor is biased in the linear region. Plane electrodes of the rest two RRAM cells (*B*, *D*) are floating. (b) A variant of half-V_{DD} scheme with a different bias design. (c) In the V_{DD}/GND scheme, a pair of *GND* and V_{DD} biases are applied to two target RRAM cells (e.g., *B*, *D*), while the select transistor is turned off. Plane electrodes of the rest two RRAM cells (*A*, *D*) are floating. (d) In the 3D-LUT scheme, multiple vertical pillars store different input and output combinations after programming with the first two schemes. Subsequent logic inputs are decoded to turn on a transistor selecting the correct vertical pillar, where the logic outputs are stored in upper layers and are ready for readout.

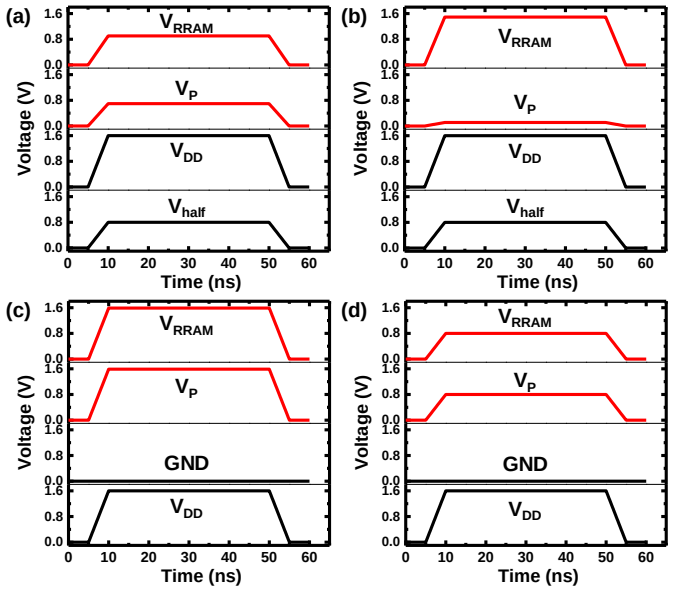

Fig. 8. Timing diagrams of (a) (b) half-V_{DD} scheme and (c) (d) V_{DD}/GND scheme from SPICE simulations. Input state is '1' for (a) (c), and is '0' for (b) (d). V_P is the voltage on the vertical pillar. V_{RRAM} is the resulting voltage across the output RRAM cell.

first. If $B = 0$, the voltage division between two RRAM cells alone will not SET or RESET either of the RRAM cells. However, if $B = 1$, most of V_{DD} will be dropped on the pillar electrode of cell *D* due to voltage division. In this way, a SET operation is performed on *D*, triggering the logic state transition from '0' to '1'. Timing diagrams of half-V_{DD} and V_{DD}/GND schemes are shown in Fig. 8. For the half-V_{DD} scheme (Fig. 8(a) and Fig. 8(b)), different input cell states result in different pillar voltage (V_P), and thereby different voltage across the output RRAM cell (V_{RRAM}). Similarly, for the V_{DD}/GND scheme (Fig. 8(c) and Fig. 8(d)), it is shown that V_{RRAM} has dependency on the input cell state in this common-pillar vertical structure.

In the 3D-LUT scheme, RRAM cells along the 3D vertical pillars are programmed to represent various logic input/output

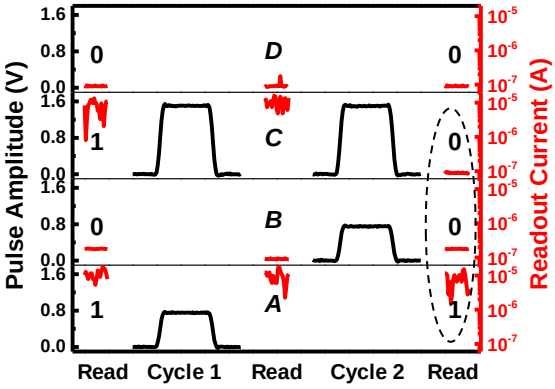

Fig. 9. Experimental demonstration of AND logic ($C = AB$) with 1T-4R 3D RRAMs, where two cycles of half-V_{DD} scheme are employed after the initialization cycle (3 cycles in total). The readout states of RRAM cells including intermediate states indicate the correct implementation of $C = AB$.

data. Therefore, for the mostly-read logic activities, logic inputs are decoded, and the corresponding transistor is turned on to select the correct vertical pillar, where the logic outputs are already stored in upper layers. Read voltage (V_{READ}) is applied to the WLs for logic evaluation, which aligns with the readout operations. As illustrated in Fig. 7(d), in this 3D-LUT structure, a certain logic function block with different input/output combinations is pre-programmed on multiple vertical pillars, which share the same BL. Meanwhile, other logic function blocks share the other BLs. The 3D-LUT scheme minimizes the write cycles on RRAM cells, which can greatly alleviate the endurance requirement.

B. Experimental Demonstrations

To verify the in-memory operation schemes and demonstrate logic primitives, electrical measurements are conducted on 1T-4R 3D RRAMs, using Keithley 4200 semiconductor characterization system with pulse measurement units. Fig. 9 shows the measured waveforms during the execution of AND logic. There are 3 cycles in total. *A* and *B* are the logic inputs

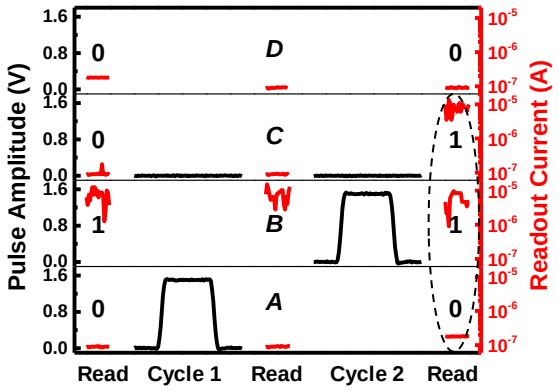

Fig. 10. Experimental demonstration of OR logic ($C = A + B$) with 1T-4R 3D RRAMs, where two cycles of V_{DD}/GND scheme are employed after the initialization cycle (3 cycles in total). The readout states of RRAM cells including intermediate states indicate the correct implementation of $C = A + B$.

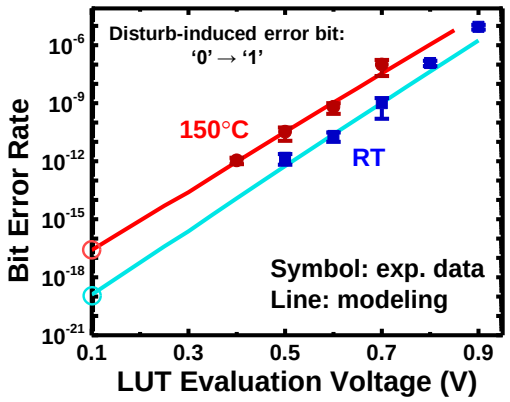

Fig. 11. Measured and modeled bit error rates (BER) as a function of evaluation voltage and operating temperature in the 3D-LUT scheme. Long-term testing is performed on 3D RRAMs that store logic outputs. Error bits are induced by undesired disturb on these RRAM cells during LUT evaluation.

and C is the logic output (initialized to '1' by forming or SET operation during the initialization cycle). Cell D is not used in this case. Then, two cycles of half-V_{DD} scheme are applied on the 1T-4R structure, addressing three RRAM devices. $V_{DD}/2$ pulses are applied on cell A and B during the two cycles, respectively, while the V_{DD} pulses are applied only on cell C. The pulse width is set to be 200 ns. These two cycles of pulses trigger the switching of C from '1' to '0', as indicated by the measured readout current, and thereby programmed the output C to yield an AND operation. Logic inputs and outputs are memorized in the non-volatile manner. A & B can be set to other input combinations, yet the AND operation is always ensured through two cycles of half-V_{DD} scheme. Half-V_{DD} scheme is suitable for AND-rich computations. As mentioned before, a variant of the half-V_{DD} scheme can be used for INV-rich computations, where the drain of the transistor is biased with V_{DD}, the input cell is biased with $V_{DD}/2$, and the output cell is grounded. Such scheme is similar to the IMP-based INV operation [17], [39]. Fig. 10 shows the measured waveforms during the execution of OR logic. There are 3 write cycles in total. $A = 0$ & $B = 1$ are initialized as the inputs and C is initialized to '0' during the initialization cycle. Here, the V_{DD}/GND scheme is employed for two cycles on the

Table I
DESIGN CHOICES OF DIFFERENT OPERATION SCHEMES

Logic / Scheme	AND-rich	OR-rich	INV IMP	Cascaded logic
Half-V_{DD}	■		■	■
V_{DD}/GND		■		■
3D-LUT	■	■	■	

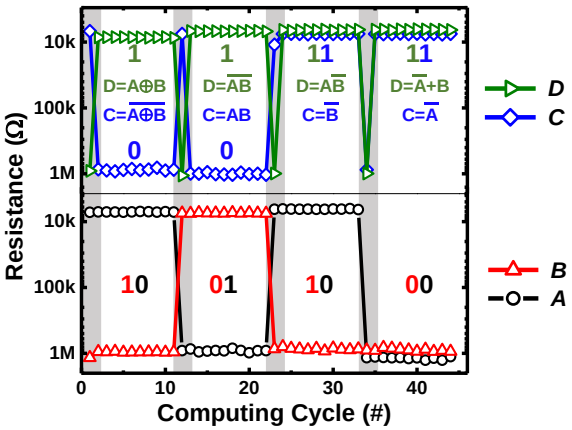

Fig. 12. Experimental demonstration of reconfigurable in-memory computing enabled by integrating all the three operation schemes. Resistance (logic) states of A to D in 1T-4R 3D RRAMs are measured during the computing. For the computing cycles in each of the gray regions, a new function is performed (as indicated by Boolean expressions) using half-V_{DD} scheme and V_{DD}/GND scheme. C/D are initialized and A/B are intentionally set to different input combinations. For the rest of computing cycles, 3D-LUT scheme is employed for logic evaluations. Eight different functions with various input combinations are implemented and measured correctly.

1T-4R 3D RRAM. The measured readout after the two-cycle operation indicate that a correct OR operation is performed, where $C = A + B$ after being switched from '0' to '1'. V_{DD}/GND scheme is convenient for OR-rich computations for a complex Boolean expression. Here we show that the developed in-memory operation schemes are available for both AND-rich and OR-rich logic.

The 3D-LUT scheme is designed to hold the logic data in a non-volatile manner for mostly-read use cases. Thus, it is important to scrutinize the reliability. Here, the reliability of the 3D-LUT scheme is examined by performing logic evaluations and measuring bit error rates (BER) on 3D RRAMs. Specifically, in the worst-case scenario, there can be unintentional resistive switching from HRS ('0') to LRS ('1') during read operations, and error bits can be induced in logic outputs by such disturb events. Long-term electrical measurements are performed for 3D-LUT evaluations and obtain the BER statistics. Fig. 11 shows the measured and modeled data of output BER, as a function of evaluation voltage and operating temperature. The 3D-LUT scheme is read-dominant, and therefore is immune to switching-induced variations. The results also indicate that the 3D-LUT scheme is robust over a wide range of operating voltages and temperatures. Furthermore, a linear reduction of LUT

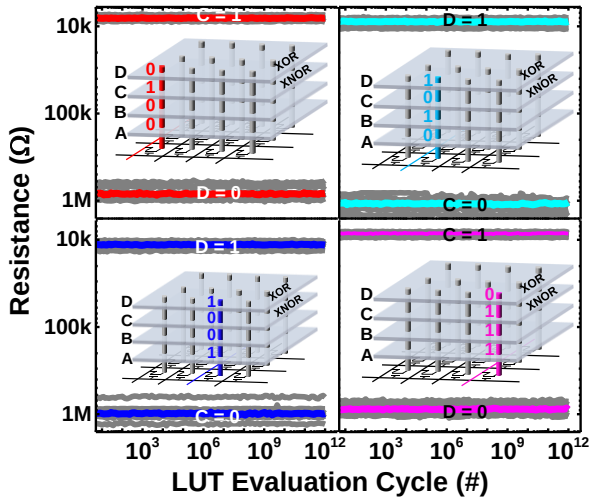

Fig. 13. RRAM-based multiplication kernel is implemented with 3D-LUT scheme. Measured data up to 10^{12} cycles indicate the correct and robust XOR evaluations of different input/output combinations. Gray lines in each subplot represent 10 independent programming events of 3D LUTs and the colored line is the median thereof.

evaluation voltage decreases the output BER exponentially, which is well reproduced by the RRAM SPICE model. This is because the disturb error bits originate from the nonlinear accumulation effect of CF growth. Such nonlinearity is manifest in the experimental data presented in Fig. 4 as well. When the LUT evaluation voltage is set to be 0.1 V (typical voltage for RRAM read operations), the projected BER at 150°C is below 10^{-15}.

Table I summarizes the design choices of different operation schemes discussed above for different cases, as a guideline for breaking down and scheduling the computation tasks for memory-centric systems. The green boxes in Table I means that the operation scheme is inherently more suitable or efficient for certain type of logic. For example, OR-rich logic can be straightforward by employing V_{DD}/GND scheme. Both half-V_{DD} scheme and V_{DD}/GND scheme (or combination of them) support cascading logic for multiple stages, which is enabled by operating on additional RRAM cells along the vertical pillar. After programming 3D RRAM pillars in an array fashion, 3D-LUT scheme leads to fast and efficient logic evaluations. It does not directly support cascading logic on the fly as the stored LUTs are static. However, the contents of the LUT can be easily reconfigured with new programming cycles. Here, reconfigurable logic is demonstrated for the use of all three operation schemes in the same structure, as shown in Fig. 12. Logic states of *A* to *D* in 1T-4R 3D RRAMs are monitored during the computation. The computing cycle marked by the gray background in Fig. 12 performs a specific logic operation with certain logic inputs using half-V_{DD} scheme and also the V_{DD}/GND scheme. For example, the first XOR/XNOR functions are implemented with a combination of AND, OR, IMP, and INV on the 1T-4R 3D structures. After programming, 3D-LUT scheme is employed for the mostly-read logic evaluations, which is verified by the readout resistance values of multiple cycles. The programmed LUTs can also be reconfigured for new logic functions with new input data, using

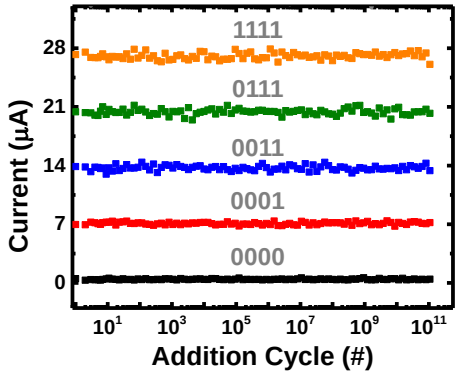

Fig. 14. Measured in-memory addition (accumulation) results on 4-layer 3D RRAMs up to 10^{11} read cycles.

a combination of the three operation schemes. Note that the operation schemes that involve resistive switching processes should be aware of the intrinsic variability of RRAM such as resistance and switching voltage variations. It is therefore desirable that device stack engineering and material optimization can improve device uniformity. However, it is also worth noting that the demonstrated logic operations on 3D RRAM are more "digital like", and a reasonable ON/OFF resistance ratio (e.g., 50~100) can help to yield relatively reliable outputs. In the next section, we will also discuss the interaction between new computation models/algorithms and system design, where the RRAM-centric systems can be resilient to even hard memory errors (endurance failure).

V. Hyperdimensional Computing: A Case Study

If the key computation primitives can be identified, novel computation models and algorithms can be efficiently mapped onto RRAM-centric computing systems. As a case study, we evaluate the use of 3D RRAM-centric architecture for hyperdimensional (HD) computing, a neural-inspired computation model that represents and processes information in high dimensionality [23]. Instead of computing with numbers, HD computing represents and processes data with high-dimensional (e.g., kilo-bit length) vectors, inspired by the remarkable correspondence of mathematical properties of high-dimensional space to human's perception, memory, and cognition [23]. Regardless of specific applications (e.g., language recognition, scene understanding) and algorithms (e.g., random indexing), HD computing requires three key operations on HD vectors: multiplication, addition, and permutation (MAP). From a hardware perspective, these vector operations are all memory-intensive. Therefore, it is desirable to perform these computation kernels for MAP operations native within the memory array without moving the data in and out of the memory array. Here we demonstrate the use of 3D RRAM to implement the MAP kernels and compare the design of an HD language recognition system with conventional CMOS logic using system-level and circuit-level simulations.

A. RRAM-Based MAP Kernels

The MAP kernels can be efficiently constructed for a 3D RRAM using the three in-memory operation schemes described

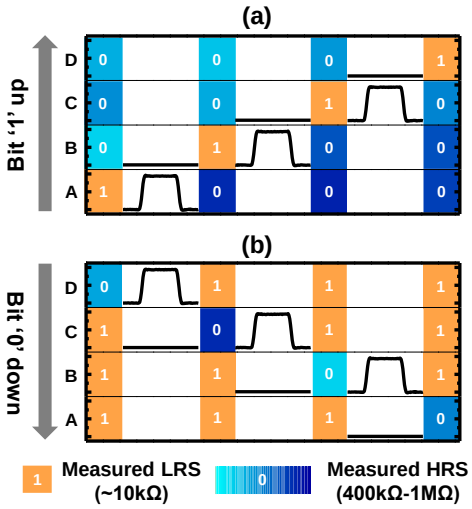

Fig. 15. Permutation is implemented via in-memory bit transfer using the V_{DD}/GND scheme. (a) Measured resistance evolution of RRAM cells moving '1' up from *A* to *D* vertically. (b) Measured resistance evolution of RRAM cells moving '0' down from *D* to *A*. After each cycle of transfer, the state of the 'source' cell is switched intentionally for better illustrating the transfer path.

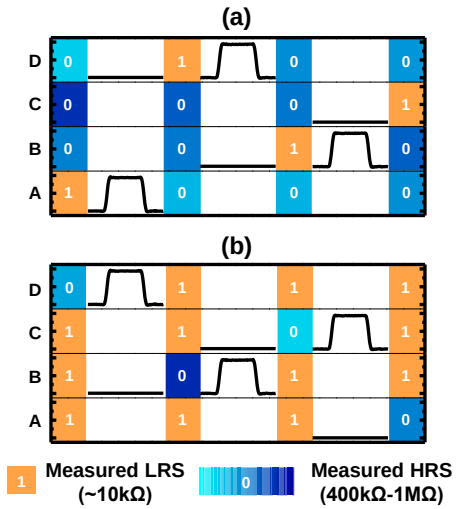

Fig. 16. Arbitrary in-memory bit transfer of (a) '1' and (b) '0' using the V_{DD}/GND scheme. Color maps represent the measured resistance states of the 4-layer 3D RRAM.

earlier. The HD vector operations are then performed directly in the 3D memory. The multiplication between two HD vectors consisting of binary bits is essentially a bit-wise XOR [24], yielding a new binary HD vector with the same dimensionality or vector length. Using the half-V_{DD} scheme and the V_{DD}/GND scheme described in Section IV, XOR logic can be programmed in the 3D RRAM. The 3D-LUT scheme is employed afterwards for read operations, as shown in Fig. 12. Measurements of XOR evaluations on different vertical pillars last for 10^{12} cycles and are repeated, showing the correct and reliable functionality. The addition operation in HD computing is bit summation or accumulation. It is naturally enabled by the 3D configuration of the 3D RRAM array, via current summing along the shared vertical pillars of 3D RRAMs. Fig. 14 shows the measured addition outputs up to 10^{11} recurrent cycles for various 4-bit vectors, which are written into 1T-4R 3D

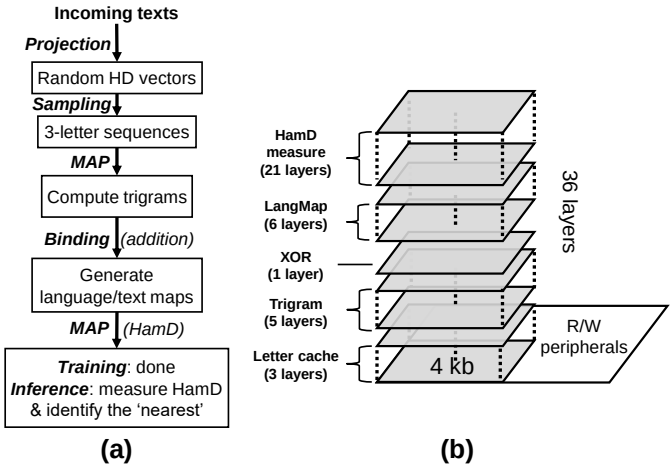

Fig. 17. (a) Algorithm pipeline for HD language recognition. (b) Schematic of a 36-layer 3D RRAM array, where different stages in the algorithm are implemented in certain layers based on simulations with device-level experimental data and compact models incorporated.

RRAMs. As long as a reasonable memory window (e.g., $R_{HRS}/R_{LRS} > 10$) is maintained, distinct levels of correct outputs can be obtained from the addition kernel. The third operation, i.e., permutation, is shifting or shuffling of bits, and can be implemented through in-memory bit transfer following the V_{DD}/GND scheme. As illustrated in Fig. 15, cell *A* is initialized to '1' via a SET operation and the rest of three cells are initialized to '0' via RESET operations. Then, the V_{DD}/GND scheme applied to a pair of RRAM locations and triggers the direct data copy from one cell to another without extra readout or write-back operations. The measured resistance state evolutions of RRAM cells illustrate the processes of moving bit '1' and bit '0' up (Fig. 15 (a)) and down (Fig. 15 (b)) in the 3D vertical structure. Moreover, arbitrary in-memory bit transfer is supported as well using the V_{DD}/GND scheme. Fig. 16 shows the measured resistance state evolution of 4-layer 3D RRAM for the transfer of bit '1' (Fig. 16(a)) and bit '0' (Fig. 16(b)) in two arbitrary orders.

B. System-Level Evaluations

To evaluate in-memory HD computing systems with RRAM-based MAP kernels, a language recognition application (recognizing/identifying a given sentence as one of the 21 European languages) is chosen and system-level evaluations are performed based on the developed RRAM model and simulation tools. Fig. 17(a) illustrates the algorithm pipeline and Fig. 17(b) shows how the algorithm stages are implemented in multiple blocks in a 36-layer 3D array. Results are obtained from simulations with device-level experimental data and compact models incorporated. Details of the algorithm pipeline can be found in [24], where a digital CMOS implementation was reported. For training, 21 sample texts are taken from Wortschatz Corpora [58]. The procedure starts with taking input letters sampled from a sample text and encoding them into HD vectors (vector length = 1 kbit). The random distribution of '1's and '0's in HD vectors is achieved by utilizing the stochastic SET properties of RRAM, as characterized and modeled in Section III. RRAM cells are

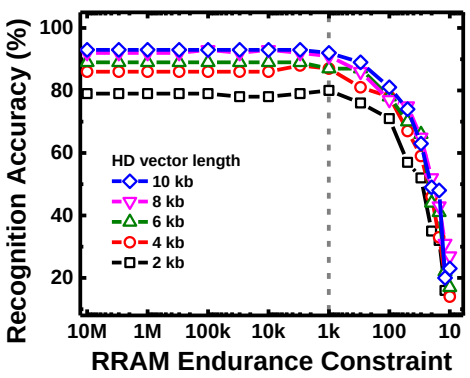

Fig. 18. Simulated recognition accuracy of the HD language recognition system as a function of RRAM endurance and HD vector length.

programmed to achieve ~50% P_{SET} with 4% D2D variations, which results in a balanced but random '1'/'0' distribution in a HD vector. With a sliding window of 3 letters (trigrams), the HD vectors are then multiplied (XORed), permuted, and added to generate trigram vectors. A single 1-kb-length language vector is produced after addition of all trigram vectors from a single sample text. The language vectors generated from 26 sample texts are stored in 6 layers (*"LangMap"*) of a 3D RRAM array (Fig. 17(b)). For inference, 21,000 unseen sentences (1,000 sentences/language) are taken from Europarl Parallel Corpus [59]. The encoded test vectors go through the same pipeline yielding test vectors for the test sentences. To identify an unseen sentence to a language, Hamming distances between learned language vectors and the test vector are measured through XOR and addition operations, and the language vector with the smallest Hamming distance points to the identified language. Simulations are performed to evaluate the system recognition accuracy while considering the constraint of RRAM endurance as write operations on RRAM cells are involved. Hard stuck-at errors are assumed when endurance failures (stuck at '1' or stuck at '0') occur on RRAM cells during the entire simulation. As shown in Fig. 18, under various levels of RRAM endurance constraints from 10^7 cycles down to 10^3 cycles, the HD system is resilient to hard errors and the simulated recognition accuracy is maintained. This error resilience originates from the properties of the HD computation model. HD representation is holographic, which means the encoded or learned information is equally distributed among all the bits in an HD vector. Thus, certain amount of error bits can be tolerant. To compare with conventional CMOS design, the 3D RRAM design with peripheral circuits is benchmarked with a digital CMOS design [24], using the same 28-nm PDK. Owing to the compact implementation of MAP kernels and the 3D architecture, greater than 400× area savings and 2× lower energy consumption are obtained using the RRAM-centric design, with 3% drop in recognition accuracy [25].

A variety of essential computation kernels can be further explored with 3D RRAM technologies. For instance, the addition kernel in this work can be regarded as a special case of dot product. Utilizing the efficient current summing in 3D RRAM, a dot product kernel can be built for a class of machine learning algorithms. With parallelism, it can be extended to a matrix-vector multiplication kernel.

VI. Conclusion

This work presents a design and modeling example of an RRAM-centric computing system, covering various design aspects including device compact modeling, novel 3D architecture and operation schemes, experimental demonstration of key computational kernels, and algorithm-hardware co-design. The RRAM SPICE model with three hierarchical levels of physics realism can facilitate energy-efficient RRAM circuit design with a more accurate representation of RRAM behaviors. In-memory logic operations with 3D RRAMs are experimentally demonstrated; they have the benefit of having a set of flexible write and read schemes that fully utilizes inherent 3D RRAM properties. Furthermore, co-designing RRAM-centric computing systems with HD computing model is explored. Results suggest that RRAM-centric cognitive systems are resilient to hard errors induced by endurance failures, making various types of RRAM feasible for memory-centric HD computing systems.

Acknowledgement

The authors would like to thank Dr. Kai-Shin Li, Dr. Chang-Hsien Lin, Dr. Juo-Luen Hsu, Dr. Wen-Cheng Chiu, Dr. Min-Cheng Chen, Dr. Tsung-Ta Wu, Dr. Jia-Min Shieh, and Dr. Wen-Kuan Yeh from National Nano Device Laboratories, Taiwan, for device fabrication. The authors also appreciate Dr. Abbas Rahimi, Mr. Miles Rusch, Dr. Pentti Kanerva, and Prof. Jan Rabaey from UC Berkeley for the help with system simulations and fruitful discussions.

References

[1] Atzori, Luigi, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787-2805, Oct. 2010.

[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] 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.

[4] B. Rogers, A. Krishna, G. Bell, K. Vu, X. Jiang, Y. Solihin, "Scaling the bandwidth wall: challenges in and avenues for CMP scaling," *ACM SIGARCH Computer Architecture News*, vol. 37, no. 3, pp. 371-382, June 2009.

[5] I. G. Baek, C. Park, H. Ju, D. J. Seong, H. S. Ahn, J. H. Kim, M. K. Yang, S. H. Song, E. M. Kim, S. O. Park, C. H. Park *et al.*, "Realization of vertical resistive memory (VRRAM) using cost effective 3D process," *IEEE International Electron Devices Meeting (IEDM)*, 2011, pp. 737-740.

[6] H.-Y. Chen, S. Yu, B. Gao, P. Huang, J. Kang, and H.-S. Philip Wong, "HfO_x based vertical RRAM for cost-effective 3D cross-point architecture without cell selector," in *IEEE International Electron Devices Meeting (IEDM)*, 2012, pp. 497–500.

[7] M. M. Shulaker, T. F Wu, A. Pal, L. Zhao, Y. Nishi, K. Saraswat, H.-S. P. Wong, and S. Mitra, "Monolithic 3D integration of logic and memory: Carbon nanotube FETs, resistive RAM, and silicon FETs," *IEEE International Electron Devices Meeting (IEDM)*, 2014, pp. 1-4.

[8] S. Jeloka, N.B. Akesh, D. Sylvester, and D. Blaauw, "A 28 nm configurable memory (TCAM/BCAM/SRAM) using push-rule 6T bit cell enabling logic-in-memory," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 4, pp. 1009-1021, April 2016.

[9] J. Zhang, Z. Wang, and N. Verma. "A machine-learning classifier implemented in a standard 6T SRAM array," *IEEE Symp. VLSI Circuits (VLSI-Circuits)*, 2016, pp. 1-2.

[10] M. Kang, M. S., Keel, N. R. Shanbhag, S. Eilert, and K. Curewitz, "An energy-efficient VLSI architecture for pattern recognition via deep embedding of computation in SRAM," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 8326-8330.

[11] X. Guo, E. Ipek, and T. Soyata. "Resistive computation: avoiding the power wall with low-leakage, STT-MRAM based computing," *ACM SIGARCH Computer Architecture News*, vol. 38, no. 3, pp. 371-382, June 2010.

[12] D. E. Nikonov and I. A. Young, "Overview of beyond-CMOS devices and a uniform methodology for their benchmarking," *Proc. IEEE*, vol. 101, no. 12, pp. 2498–2533, Dec. 2013.

[13] W. Zhao, D. Ravelosona, J. Klein, and C. Chappert, "Domain wall shift register-based reconfigurable logic," *IEEE Trans. Magn.*, vol. 47, no. 10, pp. 2966–2969, Oct. 2011.

[14] B. Behin-Aein, D. Datta, S. Salahuddin, and S. Datta, "Proposal for an all-spin logic device with built-in memory," *Nature Nanotechnol*., vol. 5, pp. 266–270, Feb. 2010.

[15] M. Cassinerio, N. Ciocchini, and D. Ielmini, "Logic computation in phase change materials by threshold and memory switching," *Adv. Mater*., vol. 25, no. 41, pp. 5975–5980, 2013.

[16] T. Hasegawa, K. Terabe, T. Tsuruoka, and M. Aono "Atomic switch: atom/ion movement controlled devices for beyond von-Neumann computers," *Adv. Mater*., vol. 24, no. 2, pp. 252–267, Jan. 2012.

[17] J. Borghetti, G. S. Snider, P. J. Kuekes, J. J. Yang, D. R. Stewart, and R. S. Williams, "'Memristive' switches enable 'stateful' logic operations via material implication," *Nature*, vol. 464, no. 7290, pp. 873–876, 2010.

[18] S. Yu, X. Guan, and H.-S. P. Wong, "On the switching parameter variation of metal oxide RRAM—Part II: Model corroboration and device design strategy," *IEEE Trans. Electron Devices*, vol. 59, no. 4, pp. 1183–1188, Apr. 2012

[19] D. Ielmini, "Modeling the universal set/reset characteristics of bipolar RRAM by field-and temperature-driven filament growth," *IEEE Trans. Electron Devices*, vol. 58, no. 12, pp. 4309-4317, Dec. 2011.

[20] R. Degraeve, A. Fantini, S. Clima, B. Govoreanu, L. Goux, Y. Chen, D. Wouters, P. Roussel, G. Kar, G. Pourtois, S. Cosemans, *et al.*, "Dynamic 'hour glass' model for SET and RESET in HfO_2 RRAM," *Symp. VLSI Tech.*, 2012, pp. 75-76.

[21] P. Sheridan, K.-H. Kim, S. Gaba, T. Chang, L. Chen, and W. Lu, "Device and SPICE modeling of RRAM devices," *Nanoscale*, vol. 3, no. 9, pp. 3833–3840, 2011.

[22] S. Ambrogio, S. Balatti, D. Gilmer, D. Ielmini, "Analytical modeling of oxide-based bipolar resistive memories and complementary resistive switches," *IEEE Trans. Electron Devices*, vol. 61, no. 7, pp. 2378-2386, July 2014.

[23] P. Kanerva, "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors," *Cognitive Computation*, vol. 1, no. 2, pp. 139-159, June 2009.

[24] A. Rahimi, P. Kanerva, and J. M. Rabaey, "A robust and energy-efficient classifier using brain-inspired hyperdimensional computing," *International Symp. Low Power Electronics and Design (ISLPED)*, 2016, pp. 64-69.

[25] H. Li, T. F. Wu, A. Rahimi, K.-S. Li, M. Rusch, C.-H. Lin, J.-L. Hsu, M. M. Sabry, S. B. Eryilmaz, J. Sohn, W.-C. Chiu, M.-C. Chen, T.-T. Wu, J.-M. Shieh, W.-K. Yeh, J. M. Rabaey, S. Mitra, and H.-S. P. Wong, "Hyperdimensional computing with 3D VRRAM in-memory kernels: device-architecture co-design for energy-efficient, error-resilient language recognition," *IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 16-1.

[26] 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.

[27] H.-S. P. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu, P. Chen, B. Lee, F.-T. Chen, and M.-J. Tsai, "Metal–oxide RRAM," *Proc. IEEE*, vol. 100, no. 6, pp. 1951-1970, June 2012.

[28] W. Shen, C. Mei, Y. Chih, S. Sheu, M. Tsai, Y. King, C. Lin, *et al.*, "High-K metal gate contact RRAM (CRRAM) in pure 28nm CMOS logic process," *IEEE International Electron Devices Meeting (IEDM)*, 2012, pp. 31-36.

[29] W. Kim, S. Park, Z. Zhang, Y. Yang-Liauw, D. Sekar, H.-S. P. Wong, and S. S. Wong, "Forming-free nitrogen-doped AlO_x RRAM with sub-μA programming current," *Symp. VLSI Tech.*, 2011, pp. 1-2.

[30] H. Y. Lee, Y. S. Chen, P. S. Chen, P. Y. Gu, Y. Y. Hsu, S. M. Wang, W. H. Liu, C. H. Tsai, S. S. Sheu, P. C. Chiang, W. P. Lin *et al.*, "Evidence and solution of over-RESET problem for HfO_x based resistive memory with sub-ns switching speed and high endurance," *IEEE International Electron Devices Meeting (IEDM)*, 2010, pp. 1-4.

[31] B. Govoreanu, G. Kar, Y. Chen, V. Paraschiv, S. Kubicek, A. Fantini, I. Radu, L. Goux, S. Clima, R. Degraeve, N. Jossart, *et al.*, "10×10 nm^2 Hf/HfO_x crossbar resistive RAM with excellent performance, reliability and low-energy operation,", *IEEE International Electron Devices Meeting (IEDM)*, 2011, pp. 729-732.

[32] M.-J. Lee, C. Lee, D. Lee, S. Lee, M. Chang, J. Hur, Y. Kim, C. Kim, D. Seo, S. Seo, U. Chung, *et al.*, "A fast, high-endurance and scalable non-volatile memory device made from asymmetric Ta_2O_{5-x}/TaO_{2-x} bilayer structures," *Nature Mater.*, vol. 10, pp. 625–630, Aug. 2011.

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

[34] T.-Y. Liu *et al.*, "A 130.7 mm^2 2-layer 32 Gb ReRAM memory device in 24 nm technology," in *IEEE ISSCC Tech. Dig. Papers*, Feb. 2013, pp. 210–211.

[35] H. Pan, K. Huang, S. Chen, P. Peng, Z. Yang, C. Kuo, Y. Chih, Y. King, and C. Lin, "1Kbit FINFET dielectric (FIND) RRAM in pure 16nm FinFET CMOS logic process," *IEEE International Electron Devices Meeting (IEDM)*, 2015, pp. 10-5.

[36] B. Govoreanu, L. Di Piazza, J. Ma, T. Conard, A. Vanleenhove, A. Belmonte, D. Radisic, M. Popovici, A. Velea, A. Redolfi, O. Richard, *et al.*, "Advanced a-VMCO resistive switching memory through inner interface engineering with wide ($> 10^2$) on/off window, tunable μA-range switching current and excellent variability," *Symp. VLSI Tech.*, 2016, pp. 1-2.

[37] F. Hsueh, C. Shen, J. Shieh, K. Li, H. Chen, W. Huang, H. Wang, C. Yang, T. Hsieh, C. Lin, B. Chen, *et al.*, "First fully functionalized monolithic 3D+ IoT chip with 0.5 V light-electricity power management, 6.8 GHz wireless-communication VCO, and 4-layer vertical ReRAM," *IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 2-3.

[38] C. Ho, T.Y. Shen, P.Y. Hsu, S.C. Chang, S.Y. Wen, M.H. Lin, P.K. Wang, S.C. Liao, C.S. Chou, K.M. Peng, C.M. Wu, *et al.*, "Random soft error suppression by stoichiometric engineering: CMOS compatible and reliable 1Mb HfO_2-ReRAM with 2 extra masks for embedded IoT systems," *Symp. VLSI Tech.*, 2016, pp. 1-2.

[39] G. C. Adam, B. D. Hoskins, M. Prezioso, D. B. Strukov, "Optimized stateful material implication logic for three-dimensional data manipulation," Nano Research, vol.9, no. 12, pp. 3914-3923, Dec. 2016.

[40] P. Huang, J. Kang, Y. D. Zhao, S. Chen, R. Han, Z. Zhou, Z. Chen, W. Ma, M. Li, L. Liu, and X. Liu, "Reconfigurable nonvolatile logic operations in resistance switching crossbar array for large-scale circuits," *Adv. Mater.*, Sep. 2016.

[41] S. Kvatinsky, D. Belousov, S. Liman, G. Satat, N. Wald, E. G. Friedman, A. Kolodny, and U. C. Weiser, "MAGIC—Memristor-Aided Logic," *IEEE Trans. Circuits Syst. II: Express Briefs*, vol. 61, no. 11, pp. 895-899, Nov. 2014.

[42] M. F. Chang, S. M. Yang, C. C. Guo, T. C. Yang, C. J. Yeh, T. F. Chen, L. Y. Huang, S. S. Sheu, P. L. Tseng, Y. S. Chen, "Set-triggered-parallel-reset memristor logic for high-density heterogeneous-integration friendly normally off applications," *IEEE Trans. Circuits Syst. II: Express Briefs*, vol.62, no. 1, pp. 80-84, Jan. 2015.

[43] H. Manem, J. Rajendran, and G. S. Rose, "Stochastic gradient descent inspired training technique for a CMOS/nano memristive trainable threshold gate array," *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 59, no. 5, pp. 1051–1060, May 2012.

[44] E. Linn, R. Rosezin, S. Tappertzhofen, U. Bottger, and R. Waser, "Beyond von Neumann—logic operations in passive crossbar arrays alongside memory operations," Nanotechnology, vol. 23, no. 30, pp. 305205, Jul. 2012.

[45] S. Balatti, S. Ambrogio, and D. Ielmini, "Normally-off logic based on resistive switches—part I: logic gates," *IEEE Trans. Electron Devices*, vol. 62, no. 6, pp. 1831-1838, May 2015.

[46] S. Shirinzadeh, M. Soeken, P. E. Gaillardon, and R. Drechsler, "Fast logic synthesis for RRAM-based in-memory computing using majority-inverter graphs," *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2016, pp. 948-953.

[47] C. W. Hsu, I. T. Wang, C. L. Lo, M. C. Chiang, W. Y. Jang, C. H. Lin, T. H. Hou, "Self-rectifying bipolar TaOx/TiO2 RRAM with superior endurance over 10^{12} cycles for 3D high-density storage-class memory," *IEEE Symp. VLSI Tech.*, 2013, pp. 166-167.

[48] Y. Wu, H. Yi, Z. Zhang, Z. Jiang, J. Sohn, S. Wong, and H.-S. P. Wong, "First demonstration of RRAM patterned by block copolymer

self-assembly,". *EEE International Electron Devices Meeting (IEDM)*, 2013, pp. 20.8.

[49] Y.-Y. Chen *et al.*, "Balancing SET/RESET pulse for > 10^{10} endurance in HfO_2/Hf 1T1R bipolar RRAM," *IEEE Trans. Electron Devices*, vol. 59, no. 12, pp. 3243-3249, Dec. 2012.

[50] P. Huang, B. Chen, Y. Wang, F. Zhang, L. Shen, R. Liu, L. Zeng, G. Du, X. Zhang, B. Gao, J. Kang and X. Liu, "Analytic model of endurance degradation and its practical applications for operation scheme optimization in metal oxide based RRAM," *IEEE International Electron Devices Meeting (IEDM)*, 2013, pp. 22.5.

[51] H. Li, Z. Jiang, P. Huang, Y. Wu, H.-Y. Chen, B. Gao, X. Liu, J. Kang, and H.-S. P. Wong, "Variation-aware, reliability-emphasized design and optimization of RRAM using SPICE model," *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2015, pp. 1425-1430.

[52] Z. Jiang, Y. Wu, S. Yu, L. Yang, K. Song, Z. Karim, and H.-S. P. Wong, "A compact model for metal–oxide resistive random access memory with experiment verification," *IEEE Trans. Electron Devices*, vol. 63, no. 5, pp. 1884-1892, May 2016.

[53] X. Guan, S. Yu, and H.-S. P. Wong, "On the switching parameter variation of metal-oxide RRAM—Part I: physical modeling and simulation methodology," *IEEE Trans. Electron Devices*, vol. 59, no. 4, pp. 1172-1182, April 2012.

[54] A. Fantini, L. Goux, R. Degraeve, D. Wouters, N. Raghavan, G. Kar, A. Belmonte, Y. Y. Chen, B. Govoreanu, and M. Jurczak, "Intrinsic switching variability in HfO_2 RRAM," in *Proc. 5th IEEE IMW*, 2013, pp. 30–33.

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

[56] H. Li, K.-S. Li, C.-H. Lin, J.-L. Hsu, W.-C. Chiu, M.-C. Chen, T.-T. Wu, J. Sohn, S. B. Eryilmaz, J.-M. Shieh, W.-K. Yeh, and H.-S. P. Wong, "Four-layer 3D vertical RRAM integrated with FinFET as a versatile computing unit for brain-inspired cognitive information processing," *Symp. VLSI Technology (VLSI-T)*, 2016, pp. 1-2.

[57] B. Gao, B. Chen, R. Liu, F. Zhang, P. Huang, L. Liu, X. Liu, J. Kang, H.-Y. Chen, S. Yu, and H.-S.P. Wong, "3-D cross-point array operation on AlO_y /HfO_x -based vertical resistive switching memory," *IEEE Trans. Electron Devices*, vol. 61, no. 5, pp. 1377–1380, May 2014.

[58] U. Quasthoff, M. Richter, C. Biemann, "Corpus portal for search in monolingual corpora," *International Conference on Language Resources and Evaluation (LREC)*, 2006, p. 21.

[59] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," *MT Summit*, 2005, pp. 79-86.