Reconfigurable 2T2R ReRAM Architecture for Versatile Data Storage and Computing In-Memory

Yuzong Chen, Lu Lu, *Student Member, IEEE*, Bongjin Kim, *Member, IEEE*, and Tony Tae-Hyoung Kim, *Senior Member, IEEE*

***Abstract*— Nonvolatile memory (NVM)-based computing in-memory (CIM) is a promising solution to data-intensive applications. This work proposes a 2T2R resistive random access memory (ReRAM) architecture that supports three types of CIM operations: 1) ternary content addressable memory (TCAM); 2) logic in-memory (LiM) primitives and arithmetic blocks such as full adder (FA) and full subtractor; and 3) in-memory dot-product for neural networks. The proposed architecture allows the NVM operations in both 2T2R and conventional 1T1R configurations. The proposed LiM full adder (LiM-FA) improves the delay, the static power, and the dynamic power by 3.2×, 1.2×, and 1.6×, respectively, compared with state-of-the-art LiM-FAs. Furthermore, based on different optimization techniques and robustness analysis, a lower precharge voltage is set for each mode. This reduces the TCAM search energy and 1T1R ReRAM access energy by 1.6× and 1.14×, respectively, compared with the case without optimizations.**

***Index Terms*— Computing in-memory (CIM), nonvolatile memory (NVM), resistive random access memory (ReRAM), robustness analysis, ternary content addressable memory (TCAM).**

I. Introduction

As MORE attention in our daily life shifts toward emerging applications such as machine learning and big-data processing, existing computing paradigms face unprecedented challenges in executing required tasks with high energy efficiency. The scaling trend of CMOS performance has slowed down because of the power wall and slower voltage scaling. Moreover, traditional von-Neumann computing architecture suffers from long latency and high energy consumption because of expensive data movements between memory and arithmetic-logic units (ALUs). This latency and energy overheads become more severe as the memory hierarchy goes from a register file to cache, main memory (e.g., DRAM), and nonvolatile storage (e.g., FLASH). As shown in Fig. 1, a typical ALU operation (e.g., 32-bit integer addition) takes less than 1 ns and consumes less than 1 pJ while a data movement from a register file and cache takes comparable latency [1] but dissipates tens of pJ [2]. When considering off-chip memory, the latency and energy can increase to tens of ns [1] and a few nJ [3], respectively.

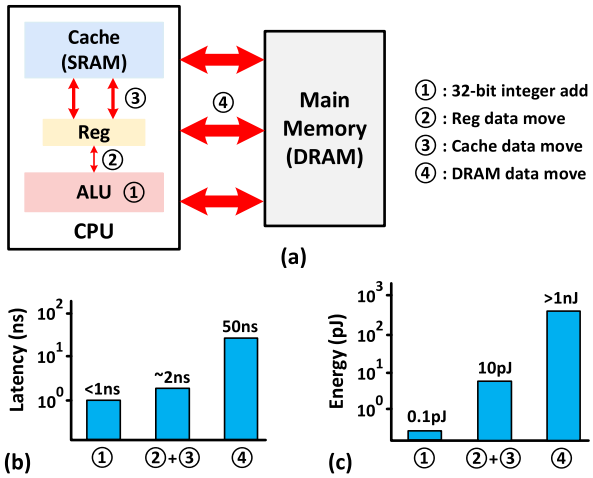


Fig. 1. Traditional von-Neumann architectures. (a) Block diagram, (b) latency, and (c) energy cost of a 32-bit integer ALU addition and different stages of data movements.

To continuously improve computing performance and energy efficiency, people have focused on beyond CMOS devices and beyond von-Neumann architectures. For example, resistive random access memory (ReRAM) [4] is an attractive nonvolatile memory (NVM) candidate for the next-generation storage system. It shows good read/write performance, low programming voltage, good scalability, and compatibility to the CMOS fabrication process. It also allows monolithic 3-D integration [5], [6] with logic devices to achieve high integration density and fine-grained connectivity between logic and memory circuits.

Meanwhile, beyond von-Neumann architecture like computing in-memory (CIM) is also under active research to tackle the expensive latency and energy cost associated with data movements. Various CIM works have been reported in the literature [7]–[20]. The ternary content addressable memory (TCAM) operations in [7] and [16]–[20] perform parallel bit-wise XOR/XNOR between the search key and the stored data to achieve fast search. The logic in-memory (LiM) operations in [7] and [10]–[12] realize Boolean logic functions between two or more words in/near memory without reading out operands and sending them to ALU. Application-specific CIM executing in-memory dot-product (IM-DP) for neural

Manuscript received May 28, 2020; revised August 31, 2020; accepted September 25, 2020. Date of publication October 20, 2020; date of current version November 24, 2020. This work was supported by RIE2020 ASTAR AME IAF-ICP Grant under Grant I1801E0030. *(Corresponding author: Yuzong Chen.)*

The authors are with the Centre for Integrated Circuits and Systems (CICS), School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: yuzong.chen@ntu.edu.sg; lulu@ntu.edu.sg; bjkim@ntu.edu.sg; thkim@ntu.edu.sg).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2020.3028848

networks are available in [8], [9], and [13]–[15]. While many CIM designs support only one functionality in addition to the normal memory functionality, several research works implement two or more types of the aforementioned CIM functions on the same array [7], [20] with various tradeoffs such as cell area and data placement strategy. Instead of designing separate CIM macros for each required task, a multifunctional CIM macro can use the same memory array with combined peripheral supports to accelerate different tasks. This improves the versatility of the memory and reduces the cost.

Given the advantages of ReRAM and CIM, it brings significant value to implement robust and high-density ReRAM-based CIM (R-CIM) systems. However, some prior works only focused on the functionality of R-CIM without using conventional memory architectures (i.e., a memory array with decoders). For example, the CIM processor in [20] is based on a field-programmable gate array (FPGA)-like architecture and requires a customized compilation framework to map the dataflow. Besides, it requires three arrays (two as row and column decoders) to implement the storage of one array in the memory mode, leading to large hardware overheads. Zheng *et al.* [16] and Chang *et al.* [17] reported reliable ReRAM-based TCAMs because of the high I_{ON}-to-I_{OFF} ratio of the additional access transistors in the bit-cell. However, the large cell area (i.e., 5T2R) prevents them from being utilized in high-density storage systems. Ly *et al.* [18] extensively characterized the robustness of a high-density 2T2R TCAM, but the analysis did not consider circuit-level variation sources such as the sense amplifier (SA) offset.

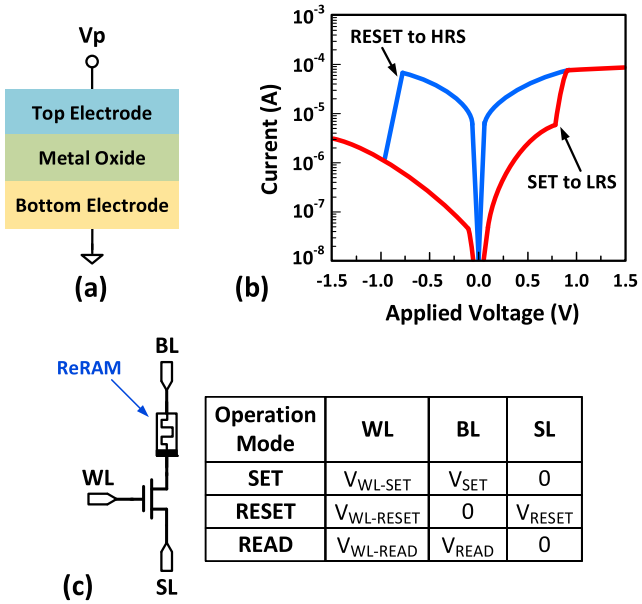
To address these limitations, this article proposes a reconfigurable 2T2R R-CIM architecture. It can support TCAM, LiM, and IM-DP operations. Besides, the 2T2R architecture can also operate as a conventional 1T1R ReRAM under the situations where CIM operations are not required. We optimize the proposed architecture using existing and novel design techniques. We propose a novel LiM full adder (LiM-FA) which is more efficient compared with state-of-the-art LiM-FA [10], [12] in terms of the delay (3.2×), the static power (1.2×), and the dynamic power (1.6×). To characterize the robustness of the proposed R-CIM system, we first provide a quantitative approach for the sensing margin with respect to the precharge voltage and the ReRAM ON/OFF ratio. Combining the proposed optimizations with robustness analysis, we can set a lower precharge voltage for each operation mode and this reduces the TCAM search energy and the 1T1R ReRAM access energy by 1.6× and 1.14×, respectively.

The rest of this article is organized as follows. Section II provides a background of ReRAM and the challenges associated with R-CIM. Section III introduces the proposed 2T2R R-CIM architecture and explains its operations including TCAM, LiM, and IM-DP operations as well as the configurable data storage. Section IV presents several optimizations for the proposed R-CIM using existing and novel design techniques. In Section V, we evaluate the proposed R-CIM system based on optimizations and robustness analysis. And finally, we conclude this article in Section VI.

II. ReRAM Basics and Design Challenges

A. ReRAM Device and 1T1R Bit-Cell

A ReRAM device is typically formed by a metal–insulator–metal stack as shown in Fig. 2(a). It can switch from a high-resistance state (HRS) to a low-resistance state (LRS) by the SET operation and LRS to HRS by the RESET operation. ReRAM devices have two types of switching: 1) unipolar switching where the switching direction depends on the amplitude of the applied programming voltage but not on the voltage polarity and 2) bipolar switching where the switching direction also depends on the polarity of the programming voltage. This article focuses on the bipolar ReRAM device.



Operation Mode	WL	BL	SL
SET	$V_{WL\text{-}SET}$	V_{SET}	0
RESET	$V_{WL\text{-}RESET}$	0	V_{RESET}
READ	$V_{WL\text{-}READ}$	V_{READ}	0

Fig. 2. (a) Three-layer ReRAM device. (b) I–V characteristic of a bipolar ReRAM device with LRS = 10 kΩ and HRS = 1 MΩ. (c) Schematic and biasing conditions for different operations of the 1T1R bipolar ReRAM bit-cell.

For simulation and analysis, we developed a Verilog-A compact model like [22] for the ReRAM device. The model is based on the conductive filament switching mechanism [23]. The I–V relationship of the ReRAM model can be expressed as

$$I = I_0 * \exp\left(-\frac{g}{g_0}\right) * \sinh\left(\frac{V}{V_0}\right) \quad (1)$$

where g is the conductive filament gap distance and V is the voltage applied to the ReRAM device. I_0, g_0, and V_0 are fitting parameters. Fig. 2(b) shows the I–V characteristic curve of a bipolar ReRAM device used in this work with LRS = 10 kΩ and HRS = 1 MΩ. The ReRAM device is integrated with transistors in 40-nm technology.

One common way to realize a ReRAM bit-cell is to integrate a ReRAM device with one transistor (1T1R). Fig. 2(c) shows the 1T1R bit-cell schematic and the biasing conditions of the word-line (WL), the bit-line (BL), and the source-line (SL) for different operations. The read operation can be done in two different sensing ways: voltage-mode and current-mode. Current-mode sensing applies prefixed read voltage to BL and measures the generated current at BL. It is faster than voltage-mode sensing for long BL length [21]. Voltage-mode sensing precharges BL to a level (V_{READ}) and discharges at different rates depending on the ReRAM's resistance state. Generally, the voltage-mode sensing consumes less energy and is suitable for low-power applications [21]. In this article, we employ the voltage-mode sensing scheme.

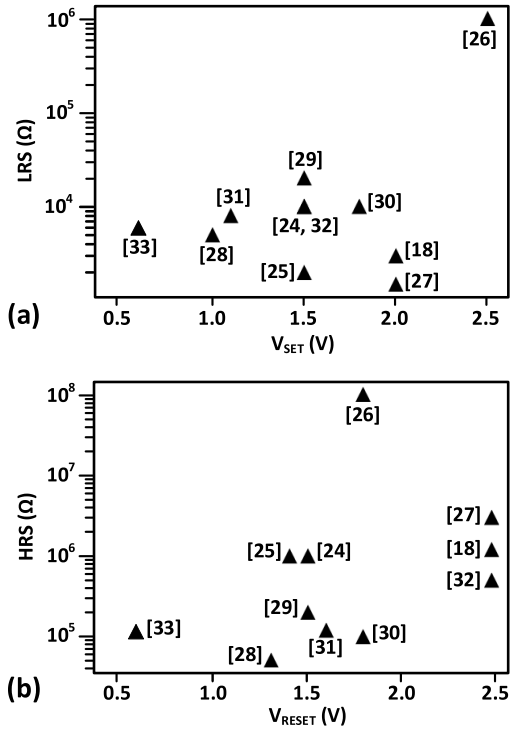


Fig. 3. (a) LRS resistance versus set voltage and (b) HRS resistance versus reset voltage in recent ReRAM technologies.

TABLE I
RERAM DEVICE PARAMETERS FOR SIMULATION

Device Parameters	Values	Circuit Parameters	Voltage (V)
HRS	1MΩ	$V_{WL-SET/RESET}$	2
LRS	10kΩ	$V_{WL-READ}$	1.2
Set Voltage	0.7 V	V_{READ} @ BL	0.3
Reset Voltage	0.7 V	V_{RESET} @ SL	1.2
		V_{SET} @ BL	1.2

B. Challenges of R-CIM

The first challenge of R-CIM comes from the limited HRS-to-LRS ratio (HLR) of the ReRAM device. Fig. 3 shows the relationship between LRS and the set voltage, and HRS and the reset voltage of the ReRAM devices in several published ReRAM works [18], [24]–[33]. All listed technologies have HLRs > 10 while several technologies present HLRs > 100 [18], [24]–[27]. Although HLR of ReRAM is typically higher than that of spin-transfer torque memory [10], it is still much smaller than that of SRAM ($>10^5$), leading to a significant degradation in the sensing margin. This phenomenon is more obvious when multiple rows are activated in CIM where the equivalent HLR is lower than that of a single ReRAM bit-cell [18], [20].

Realizing set and reset voltages lower than the nominal supply voltage of the main-stream fabrication technology is another challenge to be addressed. Recently, several works have achieved low set and reset voltages (~0.6 V) for ReRAM at advanced technology nodes [33], [34]. In this work, we set the target SET and RESET voltages as 0.7 V after comprehensive SPICE simulations using 40-nm technology. The key parameters of the ReRAM device used for simulation in this work are summarized in Table I.

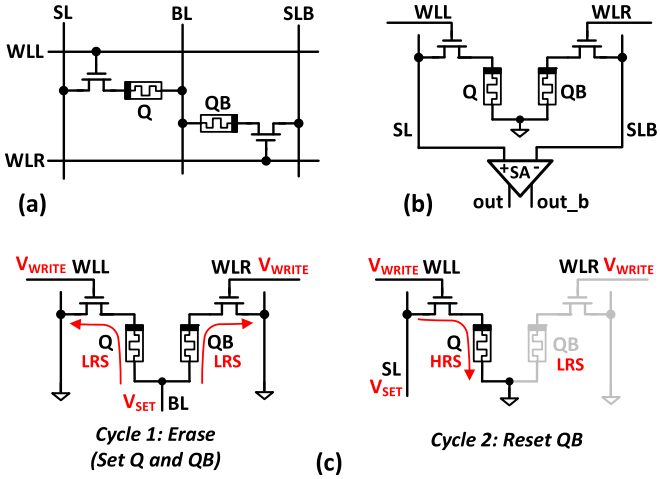


Fig. 4. Proposed 2T2R bit-cell. (a) Schematic, (b) read operation, and (c) writing operation for writing data "1."

Another challenge comes from the read-disturb. As depicted in Fig. 2(b), V_{READ} is applied to BL during a read operation. This is similar to the set condition except that V_{READ} and $V_{WL-READ}$ are smaller than V_{SET} and V_{WL-SET}, respectively. This bias condition does not affect the resistance of an ideal ReRAM device. However, the resistance of a realistic ReRAM device is partially affected by V_{READ} applied on BL, which is called a read-disturb. The read-disturb also occurs when V_{READ} is applied to SL [35], [38], too. To prevent such read disturbance, it is necessary to limit the upper bound of V_{READ}. Chien *et al.* [31] and Lv *et al.* [36] suggested an upper bound of 0.5 V for V_{READ} without having a disturbance on HRS. In [18], the TCAM search endurance is improved from 90k to 450k cycles by lowering V_{READ} from 0.6 to 0.4 V. However, V_{READ} cannot be too low when considering the BL sensing margin, as a higher V_{READ} is necessary to provide more sensing margin when the HLR is small [37].

III. PROPOSED R-CIM ARCHITECTURE

A. 2T2R ReRAM Bit-Cell

Fig. 4(a) shows the 2T2R bit-cell for the proposed R-CIM architecture. It consists of two 1T1R bit-cells sharing a common BL, which is similar to several prior works using a common-SL scheme [35], [38]. Each row in a ReRAM array shares a WL pair (WLL and WLR) while each column shares a BL and an SL pair (SL and SLB). The ReRAM device pair (Q, QB) in the bit-cell represents data "1" with (Q, QB) = (HRS, LRS) and data "0" with (Q, QB) = (LRS, HRS). For TCAM operations, the additional "don't care" state ("X") is represented by (Q, QB) = (HRS, HRS) [16]–[20]. When the proposed bit-cell is used for normal NVM storage, the complementary representation provides significant benefits such as lower bit-error rate and faster sensing compared with the conventional 1T1R ReRAM [13].

Fig. 4(b) explains the read operation of the 2T2R bit-cell. To read the data, BL is grounded, and SL and SLB are precharged to V_{READ} and left floating. Once WLL and WLR are asserted, SL and SLB will discharge at different rates depending on the stored data. After the voltages of SL and SLB develop for some time, a differential SA is enabled and detects this voltage difference to output the data.

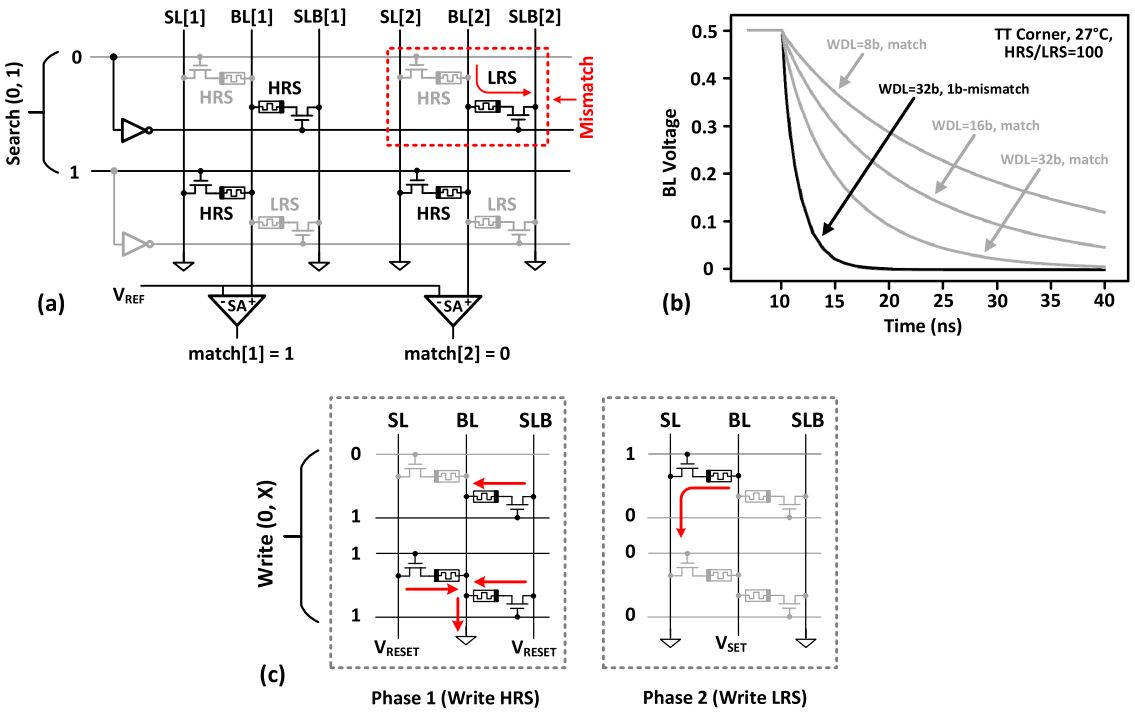


Fig. 5. TCAM operations. (a) Search example (the right column is a mismatch). (b) Simulated waveforms for match and 1-b-mismatch cases of 2T2R TCAM, the 1-b mismatch curve is only shown for WDL = 32 since it is dominated by LRS and is similar for different WDLs. (c) Write example (write data = (0, X)).

The 2T2R ReRAM bit-cell requires two cycles for writing data. Fig. 4(c) shows an example of writing data "1." Since the bit-cell contains two identical 1T1R cells, the same biasing conditions for writing the 1T1R bit-cell can be utilized. However, the shared BL architecture can disturb the ReRAM device programmed in the first cycle. To address this issue, we propose to erase both ReRAM devices by setting them in the first cycle. In the second cycle, either Q or QB is reset to HRS depending on the data to be written. The LRS device is not disturbed in the second cycle since the corresponding SL and BL are grounded.

B. TCAM Operations

TCAM is a critical component in many systems where fast searching is required. The proposed 2T2R ReRAM can operate as a 2T2R TCAM like the ones in [18] and [19] by storing words column-wise (Fig. 5). For TCAM search operation, BL is precharged while SL and SLB are grounded, which is different from the NVM access mode where SL and SLB are precharged and BL is grounded. Search data and inverted search data are applied to WLL and WLR, respectively. If all bits are matched, BL stays at the precharged level or discharges slowly because of the leakage current. If there is a mismatch, one of BL and BLB, or both BL and BLB will discharge quickly through the ReRAM device in LRS. The SA compares the BL voltage against a reference voltage V_{REF} and generates the search result.

Fig. 5(a) explains an example of searching (0, 1). The left column stores (X, 1) and the right column stores (1, 1). Since the first column stores the matched data, BL [1] will be slowly discharged through the leakage current of the bit-cells in HRS, which is recognized as match[1] = "1." However, BL [2] will be discharged below V_{REF} quickly through the bit-cell in LRS and produce a search result indicated by match [2] = "0." If more bits are mismatched, the discharge speed will be higher. Therefore, the worst case scenario is when a 1-bit mismatch occurs. It should be noted that the leakage current during the match case limits the maximum TCAM word-length (WDL). Fig. 5(b) shows the BL voltage waveforms under a 1-bit mismatch case and match cases for different WDLs when BL precharge voltage = 0.5 V. With the increased WDL, it becomes more difficult to distinguish between the match case and 1-bit mismatch case due to the reduced voltage difference.

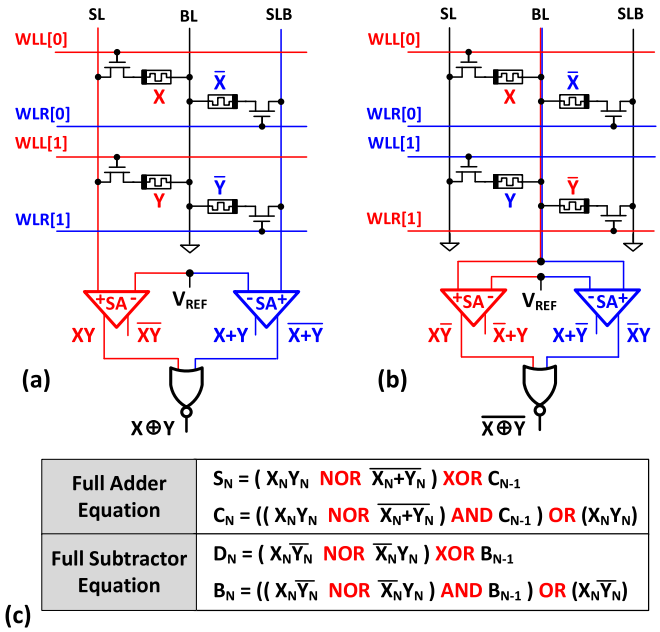
The write operation for TCAM takes two phases to write a word column-wise. HRS states are written in the first phase and LRS states are written in the second phase. An additional column decoder is necessary to select a column to be written. Since multiple cells in a column share the same BL and SL/SLB, the number of cells that can be written in parallel per cycle in each phase depends on the strength of the BL and SL/SLB drivers [39]. Therefore, it may take multiple cycles to program one TCAM word given the area constraints of TCAM write drivers. However, many TCAM applications such as neuromorphic circuits require infrequent writes, and the proposed TCAM is well-suited for such applications due to its nonvolatile feature that consumes zero standby power [18]. Fig. 5(c) illustrates writing a two-bit string (0, X). In phase 1, BL = 0 and SL = SLB = V_{RESET}. In phase 2, BL = V_{SET} and SL = SLB = 0. The WLL and WLR states in each phase are determined by the data to be written as shown in Table II.

C. LiM Operation

The proposed 2T2R structure can also compute Boolean logic functions between two words (X and Y) in memory. We utilize two address decoders to turn on two rows at the same time and use two single-ended SAs to generate LiM

TABLE II
BIASING FOR TCAM WRITE OPERATION

Data	'0'		'1'		'X'	
Phase #	1	2	1	2	1	2
WLL	GND	V_{WRITE}	V_{WRITE}	GND	V_{WRITE}	GND
WLR	V_{WRITE}	GND	GND	V_{WRITE}	V_{WRITE}	GND



(c)

Full Adder Equation	$S_N = (X_N Y_N$ NOR $\overline{X_N+Y_N})$ XOR C_{N-1} $C_N = ((X_N Y_N$ NOR $\overline{X_N+Y_N})$ AND $C_{N-1})$ OR $(X_N Y_N)$
Full Subtractor Equation	$D_N = (X_N \overline{Y_N}$ NOR $\overline{X_N} Y_N)$ XOR B_{N-1} $B_N = ((X_N \overline{Y_N}$ NOR $\overline{X_N} Y_N)$ AND $B_{N-1})$ OR $(X_N \overline{Y_N})$

Fig. 6. LiM operations between two words X and Y. (a) By precharging SL/SLB and grounding BL. (b) By precharging BL and grounding SL/SLB. (c) FA and FS equations.

primitives. As shown in Fig. 6(a), the AND/NAND operations can be performed by activating WLLs of two rows. If X or Y is in LRS, SL will discharge quickly. If both are in HRS, SL will discharge much slowly. The left SA compares the SL voltage against V_{REF} and generates the AND/NAND results. The NOR/OR operations can be performed similarly by activating WLRs of two rows. The right SA compares the SLB voltage against V_{REF} and produces the NOR/OR results. The XOR operation can be realized by connecting the AND/NOR results to a NOR gate. These primitives allow a LiM-FA to be implemented with a few additional logic gates [10], [12]. The FA operation can be computed in one cycle since the proposed 2T2R array structure simultaneously calculates all required primitives for FA as indicated by the FA equation in Fig. 6(c).

The proposed 2T2R structure can also compute the logic primitives required by a full subtractor (FS) by precharging and sensing BL, and grounding SL/SLB. As shown in Fig. 6(b), by activating WLL for row X and WLR for row Y, the two SAs can both implement $X\overline{Y}$ and $(\overline{X}+Y)$ in one cycle. Similarly, by activating WLR for row X and WLL for row Y, the two SAs can both implement $\overline{X}Y$ and $(X+\overline{Y})$ in one cycle. Note that $X\overline{Y}$ and $\overline{X}Y$ are both required as indicated by the FS equation in Fig. 6(c), but they cannot be computed simultaneously. Therefore, the FS will take two cycles to complete by latching the SA results separately (e.g., latch $X\overline{Y}$ to the left SA in the first cycle and latch $\overline{X}Y$ to the right SA in another cycle) to get all required primitives for FS.

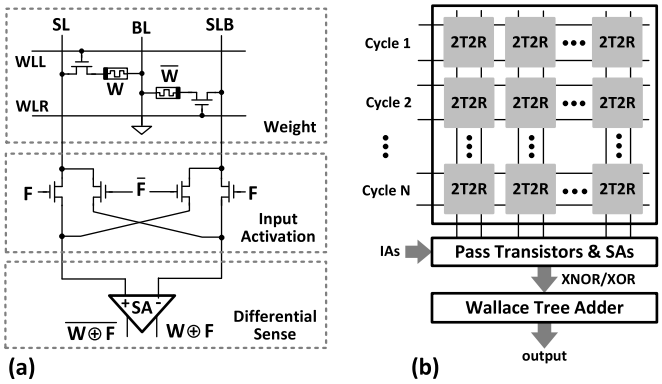


Fig. 7. IM-DP for BNNs. (a) Sensing scheme. (b) Simplified architecture.

D. IM-DP for Binary Neural Networks

Neural networks are powerful tools to achieve state-of-the-art results for many smart applications such as computer vision and natural language processing. But they consume significant storage and computation resources. Recently, binarized neural networks (BNNs) [40] have been proposed to reduce storage and computation demands by restricting weights and input activations (IAs) to $+1$ and -1. As a result, the dot-product operation (the dominant operation in neural networks) in BNNs is replaced by the simple XNOR-popcount operation. Several works have implemented IM-DP for BNNs based on the 2T2R ReRAM structure [8], [13]. Chen *et al.* [8] used multilevel current-mode SAs to accomplish this, while Bocquet *et al.* [13] used differential voltage-mode sensing. However, the above 2T2R ReRAMs fail to support other functions like the proposed architecture.

Fig. 7 explains how the proposed ReRAM architecture executes IM-DP for BNNs. The ReRAM array implemented with the proposed 2T2R bit-cells stores weights (W) in a complementary format. Additional pass transistors controlled by the BNN layer's IAs (represented by signal F in a complementary format) connect SL and SLB to a differential SA as in [13]. Therefore, the XNOR/XOR between the weight and the IA can be computed during a differential read as shown in Fig. 7(a). Fig. 7(b) shows the simplified architecture to perform the IM-DP. All BLs are grounded (not drawn in Fig. 7 for simplicity). In each cycle, one row is activated by enabling its WLL and WLR. The bitwise XNOR/XOR between weights and IAs are computed by the differential SA and the results are sent to a digital Wallace tree adder to perform the popcount operation. Although this approach has a throughput degradation compared with other implementations of IM-DP that activate many rows at the same time [8], [9], this purely digital implementation does not need analog-to-digital converters (ADCs) which incur high area overhead. For example, the work in [9] requires 64-bit-cells in every column to implement a 5-bit column ADC.

E. Configurable Data Storage

Although the proposed 2T2R structure can perform different types of CIM operations, a system may need the ReRAM as a pure storage block, e.g., program instruction storage. In such situations, it is desirable to minimize the area overhead coming from the 2T2R bit-cell structure. To address this issue, the

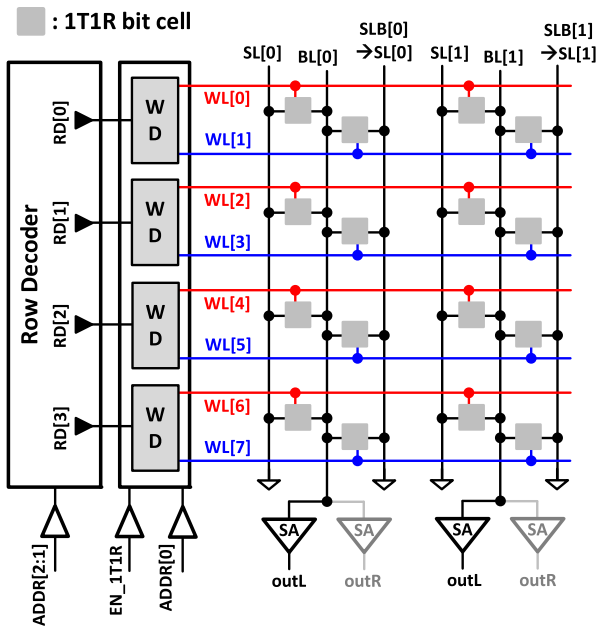


Fig. 8. Read operation of the 2T2R structure configured as a 1T1R array.

proposed 2T2R bit-cell can be reconfigured so that it can operate as two conventional 1T1R bit-cells. Fig. 8 shows a diagram of the read operation in the 1T1R mode. While the array has 4 rows × 2 columns in the 2T2R mode, it is configured into 8 rows × 2 columns in the 1T1R mode by configuring one 2T2R bit-cell into two symmetric 1T1R bit-cells sharing one BL. The SL pair SL/SLB in the 2T2R mode becomes SLs in the 1T1R mode. Although the two SLs in one column are not physically connected, they have the same column index.

The read operation follows that of the conventional 1T1R ReRAM by grounding SL and precharging BL. Since the number of rows is doubled, one additional row-address bit is necessary to differentiate WLR from WLL. In Fig. 8, the 2-bit address ADDR[2:1] is sent to the row decoder as for 2T2R read. The additional address bit ADDR[0] is sent to the word-line driver (WD) that contains WLL/WLR logic. If ADDR[0] = "0," WLL (red line) is turned on in that row; if ADDR[0] = "1," WLR (blue line) is turned on in that row. Although BL is connected to two single-ended SAs, we can use only one of them. The biasing for the write operation of the 1T1R configuration is identical to that of the 1T1R ReRAM as shown in Fig. 2(c).

IV. Optimization of the Proposed R-CIM

A. Reconfigurable SA

As explained in Section III, the interconnects between the R-CIM array and the SAs need to be controlled based on the selected functions. Fig. 9 shows the schematic of the reconfigurable SA modified from [7]. It consists of two pairs of cross-coupled inverters (SA1 and SA2). The 2:1 MUX connected to SL/SLB and BL is controlled by a select signal (BL_sel) and connect them to the SA. The output latches (L and R) are controlled by additional control signals "en_L" and "en_R" to sample the results of two single-ended SAs separately or together as required by the LiM-FA or FS

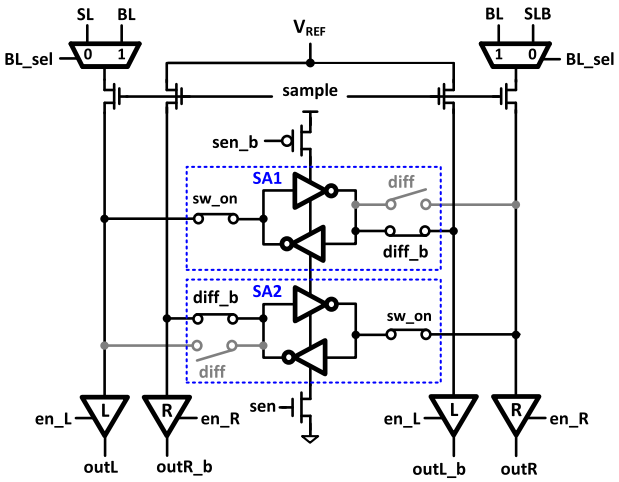


Fig. 9. Schematic of the reconfigurable SA modified from [7] in the TCAM or LiM mode. Two switches "sw_on" are always ON.

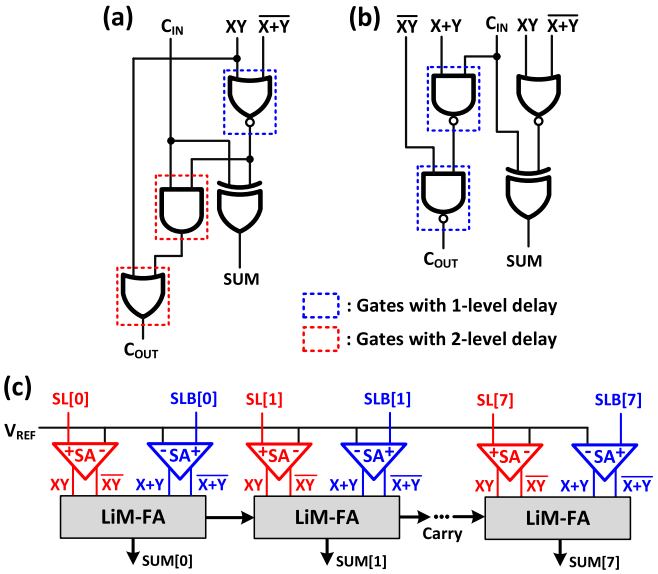


Fig. 10. LiM-FA designs (a) before and (b) after optimization. (c) Implementation of an 8-bit RCA using LiM-FAs.

operation. In the TCAM and 1T1R storage modes (i.e., BL_sel = "1" and diff = "0"), BL is compared with V_{REF} in the SA as depicted in Figs. 5(a) and 8, respectively. In the LiM mode (i.e., diff = "0"), either SL/SLB or BL is compared with V_{REF} as shown in Fig. 6(a) and (b), respectively. For 2T2R ReRAM read and IM-DP operations (i.e., BL_sel = "0" and diff = "1"), SA1 and SA2 are connected in parallel to operate as a differential SA.

The original design in [7] uses PMOS transistors controlled by "diff" and "diff_b." Since PMOS can only pass a weak logic "0," we use transmission-gate (TG) switches in our design to obtain strong differential outputs for both cross-coupled inverters. These four outputs provide all the primitives required for the LiM-FA and FS. Note that the TGs controlled by "diff" and "diff_b" in each small SA have a limited size for area efficiency. Therefore, their ability to pass voltages during sensing will affect the SA offset. Since one of them is ON during the sensing time, we put one additional switch (sw_on) that is always ON in each SA to compensate for the offset.

B. Optimization of FA

Prior LiM-FA design [10], [12], as shown in Fig. 10(a), is based on the Boolean equations in Fig. 6(c). However, the carry propagation is the critical path in ripple-carry adder (RCA), therefore should be optimized to reduce latency. This work improves the latency of the carry-out by modifying the carry-out expression compatible with the proposed 2T2R array architecture. The Boolean equation for C_{OUT} can be written as a sum-of-product

$$\begin{aligned} C_{\text{OUT}} &= XY + XC_{\text{IN}} + YC_{\text{IN}} \\ &= XY + (X+Y)C_{\text{IN}} \end{aligned} \tag{2}$$

where X and Y are two inputs, and C_{IN} is carry-in. Using De Morgan's theorem, (2) can be reorganized as

$$\begin{aligned} C_{\text{OUT}} &= \overline{\overline{XY + (X+Y)C_{\text{IN}}}} \\ &= \overline{\overline{XY} * \overline{(X+Y)C_{\text{IN}}}}. \end{aligned} \tag{3}$$

This shows that two primitives, $\overline{XY}$ and $(X + Y)$, can implement C_{OUT} using only two additional NAND gates. Note that the above primitives and the two primitives for computing the FA's sum (XY and $\overline{X+Y}$) are all available at the reconfigurable SA's outputs as shown in Fig. 6(a). Fig. 10(a) and (b) shows the LiM-FA designs before and after optimization, respectively. The design before optimization requires a five-level delay for carry propagation while the optimized design requires only a two-level delay. The optimized FA can also be reused as an FS without changing the wiring between the reconfigurable SA's outputs and the additional four logic gates. By chaining LiM-FAs, an RCA can be implemented as shown in Fig. 10(c).

C. Inherent SA Redundancy for TCAM and 1T1R Read

For 2T2R TCAM search operations, the sensing margin is severely degraded compared to the conventional CMOS TCAMs because of the large leakage current through the HRS devices in the match case [18], [19]. For 1T1R ReRAM read operations, the sensing margin also needs to be large due to single-ended sensing. Therefore, it is important to reduce the SA offset to enhance the robustness of TCAM search and 1T1R read operations.

In the proposed reconfigurable SA, BL is connected to two single-ended SAs in the TCAM mode and the 1T1R storage mode. This inherently allows the use of SA redundancy [41]. The main benefit of SA redundancy is that the exclusive selection of multiple small SAs can improve the offset compared to one large SA. Fig. 11 shows simulated SA offsets based on 10 000 Monte-Carlo points. The small SA is the single-ended SA1 or SA2 in the reconfigurable SA presented in Fig. 9. The large SA is the reconfigurable SA in differential mode (i.e., SA1 and SA2 connected in parallel). Each of the two small SAs has an offset of 12.5 mV while the large SA has an offset of 8.9 mV. By employing SA redundancy and selecting one SA with a smaller offset from the two small SAs, the new offset becomes 7.5 mV. The SA redundancy improves the offset by 1.67× and 1.19× compared to the small SA and the large SA, respectively. The improved offset obtained from the SA redundancy will be used for the evaluation of the proposed R-CIM in Section V.

TABLE III
WLL/WLR LOGIC TABLE

OPC (AB)	*Operations*	*RD_1[jj] (X)*	*RD_2[jj] (Y)*	*WLL[jj] (AX+BY)*	*WLR[jj] (AY+BX)*
11	2T2R NVM, LiM-FA, IM-DP	1	0	**1**	**1**
		0	1	**1**	**1**
10	LiM-FS cycle 1, 1T1R NVM	1	0	**1**	**0**
		0	1	**0**	**1**
01	LiM-FS cycle 2	1	0	**0**	**1**
		0	1	**1**	**0**
00	Disable WLs	-	-	**0**	**0**

D. Additional Circuits for Versatile CIM Functions

Implementation of versatile CIM functions on a conventional ReRAM array requires additional architectural support. We propose several architectural optimizations below.

1) WLL/WLR Logic: This logic block should be designed carefully since it is replicated in every row. In this work, we use a 2-bit operation code (OPC) and two row decoder outputs (RD_1[j] and RD_2[j]) to control WLL and WLR as shown in Fig. 12(a). RD_1[j] and RD_2[j] represent the outputs from two row decoders for row j. Table III shows the detailed logic table for WLL/WLR. The simplified logic expressions for WLL and WLR are "AX+BY" and "AY+BX," respectively. Here, "A" and "B" represent the 2-bit OPC while "X" and "Y" are the outputs from the row decoders. These two simple representations allow the WLL/WLR logic to be designed with only a few logic gates and the area overhead can be small.

2) Multicycle TCAM Search: One challenge for high-density ReRAM arrays is that the SA cannot fit into the column pitch. Therefore, column multiplexing is generally employed. In this work, a column multiplexing ratio of four is used for a 256 × 128 array (Section V) to read out 32-bit words. Such a design choice increases the area efficiency but prevents parallel search in the TCAM mode. To better utilize the column multiplexing feature, we propose to store each TCAM word across multiple columns, which requires multiple cycles for a search operation. In addition, the maximum WDL in the TCAM mode depends on the HLR of the ReRAM device. For HLR = 100, WDL can be 32 [19]. Therefore, a 128-bit TCAM word needs to be stored across the first 32 rows and 4 columns in the 256 × 128 array. Other rows can still be used for other CIM operations. Note that this design choice depends on the application (the required TCAM WDL), the HLR of ReRAM device, and the memory size. For a complete search, partial search results from four cycles are combined [39]. Fig. 12(b) shows an example where a 64-bit word is searched over two cycles. Two columns store the 64-bit word (i.e., 32 bits in each column). The partial result combiner after the SA consists of just an SR flip-flop and a NOR gate.

E. Sensing Margin Considerations

Fig. 13 shows the simulated read power breakdown of a conventional 1T1R ReRAM array using the DESTINY simulator [42]. The array size is 256 × 256 with a column multiplexing ratio of two. The BL precharge voltage is set to 0.3 V at room temperature. Due to the large BL capacitance,

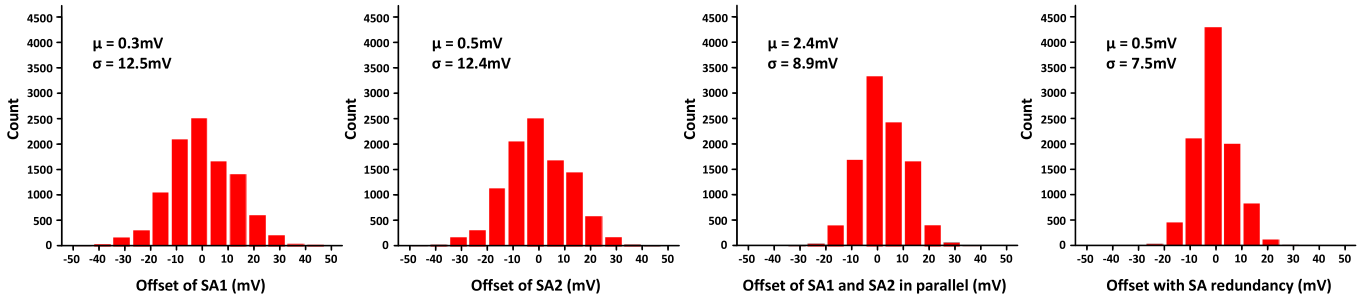


Fig. 11. SA offsets for two small SAs (SA1 and SA2 in Fig. 9), one large SA (SA1 and SA2 connected in parallel), and SA redundancy.

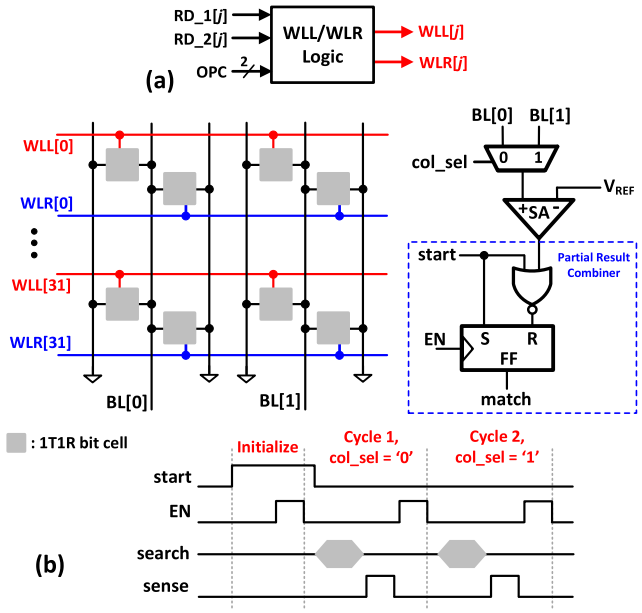


Fig. 12. Architectural support. (a) WLL/WLR logic block. It receives an OPC and results from two row decoders to control WLL and WLR of row j. (b) Example for multicycle TCAM that searches a 64-bit word in two cycles and combines partial results.

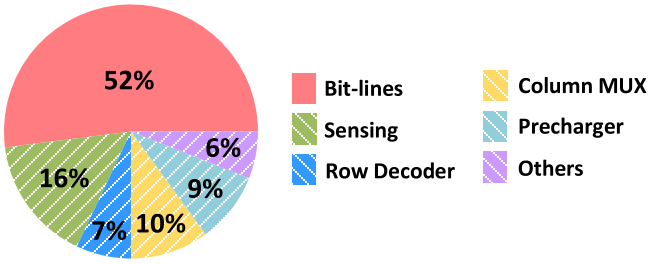


Fig. 13. Simulated power breakdown of ReRAM read with 0.3-V BL precharge voltage at room temperature.

BL precharge and discharge consume more than half of the read power. For a low precharge voltage, a full BL swing is generally required to get the maximum sensing margin across process variations [43]. Therefore, the energy consumed by BL switching can be expressed as

$$E_{\mathrm{BL}} = C_{\mathrm{BL}} \times V_{\mathrm{READ}}^2 \tag{4}$$

where C_{BL} is the BL capacitance and V_{READ} is the BL precharge voltage. As described by (4), the reduction in V_{READ} can reduce the BL switching power. But this comes at the cost of BL sensing margin degradation. As explained in Section II, V_{READ} also disturbs the ReRAM device resistance during read operations and affects reliability. Therefore, V_{READ} needs to be determined carefully after considering power, sensing margin, speed, and read-disturb.

In the voltage sensing mode, the sensing margin (V_{SM}) is determined by the voltage difference between two BL voltage levels that need to be distinguished. These two voltage levels come from two RC discharging circuits with high resistance and low resistance, respectively, which is given in the following:

$$V_{\mathrm{SM}}(t) = V_{\mathrm{READ}} * \left(\exp\left(\frac{-t}{R_{\mathrm{H}} C_{\mathrm{BL}}}\right) - \exp\left(\frac{-t}{R_{\mathrm{L}} C_{\mathrm{BL}}}\right)\right). \tag{5}$$

Here, $V_{\mathrm{SM}}(t)$ is the sensing margin with respect to time, V_{READ} is the precharge voltage, R_{H} and R_{L} are the high resistance and the low resistance seen from BL, and C_{BL} is the BL capacitance. Note that R_{H} and R_{L} may not be the same as HRS and LRS. For example, for the TCAM search operation, R_{H} and R_{L} are the worst case equivalent resistance values seen from BL in the match case and the 1-bit mismatch case, respectively. When considering n-bit search data, R_{H} is obtained from n HRS cells while R_{L} is from $(n-1)$ HRS cells and one LRS cell. The derivative of (5) with respect to time t is as follows:

$$\frac{dV_{\mathrm{SM}}}{dt} = V_{\mathrm{READ}} * \exp\left(\frac{-t}{R_{\mathrm{L}} C_{\mathrm{BL}}}\right) * \frac{1}{R_{\mathrm{L}} C_{\mathrm{BL}}} - V_{\mathrm{READ}} * \exp\left(\frac{-t}{R_{\mathrm{H}} C_{\mathrm{BL}}}\right) * \frac{1}{R_{\mathrm{H}} C_{\mathrm{BL}}}. \tag{6}$$

To calculate the maximum V_{SM}, (6) should be set to 0, which generates the delay to get the maximum V_{SM}

$$t_{\max(V_{\mathrm{SM}})} = \frac{R_{\mathrm{H}} C_{\mathrm{BL}}}{R_{\mathrm{H}}/R_{\mathrm{L}} - 1} * \ln\left(\frac{R_{\mathrm{H}}}{R_{\mathrm{L}}}\right). \tag{7}$$

From (7), $t_{\max(V_{\mathrm{SM}})}$ only relies on $R_{\mathrm{H}}/R_{\mathrm{L}}$, not on V_{READ}. Fig. 14(a) shows the normalized $t_{\max(V_{\mathrm{SM}})}$ versus $R_{\mathrm{H}}/R_{\mathrm{L}}$ with a specific time constant $R_{\mathrm{H}} C_{\mathrm{BL}}$. It can be observed that $t_{\max(V_{\mathrm{SM}})}$ decreases exponentially when $R_{\mathrm{H}}/R_{\mathrm{L}}$ increases. This indicates that when $R_{\mathrm{H}}/R_{\mathrm{L}}$ is small, the SAs need more delay and the enabling time is sensitive to the variations in $R_{\mathrm{H}}/R_{\mathrm{L}}$. However, when $R_{\mathrm{H}}/R_{\mathrm{L}}$ is relatively large, the SAs can be enabled earlier and $t_{\max(V_{\mathrm{SM}})}$ is not sensitive to the variations in $R_{\mathrm{H}}/R_{\mathrm{L}}$, which is more desirable. After applying (7) back to (6), the maximum V_{SM} becomes

$$\max(V_{\mathrm{SM}}) = V_{\mathrm{READ}} * \left(\left(\frac{R_{\mathrm{H}}}{R_{\mathrm{L}}}\right)^{\frac{1}{1-\frac{R_{\mathrm{H}}}{R_{\mathrm{L}}}}} - \left(\frac{R_{\mathrm{H}}}{R_{\mathrm{L}}}\right)^{\frac{1}{\frac{R_{\mathrm{L}}}{R_{\mathrm{H}}}-1}}\right). \tag{8}$$

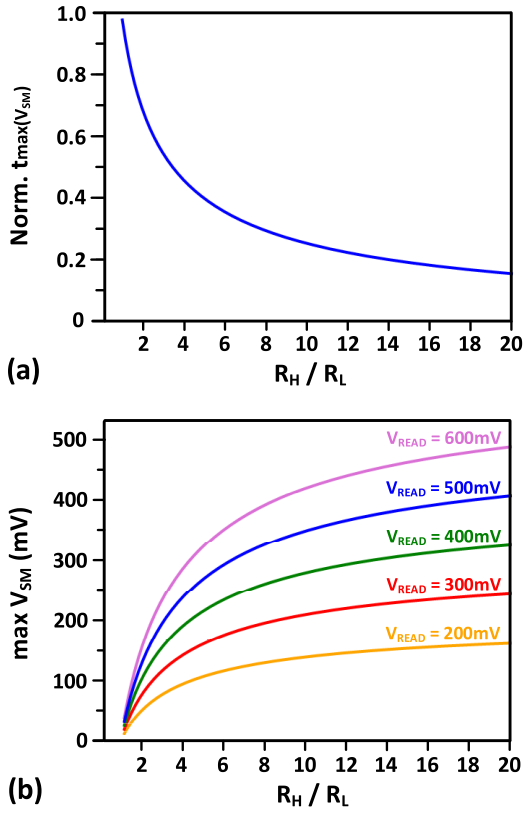


Fig. 14. (a) Normalized $t_{\max(V_{SM})}$ versus R_H/R_L. (b) $\max(V_{SM})$ versus R_H/R_L at different V_{READ}.

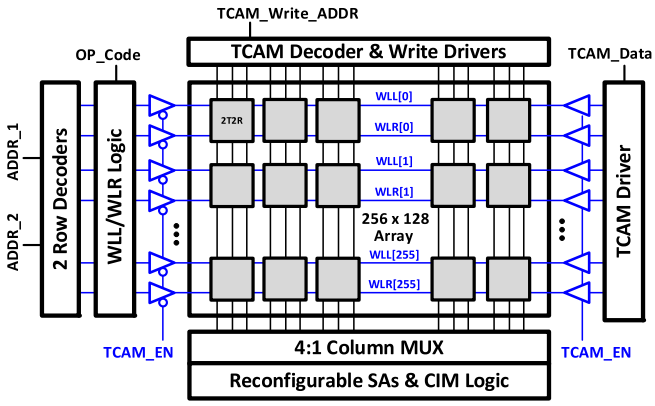


Fig. 15. Block diagram of the 256 × 128 array for simulation.

TABLE IV
OFFSETS OF DIFFERENT SA CONFIGURATIONS

Config	*Meaning*	*Mean (mV)*	*Sigma (mV)*
SA1	One small SA	0.3	12.5
SA2	One small SA	0.5	12.4
SA1 + SA2	Two small SAs connected in parallel	2.4	8.9
min(SA1, SA2)	The small SA with smaller offset	0.5	7.5

Fig. 14(b) shows the relationship between $\max(V_{SM})$ and R_H/R_L at different V_{READ} levels. When R_H/R_L is low, V_{SM} shows a larger sensitivity to R_H/R_L. When R_H/R_L is high, V_{SM} is less sensitive to R_H/R_L. In summary, it is desirable to provide high R_H/R_L for reliable BL sensing.

V. EVALUATION OF THE PROPOSED 2T2R RERAM

In this section, we evaluate the proposed 2T2R ReRAM under different operating modes. We first analyze the area and energy overheads of the proposed R-CIM architecture using Hspice and a modified version of DESTINY [42]. Then we present how different optimization techniques help improve the robustness of the R-CIM system and reduce the energy consumption by setting a lower V_{READ}. The ReRAM device is modeled with Verilog-A and the model is calibrated using the data from [24] to get default LRS = 10 kΩ and HRS = 1 MΩ. The modeled ReRAM devices are integrated with transistors in 40-nm CMOS technology. We design an R-CIM array of 256 × 128 as shown in Fig. 15. Note that WLLs/WLRs are controlled by tristate buffers and signal "TCAM_EN." In TCAM mode, WLLs/WLRs are driven by TCAM drivers instead of row decoders and the maximum allowed search WDL is 32 as explained in Section IV-D.

The sensing margin must be larger than the SA offset for reliable sensing. The SA offset from Fig. 10 are summarized in Table IV. We consider SA offset variations of 4-sigma for an acceptable yield. Besides the SA offset, variations of access transistors of the 2T2R cell also need to be considered. We perform 1000 Monte-Carlo simulations for TCAM search and ReRAM access operations for different V_{READ}. Fig. 16(a) and (b) shows the simulation results for TCAM (V_{READ} = 0.5 V) and ReRAM access (V_{READ} = 0.3 V), respectively. The maximum variations of 32-bit match and 1-bit mismatch for TCAM have standard deviations equal to 5.8 and 7.3 mV, respectively. The variations due to access transistors for normal ReRAM read have smaller standard deviations. The transistor variations for LiM is similar to that of ReRAM read because they only access a small number of cells. The standard deviations for accessing HRS and LRS devices are 0.3 and 6.5 mV, respectively. As a result, TCAM has the worst V_{SM} degradation due to transistor variations of the 2T2R bit-cell.

In the following analysis, results for the achievable V_{SM} are based on the voltage difference developed between the worst Monte-Carlo case of two voltage levels that need to be distinguished, e.g., the lowest curve of 32-bit match and the highest curve of 1-bit mismatch in Fig. 16(a). To achieve the maximum sensing margin, the time to enable sensing is calculated using (7). Note that (7) can be close to the optimal sensing time since the resistance due to access transistors is much smaller compared with the ReRAM resistances. For single-ended sensing, the reference voltage is put in the middle of two different voltage levels after determining the sensing time.

For ReRAM variations, the LRS value is 10 kΩ with a 20% variation through all simulations. We use different HRS values to generate HLRs of 150, 100, and 50 to evaluate V_{SM} with respect to HRS variations. This HRS variation is adopted from [18] that characterizes a 2T2R TCAM and the reported $\pm 2.5\sigma$ HRS corresponds to 50% variation with respect to the

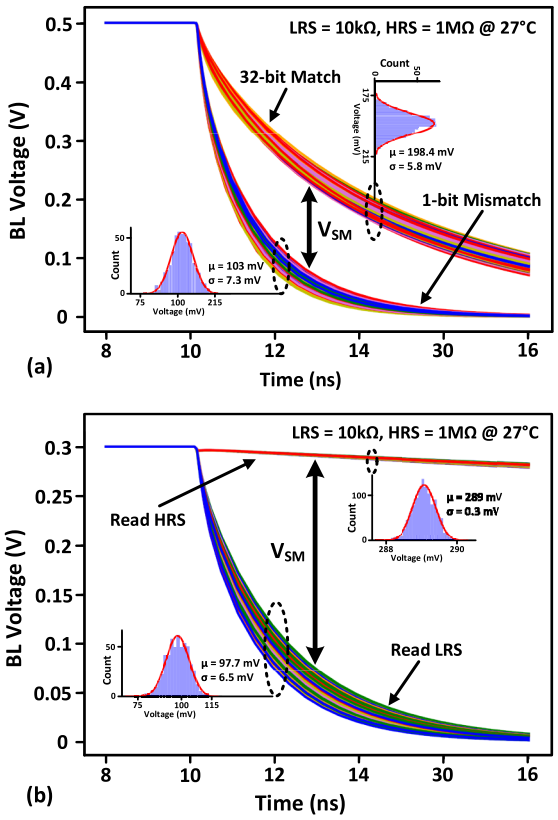


Fig. 16. BL discharge variations due to access transistors for (a) TCAM and (b) ReRAM access operations.

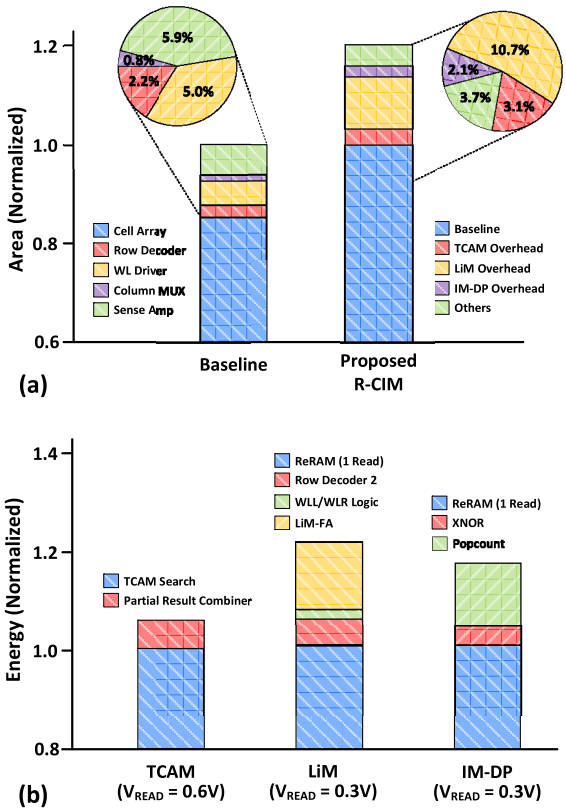


Fig. 17. Overhead of the proposed R-CIM. (a) Area. (b) Energy.

mean HRS value. Note that our assumption on HRS is more pessimistic compared with [18] since we use the worst 50% variation for all HRS when HLR = 50.

A. Area and Energy Analysis

The proposed R-CIM system supports versatile operations at the cost of additional hardware resources and energy. Fig. 17(a) shows the area breakdown of the baseline (a 2T2R array without CIM support) and the proposed R-CIM architecture. Compared with the baseline, the R-CIM overheads (16.6%) due to TCAM, LiM, and IM-DP operations are 3.1%, 10.7%, and 2.1%, respectively. An additional 3.7% overhead comes from the read voltage generators and the output MUX logic that sends SA outputs to different near-memory blocks in different modes. TCAM overhead includes the tristate buffers (0.6%), the column decoder (1.2%), and the partial result combiner (1.3%). LiM overhead includes the optimized LiM-FA circuit (0.8%), the additional row decoder (2.2%), and the WLL/WLR logic block (7.7%). IM-DP overhead includes the XNOR circuit (0.3%) and the popcount circuit (1.8%). Note that the WLL/WLR logic incurs the highest area overhead since it needs to be replicated in every row. However, this overhead can be further mitigated if the number of columns increases.

Fig. 17(b) shows the energy overhead of various circuit blocks for different CIM operations. For TCAM search operations, the overhead comes from the partial result combiner is 6.3% at V_{READ} = 0.6 V. LiM operations at V_{READ} = 0.3 V incur a slightly higher energy overhead (23.8%) compared to the baseline (one ReRAM read) due to the additional row decoder (7%), the WLL/WLR logic (1.5%), and the LiM-FA circuit (15.3%). However, the LiM energy overhead allows versatile LiM operations between two operands. With a naïve implementation that reads out two operands and sends them to ALUs, these operations will cost at least the energy of two ReRAM reads. For IM-DP at V_{READ} = 0.3 V, the energy overhead (17.4%) due to the XNOR circuit and the popcount circuit is 4% and 13.4%, respectively.

B. TCAM Robustness Evaluation

Fig. 18(a) shows the achievable V_{SM} for different V_{READ} in TCAM mode. As expected, increasing HLR improves V_{SM} at a given V_{READ}. Also, V_{SM} is more sensitive to HLR in TCAM mode since TCAM provides relatively small R_H/R_L as explained in Section IV-D. Without employing SA redundancy, the minimum voltage difference between BL and V_{REF} needs to be at least 50 mV to overcome the SA offset (i.e., 12.5 mV, 4-sigma variation). Therefore, the minimum voltage difference between the match case and the 1-bit mismatch case should be twice the voltage difference between BL and V_{REF}, which equals 100 mV. When employing the SA redundancy, the minimum V_{SM} becomes 60 mV, which is a 1.67× improvement.

Since TCAM search gives low R_H/R_L, we further characterize its robustness under temperature variations. Fig. 18(b) shows the achievable V_{SM} at different temperatures considering the nominal case (HLR = 100) and HRS degradation (HLR = 50). The achievable V_{SM} degrades as the temperature

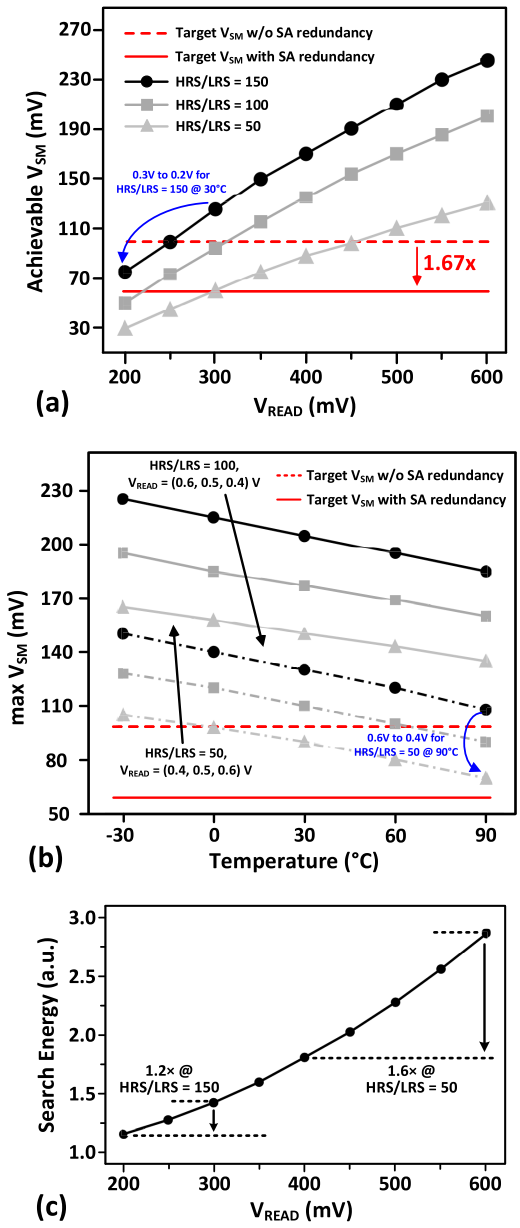


Fig. 18. TCAM robustness evaluation. (a) V_{SM} versus V_{READ} for different HLRs at room temperature. (b) V_{SM} at different temperatures and HLRs. (c) Search energy (normalized) with respect to V_{READ}.

increases. With nominal HLR, low V_{READ} (e.g., 400 mV) still gives enough V_{SM} at high temperatures. However, if considering HRS degradation, a high V_{READ} (e.g., 600 mV) must be used to provide enough V_{SM} at high temperatures if not employing SA redundancy. On the contrary, our optimizations by employing the SA redundancy reduces the V_{SM} requirement and allows TCAM to operate at $V_{READ} = 400$ mV under high temperatures. Note that TCAM has the worst V_{SM}, requiring higher V_{READ}. The other operations of the proposed ReRAM architecture can have lower V_{READ}.

The normalized TCAM search energy (including peripheral logics) with respect to V_{READ} is presented in Fig. 18(c). When HLR = 150 at room temperature, SA redundancy allows V_{READ} to be lowered from 300 to 200 mV as depicted in Fig. 18(a). This improves the TCAM search energy by 1.2×. When the HLR decreases because of HRS degradation, the SA

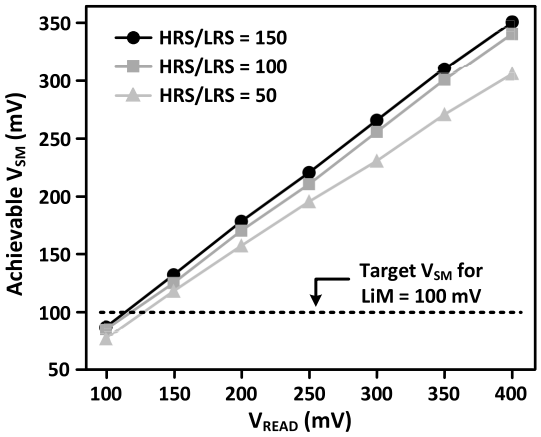


Fig. 19. V_{SM} versus V_{READ} for LiM primitive operations.

redundancy can lower V_{READ} from 600 to 400 mV at 90 °C as shown in Fig. 18(b), improving the TCAM search energy by 1.6×. Note that a V_{READ} of 600 mV will generate significant read disturbance and degrade TCAM search endurance [18]. Therefore, instead of setting a default V_{READ}, it is necessary to adjust V_{READ} smartly depending on the HLR and its degradation to avoid the energy overhead coming from the overdesign.

C. LiM Evaluation

We evaluate the robustness of the proposed LiM primitive operations (e.g., in-memory AND/NOR) as well as the performance and energy consumption of the LiM-FA. Fig. 19 shows the achievable V_{SM} versus V_{READ} for LiM primitive operations. The V_{SM} of LiM primitive operations is much larger than that of the TCAM search operation because of the increased R_H/R_L ratio obtained from the same HRS and LRS values. For example, when computing the in-memory AND function, R_H is obtained from two HRS in parallel while R_L is one LRS and one HRS in parallel which is approximately one LRS. This causes only a 2× reduction in R_H/R_L compared to the HLR of the ReRAM device. Moreover, V_{SM} is much less sensitive to HRS degradation because of the high R_H/R_L ratio (>15). Based on the proposed SA analysis, the target V_{SM} is 100 mV without employing the SA redundancy, giving the minimum V_{READ} of 150 mV for HLR = 50.

Table V compares various FA designs. The LiM-FA [10], [12] and CMOS FA [47] are simulated with 40-nm technology. Data for other designs are directly obtained from the relevant articles. Compared with other LiM-FAs [10], [12], the proposed LiM-FA achieves 3.2×, 1.2×, and 1.6× improvements in the delay, the static power, and the dynamic power, respectively. Compared with CMOS FA, the proposed LiM-FA has slightly worse dynamic power due to more levels of transition, but the delay and the static power are 1.34× and 8.9× better with fewer transistors. We also compare our LiM-FA with several FAs based on nonvolatile devices such as magnetic tunnel junctions (MTJs) [44], ferroelectric tunnel junctions (FTJs) [45], and ferroelectric field-effect transistors (FeFETs) [46]. Regarding performance, our LiM-FA is only slightly worse than the FA based on FeFET [46]. This is because the FA in [46] uses the dynamic logic design style with only a pull-down NMOS network, therefore less capaci-

TABLE V
DELAY AND POWER CONSUMPTION OF DIFFERENT FA DESIGNS

	This Work	[10][12]	[44]	[45]	[46]	[47]
Design Style	LiM, Fig. 10(b)	LiM, Fig. 10(a)	MTJ DyCML*	FTJ DyCML*	FeFET DyCML*	Complementary CMOS
Technology	40nm	40nm	40nm	40nm	45nm	40nm
Transistor Count	20 MOS†	24 MOS†	34 MOS + 4 MTJs	30 MOS + 4 FTJs	28MOS + 4 FeFETS	28 MOS
V_{DD} (V)	1.2	1.2	1.2	1.2	1	1.2
T_d (ps) / Normalized	28.7 / 1×	91.4 / 3.2×	87.4 / 3×	500 / 17.4×	20.3 / 0.7×	38.4 / 1.4×
P_{static} (nW) / Normalized	3.1 / 1×	3.6 / 1.2×	N.A.	N.A.	133.6 / 43.1×	27.7 / 8.9×
P_{DYN} (μW) / Normalized	0.68 / 1×	1.07 / 1.6×	1.98 / 2.9×	1.70 / 2.5×	1.10 / 1.6×	0.63 / 0.9×

∗ DyCML stands for "Dynamic Current Mode Logic", simulation based on 500MHz clock frequency.

† XOR gate in Fig. 9 is designed with two transmission gates, plus two inverters for two inputs. Totally 8 transistors.

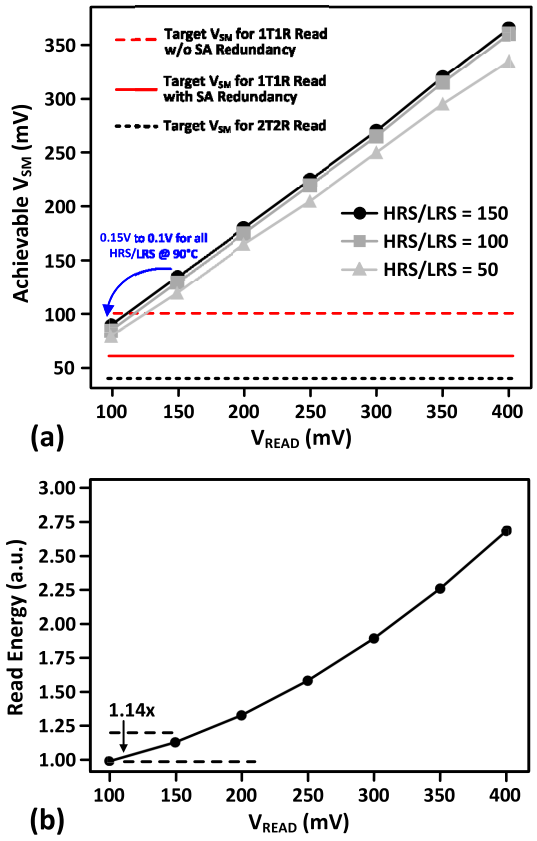


Fig. 20. ReRAM read evaluation. (a) V_{SM} versus V_{READ} at room temperature. (b) Normalized 1T1R read energy versus V_{READ}.

tive load. The drawback is that it requires an additional clock signal. Regarding power consumption, the proposed LiM-FA consumes the least power among the compared FAs based on nonvolatile devices.

D. ReRAM Read and IM-DP Evaluation

Fig. 20(a) shows the achievable V_{SM} versus V_{READ} for ReRAM access. Like LiM primitive operations, the V_{SM} for 1T1R ReRAM access is larger than that of TCAM. For 1T1R ReRAM read, only single-ended sensing using a small SA is allowed. Considering 4-sigma variation for SA offset, the target V_{SM} for 1T1R access is 100 mV without employing SA redundancy, which is achievable with V_{READ} of 150 mV. However, if SA redundancy is employed, the target V_{SM} becomes 60 mV and V_{READ} can be reduced to 100 mV. For 2T2R ReRAM read and IM-DP, the reconfigurable SA allows differential sensing. Therefore, the voltage difference between SL and SLB can be directly utilized as V_{SM} rather than degrades by a factor of 2 for single-ended sensing. In this case, the SA offset sigma becomes 8.9 mV for a large differential SA and the target V_{SM} is ~40 mV. Therefore, V_{READ} can be set to 100 mV which is the same as 1T1R read employing SA redundancy. Fig. 20(b) shows the normalized energy (including peripheral logics) versus V_{READ} for 1T1R ReRAM access. For 1T1R single-ended sensing, SA redundancy reduces V_{READ} from 150 to 100 mV and achieves 1.14× lower access energy.

TABLE VI
COMPARISON WITH RECENT RERAM AND R-CIM WORKS

		This Work	[13]	[18]	[35]
Operations		NVM, TCAM, LiM, IM-DP	NVM, IM-DP	TCAM	NVM
Technology		40nm	130nm	130nm	40nm
On/Off Ratio		~100	N.A.	> 200	N.A.
V_{READ} (V)	TCAM	0.4	-	0.6	-
	LiM	0.15	-	-	-
	IM-DP	0.1	0.1	-	-
	NVM Read	0.1	0.1	-	0.18 - 0.3

E. Discussion and Comparison With Prior Works

Table VI summarizes the V_{READ} used in this work and recent ReRAM and R-CIM works based on voltage-mode sensing. The proposed R-CIM structure supports versatile CIM operations using voltage-mode sensing. There are three V_{READ} required in this work and the area overhead for generating different V_{READ} is found to be only 1.2%.

The work in [13] uses 2T2R ReRAM cells to implement IM-DP during normal memory access with $V_{READ} = 0.1$ V. Although we adopt the IM-DP method in [13] and use the same V_{READ}, we show that $V_{READ} = 0.1$ V is sufficient to provide enough sensing margin for ReRAM read (in both 1T1R and 2T2R configurations) and IM-DP after considering

different types of variations. Note that the lowest $V_{\mathrm{READ}} = 0.1$ V for NVM read will affect the read speed [35]. Fortunately, 0.1 V is far less than the SET/RESET voltage of ReRAM devices (>0.5 V) and it leaves a large space for the tradeoff between speed and power while ensuring the ReRAM reliability. As reported in [35], which is also based on ReRAM integrated with 40-nm technology, the read speed is improved by more than 2× when V_{READ} increases from 0.18 to 0.26 V.

The work in [18] uses $V_{\mathrm{READ}} = 0.6$ V with $R_{\mathrm{H}}/R_{\mathrm{L}}$ ratio $\cong$ during TCAM search. However, with $V_{\mathrm{READ}} = 0.6$ V, significant read disturbance may occur during TCAM search and degrades the ReRAM reliability [31], [36]. On the contrary, we perform optimizations for TCAM using SA redundancy and lower the required V_{READ} to 0.4 V. Although a lower V_{READ} will decrease the search speed, it improves the reliability of ReRAM devices and reduces the search energy. When the robustness of R-CIM systems is of major concern, designers should not just care about speed and power consumption since the reliability of ReRAM devices is also important when performing different CIM operations. Therefore, it is necessary to perform different optimizations to ensure that optimal V_{READ} can be selected without significantly disturbing ReRAM devices. This work provides a guide for such optimizations.

VI. Conclusion

In this article, we proposed a reconfigurable 2T2R ReRAM architecture to support three types of CIM operations: 1) TCAM; 2) LiM; and 3) IM-DP. We proposed a configurable data storage strategy to allow the 2T2R ReRAM to operate as conventional 1T1R ReRAM in situations that CIM is not required. We performed optimizations for the proposed R-CIM using existing and novel design techniques to improve its robustness and efficiency. We quantitatively analyzed the robustness of the proposed R-CIM with respect to the precharge voltage (V_{READ}) and the ReRAM ON/OFF ratio. With the proposed optimizations, the TCAM search energy can be reduced by 1.6× with better reliability thanks to the lower V_{READ}. The proposed LiM-FA improves the delay (3.2×), the static power (1.2×), and the dynamic power (1.6×) compared with the state-of-the-art LiM-FA. Combining optimizations with robustness analysis, the same V_{READ} for ReRAM access can be set in 2T2R and 1T1R configurations. A lower V_{READ} in 1T1R configuration gives 1.14× lower access energy.

References

[1] T.-K.-J. Ting *et al.*, "An 8-channel 4.5 Gb 180GB/s 18ns-row-latency RAM for the last level cache," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2017, pp. 404–405.

[2] J. Wang *et al.*, "A 28-nm compute SRAM with bit-serial logic/arithmetic operations for programmable in-memory vector computing," *IEEE J. Solid-State Circuits*, vol. 55, no. 1, pp. 76–86, Jan. 2020.

[3] M. F. Ali, A. Jaiswal, and K. Roy, "In-memory low-cost bit-serial addition using commodity DRAM technology," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 1, pp. 155–165, Jan. 2020.

[4] Y. Chen, "ReRAM: History, status, and future," *IEEE Trans. Electron Devices*, vol. 67, no. 4, pp. 1420–1433, Apr. 2020.

[5] M. M. S. Aly *et al.*, "The N3XT approach to energy-efficient abundant-data computing," *Proc. IEEE*, vol. 107, no. 1, pp. 19–48, Jan. 2019.

[6] T. F. Wu *et al.*, "A 43pJ/cycle non-volatile microcontroller with 4.7μs shutdown/wake-up integrating 2.3-bit/cell resistive RAM and resilience techniques," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2019, pp. 226–228.

[7] 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 J. Solid-State Circuits*, vol. 51, no. 4, pp. 1009–1021, Apr. 2016.

[8] W.-H. Chen *et al.*, "A 65 nm 1 Mb nonvolatile computing-in-memory ReRAM macro with sub-16 ns multiply- and-accumulate for binary DNN AI edge processors," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2018, pp. 494–495.

[9] C. Yu, T. Yoo, T. T.-H. Kim, K. C. T. Chuan, and B. Kim, "A 16K current-based 8T SRAM compute-in-memory macro with decoupled read/write and 1-5bit column ADC," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Mar. 2020, pp. 1–4.

[10] S. Jain, A. Ranjan, K. Roy, and A. Raghunathan, "Computing in memory with spin-transfer torque magnetic RAM," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 26, no. 3, pp. 470–483, Mar. 2018.

[11] D. Reis, M. Niemier, and X. S. Hu, "Computing in memory with FeFETs," in *Proc. Int. Symp. Low Power Electron. Design*, Jul. 2018, pp. 1–6.

[12] S. K. Thirumala, S. Jain, A. Raghunathan, and S. K. Gupta, "Non-volatile memory utilizing reconfigurable ferroelectric transistors to enable differential read and energy-efficient in-memory computation," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design (ISLPED)*, Jul. 2019, pp. 1–6.

[13] M. Bocquet *et al.*, "In-memory and error-immune differential RRAM implementation of binarized deep neural networks," in *IEDM Tech. Dig.*, Dec. 2018, pp. 20-1–20-6.

[14] C.-X. Xue *et al.*, "A 1 Mb multibit ReRAM computing-in-memory macro with 14.6 ns parallel MAC computing time for CNN based AI edge processors," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2019, pp. 388–389.

[15] W. Wan *et al.*, "A 74 TMACS/W CMOS-RRAM neurosynaptic core with dynamically reconfigurable dataflow and *in-situ* transposable weights for probabilistic graphical models," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2020, pp. 498–499.

[16] L. Zheng, S. Shin, and S.-M.-S. Kang, "Memristors-based ternary content addressable memory (mTCAM)," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Jun. 2014, pp. 2253–2256.

[17] M.-F. Chang *et al.*, "Designs of emerging memory based non-volatile TCAM for Internet-of-Things (IoT) and big-data processing: A 5T2R universal cell," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2016, pp. 1142–1145.

[18] D. R. B. Ly *et al.*, "In-depth characterization of resistive memory-based ternary content addressable memories," in *IEDM Tech. Dig.*, Dec. 2018, pp. 20-1–20-3.

[19] J. Li, R. K. Montoye, M. Ishii, and L. Chang, "1 Mb 0.41 μm^2 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing," *IEEE J. Solid-State Circuits*, vol. 49, no. 4, pp. 896–907, Apr. 2014.

[20] Y. Zha, E. Nowak, and J. Li, "Liquid silicon: A nonvolatile fully programmable processing-in-memory processor with monolithically integrated ReRAM for big data/machine learning applications," in *Proc. Symp. VLSI Circuits*, Jun. 2019, pp. C206–C207.

[21] M.-F. Chang *et al.*, "Challenges and circuit techniques for energy-efficient on-chip nonvolatile memory using memristive devices," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 5, no. 2, pp. 183–193, Jun. 2015.

[22] 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, Dec. 2015.

[23] R. Waser, R. Dittmann, G. Staikov, and K. Szot, "Redox-based resistive switching memories–nanoionic mechanisms, prospects, and challenges," *Adv. Mater.*, vol. 21, nos. 25–26, pp. 2632–2663, Jul. 2009.

[24] K.-S. Li *et al.*, "Utilizing sub-5 nm sidewall electrode technology for atomic-scale resistive memory fabrication," in *Symp. VLSI Technol. (VLSI-Technol.), Dig. Tech. Papers*, Jun. 2014, pp. 1–2.

[25] H. Y. Lee *et al.*, "Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO_2 based RRAM," in *IEDM Tech. Dig.*, Dec. 2008, pp. 1–4.

[26] W. Kim *et al.*, "Forming-free nitrogen-doped AlOX RRAM with sub-μA programming current," in *IEEE Symp. VLSI Technol. Dig. Tech. Papers*, Jun. 2011, pp. 22–23.

[27] A. Grossi *et al.*, "Fundamental variability limits of filament-based RRAM," in *IEDM Tech. Dig.*, Dec. 2016, pp. 4-1–4-7.

[28] E. Vianello *et al.*, "Resistive memories for ultra-low-power embedded computing design," in *IEDM Tech. Dig.*, Dec. 2014, pp. 6-1–6-3.

[29] A. Fantini *et al.*, "Intrinsic program instability in HfO_2 RRAM and consequences on program algorithms," in *IEDM Tech. Dig.*, Dec. 2015, pp. 7-1–7-5.
[30] 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.
[31] W. C. Chien *et al.*, "A forming-free WO_x resistive memory using a novel self-aligned field enhancement feature with excellent reliability and scalability," in *IEDM Tech. Dig.*, Dec. 2010, pp. 19-1–19-2.
[32] B. Govoreanu *et al.*, "10×10 nm^2 Hf/HfO_x crossbar resistive ram with excellent performance, reliability and low-energy operation," in *IEDM Tech. Dig.*, Dec. 2011, pp. 31-1–31-6.
[33] P. Jain *et al.*, "A 3.6 Mb 10.1Mb/mm^2 embedded non-volatile ReRAM macro in 22 nm FinFET technology with adaptive forming/set/reset schemes yielding down to 0.5 V with sensing time of 5ns at 0.7 V," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2019, pp. 212–213.
[34] O. Golonzka *et al.*, "Non-volatile RRAM embedded into 22FFL FinFET technology," in *Proc. Symp. VLSI Technol.*, Jun. 2019, pp. 230–231.
[35] C.-C. Chou *et al.*, "An N40 256K × 44 embedded RRAM macro with SL-precharge SA and low-voltage current limiter to improve read and write performance," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2018, pp. 478–479.
[36] H. Lv *et al.*, "BEOL based RRAM with one extra-mask for low cost, highly reliable embedded application in 28 nm node and beyond," in *IEDM Tech. Dig.*, Dec. 2017, pp. 2–4.
[37] M.-F. Chang *et al.*, "Embedded 1Mb ReRAM in 28 nm CMOS with 0.27-to-1 V read using swing-sample-and-couple sense amplifier and self-boost-write-termination scheme," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2014, pp. 332–333.
[38] Q. Liu *et al.*, "A fully integrated analog ReRAM based 78.4TOPS/W compute-in-memory chip with fully parallel MAC computing," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2020, pp. 500–501.
[39] Q. Guo, X. Guo, Y. Bai, and E. Ipek, "A resistive TCAM accelerator for data-intensive computing," in *Proc. 44th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Dec. 2011, pp. 339–350.
[40] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in *Proc. 30th Conf. Neural Inf. Process. Syst. (NIPS)*, Dec. 2016, pp. 4107–4115.
[41] N. Verma and A. P. Chandrakasan, "A 65 nm 8T sub-Vt SRAM employing sense-amplifier redundancy," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2007, pp. 328–329.
[42] M. Poremba, S. Mittal, D. Li, J. S. Vetter, and Y. Xie, "DESTINY: A tool for modeling emerging 3D NVM and eDRAM caches," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2015, pp. 1543–1546.
[43] M. E. Sinangil and A. P. Chandrakasan, "Application-specific SRAM design using output prediction to reduce bit-line switching activity and statistically gated sense amplifiers for up to 1.9× lower energy/access," *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 107–117, Jan. 2014.
[44] S. Matsunaga *et al.*, "Fabrication of a nonvolatile full adder based on logic-in-memory architecture using magnetic tunnel junctions," *Appl. Phys. Express*, vol. 1, Aug. 2008, Art. no. 091301.
[45] Z. Wang *et al.*, "A physics-based compact model of ferroelectric tunnel junction for memory and logic design," *J. Phys. D, Appl. Phys.*, vol. 47, no. 4, Dec. 2013, Art. no. 045001.
[46] X. Yin, X. Chen, M. Niemier, and X. S. Hu, "Ferroelectric FETs-based nonvolatile logic-in-memory circuits," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 1, pp. 159–172, Jan. 2019.
[47] Deepa and V. S. Kumar, "Analysis of energy efficient PTL based full adders using different nanometer technologies," in *Proc. 2nd Int. Conf. Electron. Commun. Syst. (ICECS)*, Feb. 2015, pp. 310–315.

Yuzong Chen received the B.Eng. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2019.

He is currently a Project Officer with the Centre for Integrated Circuits and Systems (CICS), Nanyang Technological University. His research interests include resistive random access memory (ReRAM) circuits design and in-memory computing.

Lu Lu (Student Member, IEEE) received the B.E. degree from the School of Computer and Information, Hefei University of Technology, Hefei, China, in 2007, and the M.E. degree from the School of Microelectronics and Solid-State Electronics, Xiamen University, Xiamen, China, in 2010. She is currently working toward the Ph.D. degree at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

Her research interests include low-power SRAM and SRAM-based physical unclonable function (PUF).

Ms. Lu was a recipient of the IEEE SSCS Singapore Chapter Award in 2018.

Bongjin Kim (Member, IEEE) received the B.S. and M.S. degrees from POSTECH, Pohang, South Korea, in 2004 and 2006, respectively, and the Ph.D. degree from the University of Minnesota, Minneapolis, MN, USA, in 2014.

He spent two years with Rambus, Sunnyvale, CA, USA, where he was a Senior Staff Member and worked on the research of high-speed serial link circuits and microarchitectures. He worked as a Postdoctoral Fellow with Stanford University, Stanford, CA, for a year. From 2006 to 2010, he was with Samsung Electronics, Yongin, South Korea, where he performed the research on clock generators for high-speed serial links. He was also a Research Intern with Texas Instruments, Dallas, TX, USA, IBM TJ Watson Research, Yorktown Heights, NY, USA, and Rambus, during his Ph.D., from 2012 to 2014. He joined Nanyang Technological University (NTU), Singapore, in September 2017, as an Assistant Professor. His current research interests include memory-centric computing circuits and architectures, hardware accelerators, and mixed-signal circuit design techniques and methodologies.

Dr. Kim was a recipient of the Prestigious Doctoral Dissertation Fellowship Award for his Ph.D. study, the Low Power Design Contest Award from ISLPED, and the Intel/IBM/Catalyst Foundation Award from CICC Conference. His research works appeared at top circuit conferences and journals including ISSCC, IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, CICC, ESSCIRC, and IEEE JOURNAL OF SOLID-STATE CIRCUITS (JSSC).

Tony Tae-Hyoung Kim (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Korea University, Seoul, South Korea, in 1999 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Minnesota, Minneapolis, MN, USA, in 2009.

From 2001 to 2005, he was with Samsung Electronics, Hwasung, South Korea, where he performed the research on the design of high-speed SRAM memories, clock generators, and IO interface circuits. From 2007 to 2009, he was with the IBM T. J. Watson Research Center, Yorktown Heights, NY, USA, and Broadcom Corporation, Edina, MN, USA, where he performed the research on circuit reliability, low-power SRAM, and battery-backed memory design. In 2009, he joined Nanyang Technological University, Singapore, where he is currently an Associate Professor. He has authored or coauthored over 160 journal and conference articles and holds 17 U.S. and Korean patents registered. His current research interests include low-power and high-performance digital, mixed-mode, and memory circuit design, ultralow-voltage circuits and systems design, variation and aging-tolerant circuits and systems, and circuit techniques for 3-D ICs.

Dr. Kim received the Best Demo Award at APCCAS2016, the Low Power Design Contest Award at ISLPED2016, the best paper awards at 2014 and 2011 ISOCC, the AMD/CICC Student Scholarship Award at the IEEE CICC2008, the Departmental Research Fellowship from the University of Minnesota in 2008, the DAC/ISSCC Student Design Contest Award in 2008, the Samsung Humantec Thesis Award in 2008, 2001, and 1999, and the ETRI Journal Paper of the Year Award in 2005. He was the Chair of the IEEE Solid-State Circuits Society Singapore Chapter. He has served on numerous conferences as a Committee Member. He serves as an Associate Editor for the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, IEEE ACCESS, and the IEIE *Journal of Semiconductor Technology and Science*.