IEEE Open Journal of
Circuits and Systems

Received 25 July 2020; revised 10 November 2020; accepted 27 November 2020. Date of current version 26 January 2021.

Digital Object Identifier 10.1109/OJCAS.2020.3042550

A Reconfigurable 4T2R ReRAM Computing In-Memory Macro for Efficient Edge Applications

YUZONG CHEN [1], LU LU[1] (Student Member, IEEE), BONGJIN KIM [2] (Member, IEEE), AND TONY TAE-HYOUNG KIM [1] (Senior Member, IEEE)

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798

[2]Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106, USA

This article was recommended by Associate Editor X. Zhang.

CORRESPONDING AUTHOR: Y. CHEN (e-mail: yuzong.chen@ntu.edu.sg)

This work was supported by a RIE2020 ASTAR AME IAF-ICP under Grant I1801E0030.

ABSTRACT Resistive random access memory (ReRAM)-based computing in-memory (CIM) is a promising solution to overcome the von-Neumann bottleneck in conventional computing architectures. We propose a reconfigurable ReRAM architecture using a novel 4T2R bit-cell that supports non-volatile storage and two types of CIM operations: i) ternary content addressable memory (TCAM) and ii) in-memory dot product (IM-DP) for neural networks. The proposed 4T2R cell occupies a smaller area than prior SRAM-based CIM bit-cells. A 128 × 128 ReRAM macro is designed in 40nm CMOS technology. For TCAM operations, it allows a search word-length of 128 bits. For IM-DP operations, it can compute parallel dot products using binary inputs and ternary weights. The simulated search delay for TCAM operation is 0.92 ns at VDD = 0.9 V and the simulated energy efficiency for IM-DP operation is 223.6 TOPS/W at VDD = 0.7 V. Monte-Carlo simulations show a standard deviation of 4.9% in accumulate operation for IM-DP which corresponds to a classification accuracy of 95.7% on the MNIST dataset and 81.7% on the CIFAR-10 dataset.

INDEX TERMS Resistive random access memory (ReRAM), ternary content addressable memory (TCAM), computing in-memory (CIM), reconfigurable architecture.

I. INTRODUCTION

THE TREMENDOUS demand for data-intensive applications such as machine learning and big-data processing in our daily life has motivated the development of efficient edge computing which has limited delay and energy budgets. Conventional von-Neumann architectures suffer from long latency and high power consumption due to data movements between off-chip memory and arithmetic-logic units (ALUs). As shown in Fig. 1, a typical ALU operation (e.g., 32-bit integer addition) takes less than 1ns and consumes less than 1pJ while a data movement from off-chip memory can cost tens of nanoseconds and a few nanojoules [1], [2].

In recent years, alternative solutions for more efficient computing such as beyond CMOS technologies and beyond von-Neumann architectures have received much attention. Among those, ReRAM [3] is a potential non-volatile memory (NVM) candidate for the next-generation storage system with fast read/write speed, low programming voltage, and good scalability. Computing in-memory (CIM) [4]–[13] is an attractive solution to reduce the energy and latency cost of memory access by performing specific computations directly inside the memory macro without reading out operands and sending to ALUs. For example, ternary content addressable memory (TCAM) [4]–[8] is a critical component to achieve fast searches. Rather than reading out data row-by-row and sending to ALUs for comparison, TCAM performs bit-wise XOR/XNOR between the search key and all stored data to get the match result in one cycle. Besides TCAM, deep neural networks (DNNs) also cost considerable delay and power using traditional computing paradigms due to frequent memory fetch to perform the dot product (also called multiply-and-accumulate, or MAC) operation. CIM solutions for DNNs [9]–[13] can improve the throughput and energy efficiency by performing massively parallel MAC operations inside the memory array, eliminating costly data transfer.

Given the benefits of ReRAM and CIM, it brings significant value to implement reliable ReRAM-based CIM (R-CIM) systems that can accelerate versatile functions. However, many prior CIM works [5]–[12] only focus on one specific CIM function. Moreover, most ReRAM-based TCAM [6]–[8] cannot support row-wise memory access due to the conflicts between shared horizontal match-lines in TCAM and shared vertical bit-lines in conventional memory architectures. Although conventional 6T SRAM can support TCAM operations by adding additional access transistors, the large cell area overhead (e.g., two 10T cells to store 1b TCAM data in [5]) brings a great challenge to high-capacity TCAM systems. An interesting idea in [4] implements TCAM using standard push-rule 6T SRAM to save area and it reuses memory bit-lines as match-lines in TCAM mode. But it requires data words to be stored row-wise in SRAM mode and column-wise in TCAM mode, resulting in complicated word placement and data reorganization when performing different types of operations.

To tackle these challenges, we propose a reconfigurable R-CIM macro using a novel 4T2R bit-cell. The key idea is to add two access transistors to a differential 2T2R ReRAM bit-cell that has been well studied [7], [14]. The proposed R-CIM structure can function as a non-volatile storage system as well as accelerators for TCAM and IM-DP operations. Moreover, it stores data words row-wise for all types of operations, eliminating complicated data organization, and increasing the flexibility. We particularly consider the reliability issue associated with R-CIM due to the pseudo-write condition during read operations and present a solution to handle such an issue. As a result, we can use a higher VDD for CIM operations, improving the performance without disturbing ReRAM devices. The proposed R-CIM structure is designed in 40nm technology. Simulation results of a 128×128 array show that the TCAM search delay is 0.92 ns at VDD $=$ 0.9 V. The energy efficiency of IM-DP operations is 223.6 TOPS/W at VDD $=$ 0.7 V. Monte-Carlo simulation is performed for the accumulate operation in IM-DP mode and shows a standard deviation of 4.91% which corresponds to 95.7% accuracy for the MNIST dataset [15].

The rest of this article is organized as follows: Section II provides a background of ReRAM technology and existing CIM works. Section III describes the proposed reconfigurable 4T2R R-CIM architecture and explains its operations in different modes. In Section IV, we propose some optimizations at circuit and device levels to the proposed R-CIM architecture. In Section V, comprehensive simulation results are presented based on a 128 x 128 array. Finally, we conclude this article in Section VI.

II. BACKGROUND

A. RERAM TECHNOLOGY AND SIMULATION MODEL

A ReRAM device is typically a 3-layer device 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 has two types of switching modes: i) unipolar switching where the switching direction depends on the amplitude of the applied programming voltage but not on the voltage polarity; ii) bipolar switching where the switching direction depends on both the amplitude and the polarity of the programming voltage. This article will focus on the bipolar ReRAM device. One common way to realize a ReRAM bit-cell is to integrate a ReRAM device with one transistor (1T1R). Fig. 2(b) 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. Fig. 2(c) shows an example I-V curve for a bipolar HfOx-based ReRAM device [16].

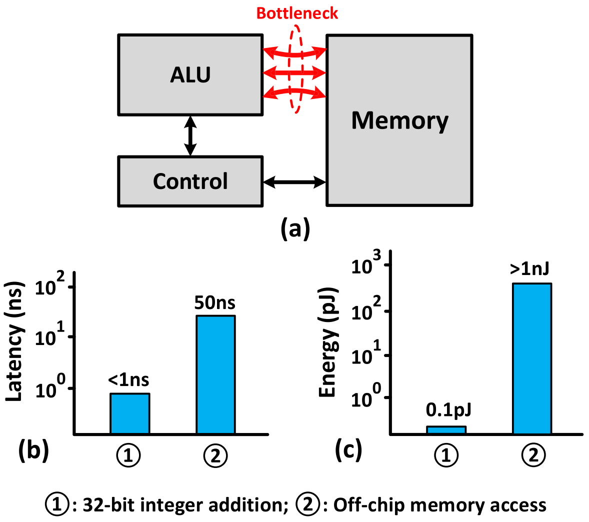
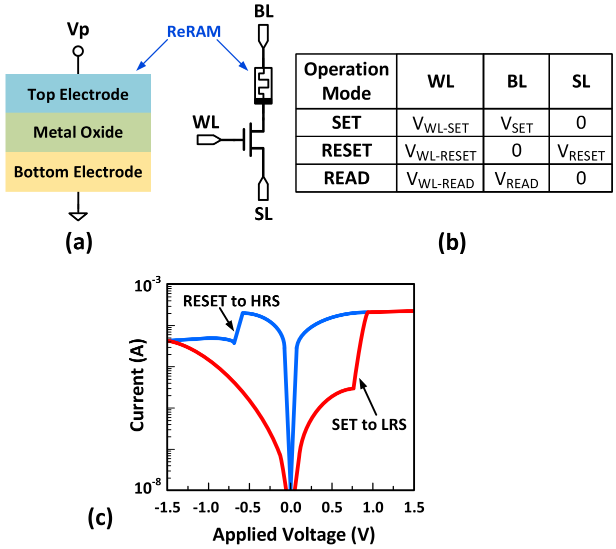


FIGURE 1. Traditional von-Neumann architectures: (a) block diagram, (b) latency and (c) energy cost of a 32-bit integer ALU addition and data movements.



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

FIGURE 2. (a) 3-layer ReRAM device; (b) schematic and biasing conditions for different operations of the 1T1R bipolar ReRAM bit-cell; (c) bipolar resistance switching characteristic of an HfOx ReRAM device [16].

TABLE 1. Key ReRAM parameters.

Device Parameters	Values	Circuit Parameters	Values
HRS	1MΩ	$V_{WL\text{-}SET/RESET}$	1.5 V / 1.4V
LRS	10kΩ	$V_{WL\text{-}READ}$	1.2 V
Set Voltage	0.7 V	V_{READ}@ BL	0.3 V
Reset Voltage	0.7 V		

For circuit-level simulation and analysis, we developed a Verilog-A model for the ReRAM device based on the conductive filament switching mechanism [17]. The I-V relationship of the ReRAM model can be expressed as:

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

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 that can be tuned for a specific ReRAM I-V characteristic. The ReRAM device parameters used in this work are adopted from [16] and the ReRAM devices are integrated with transistors in 40nm technology to form a 4T2R bit-cell. Key parameters for simulation are summarized in Table 1.

B. RELATED WORKS

Many works have been proposed to implement efficient CIM systems based on volatile memories such as SRAM and embedded DRAM (eDRAM). Do *et al.* [5] use two 10T SRAM cells to implement TCAM and employ an efficient match-line scheme to reduce power consumption. Jeloka *et al.* [4] use standard push-rule 6T SRAM cells as TCAM to reduce the cell area, For IM-DP, several novel bit-cells have been proposed. XNOR-SRAM [9] employs a 12T bit-cell to compute MAC based on the resistive voltage divider formed by access transistors. Yu *et al.* [10] use an 8T bit-cell to support current-mode accumulation. An interesting idea in [11] takes advantage of the compact cell size of eDRAM and uses a small 4T dual eDRAM cell to implement IM-DP.

Several CIM works based on non-volatile memories such as ReRAM have also been reported. Huang *et al.* [6] propose a 4T2R TCAM cell based on RC-filters to reduce the ReRAM stress during search operations. Ly *et al.* [7] extensively characterize a ReRAM-based 2T2R TCAM circuit. Chang *et al.* [8] propose a 3T1R TCAM based on multi-level ReRAM cell to achieve high density. Regarding IM-DP, Chen *et al.* [12] use two 1T1R ReRAM bit-cells to store ternary weights (+1, −1, 0) of neural networks and implements MAC based on current accumulation. Zha *et al.* [13] propose a multi-functional R-CIM system with customized data mapping to support ReRAM, TCAM and IM-DP operations.

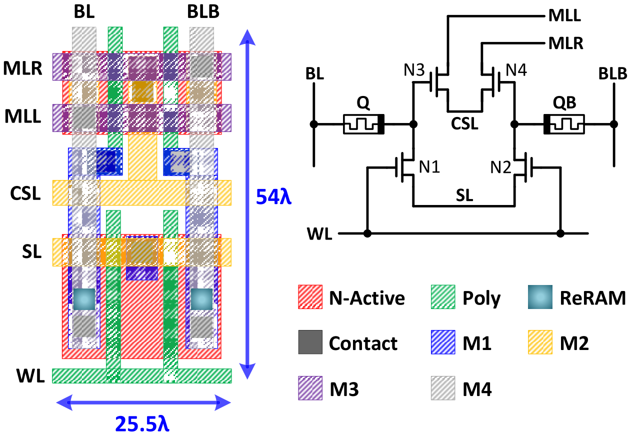


FIGURE 3. Layout (left) and schematic (right) of the proposed 4T2R ReRAM bit-cell. λ: half feature-size.

Most of these prior works only focus on CIM for one specific function [5]–[12]. For SRAM-based CIM, the large cell area of SRAM bit-cell (6T to 12T) prevents it from supporting large-capacity CIM systems. Moreover, the volatile nature of SRAM necessitates a constant power supply to maintain the stored data, resulting in high leakage power during the standby mode. On the contrary, ReRAM-based CIM can be completely powered off during standby mode thanks to the non-volatile feature of ReRAM. However, prior ReRAM-based TCAM [6]–[8] cannot operate as a conventional memory system due to conflicts between shared horizontal match-lines in TCAM and shared vertical bit-lines in normal memory architectures.

III. PROPOSED 4T2R R-CIM ARCHITECTURE

This section presents the detailed structure of the proposed 4T2R R-CIM system. It first describes the structure of the novel 4T2R bit-cell and then explains how different operations (NVM, TCAM and IM-DP) of the 4T2R R-CIM system can be achieved.

A. STRUCTURE OF 4T2R RERAM BIT-CELL

Fig. 3 shows the layout and schematic of the proposed 4T2R bit-cell. In the schematic, two transistors (N1 and N2) and two ReRAM devices (Q and QB) form a differential 2T2R bit-cell which has been employed in CIM by several prior works [7], [14]. The differential 2T2R bit-cell has two bit-lines (BL and BLB) shared by each column while the source-line (SL) and the word-line (WL) are shared by each row. The ReRAM device pair in the bit-cell represents data ‘1’ with (Q, QB) = (HRS, LRS) and data ‘0’ with (Q, QB) = (LRS, HRS). Two additional access transistors (N3 and N4) give two horizontal match-lines (MLL and MLR) and a compute source-line (CSL) shared by each row. To satisfy the write current requirements for large R-ratio (the ratio between HRS and LRS of the ReRAM device) and long retention, N1 and N2 need to exceed the minimum transistor size. They need to be sized to provide the typical ReRAM programming current of 50 μA [18]. In this work, we choose W/L = 3 for N1 and N2 in the employed 40nm technology. For N3 and N4, we choose W/L = 2 to improve the performance and prevent CIM operations from being

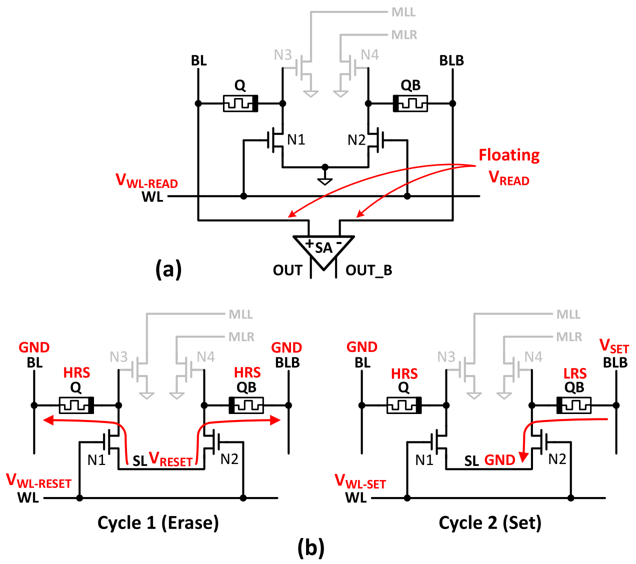


FIGURE 4. (a) Read operation of the 4T2R bit-cell using voltage-mode sensing; (b) two-cycle write operation for writing data '1.'

sensitive to local process variations. Nevertheless, the used transistor sizes still make the proposed 4T2R bit-cell smaller than other SRAM-based CIM bit-cells because of fewer transistors. The area of the 4T2R bit-cell is 0.55μm^2 (i.e., 54$\lambda \times$ 25.5λ where λ is the half feature-size) using the logic design rule. Compared with the XNOR-SRAM bit-cell [9], it is 2.7× smaller (normalized). Besides, the proposed bit-cell is 1.27x smaller (normalized) when compared with the 8T SRAM bit-cell in [10].

B. READ/WRITE OPERATIONS FOR NVM MODE

Since the 4T2R bit-cell is built on top of the differential 2T2R bit-cell [7], [14] which is formed by two 1T1R bit-cells that share a common SL, the biasing conditions for read and write operations in NVM mode are similar to that of the 1T1R bit-cell. N3 and N4 are not used in NVM mode so they can be turned off by grounding MLL/MLR and CSL. Fig. 4(a) explains the read operation using voltage-mode sensing. To read a word, BL and BLB are precharged to V_{READ} and then left floating. When WL is turned on, BL and BLB will discharge at different rates based on the states of Q and QB. A differential sense amplifier (SA) detects the voltage difference between BL and BLB and outputs the data result.

The write operation takes two cycles to program the ReRAM pair. Fig. 4(b) shows an example where data '1' is written to the 4T2R bit-cell. The same biasing conditions for writing the 1T1R bit-cell can be applied. But since two 1T1R bit-cells share a common SL, writing the second ReRAM device may disturb the first ReRAM device which has been programmed in the previous cycle. Therefore, we propose to erase both ReRAM devices by resetting them in cycle 1. In cycle 2, depending on the data to be written, either Q is set to LRS if writing data '0' or QB is set to LRS if writing data '1'. The HRS device in cycle 2 is not disturbed since the corresponding bit-line and SL are grounded.

TABLE 2. Biasing and mismatch behavior for TCAM search.

Search Data	BL	BLB	Mismatch Behavior
'1'	VDD	0V	MLL discharges
'0'	0V	VDD	MLR discharges

C. TCAM OPERATIONS

The proposed 4T2R bit-cell can operate as a TCAM to achieve fast search. The TCAM words are stored row-wise as in NVM mode. This offers better flexibility than [4] that requires row-wise storage in SRAM mode and column-wise storage in TCAM mode. In addition to states '1' and '0' that can be represented in the same way as in NVM mode, the "don't care" state 'X' in TCAM mode is represented by (Q, QB) = (HRS, HRS). To write 'X', only the erase operation (cycle 1 of the write operation in NVM mode) is performed. This means that the write circuits can be shared by both NVM and TCAM modes, reducing the hardware overhead.

Fig. 5 illustrates a search operation where the stored data is "X-0" and the search data is "0-1". In standby mode, MLL and MLR are precharged to VDD while other signals are grounded. If the search operation is not frequent, CSL can also be biased to VDD to reduce leakage power of N3 and N4. During search operations, MLL and MLR are disconnected from the precharge circuit. SL and CSL are grounded while WL is turned on. Then search data and inverted search data are applied to BL and BLB, respectively as shown in Table 2. One 1T1R path (Q-N1 or QB-N2) will form a voltage divider to bias the gate voltage of N3 or N4 while the other 1T1R path is off. If stored data and search data are matched, the 1T1R voltage divider will contain an NMOS and an HRS device so that the gate voltage of N3 or N4 is below the threshold voltage of NMOS (V_G, N4[0] in Fig. 5). Therefore, both MLL and MLR maintain at the precharged voltage. If there is a mismatch, the 1T1R voltage divider will contain an NMOS and an LRS device so that the gate voltage of N3 or N4 is above the threshold voltage of NMOS ($V_{G,N3[1]}$ in Fig. 5). Therefore, either MLL or MLR or both MLL and MLR will discharge to the ground. Two SAs compare MLL and MLR voltages with a reference voltage V_{REF} separately and the comparison results are connected to an AND gate to produce the match result.

The separation of MLL and MLR provides considerable benefits to the TCAM performance. In the worst 1-bit mismatch case, only one NMOS device discharges in a row. In the conventional SRAM-based TCAM (sTCAM) implementation [5], there is only one match-line (ML) in each row and each bit-cell has two NMOS transistors connected to the ML, resulting in larger ML capacitance. On the other hand, the proposed 4T2R TCAM cell separates MLL and MLR so each bit-cell only has one NMOS connected

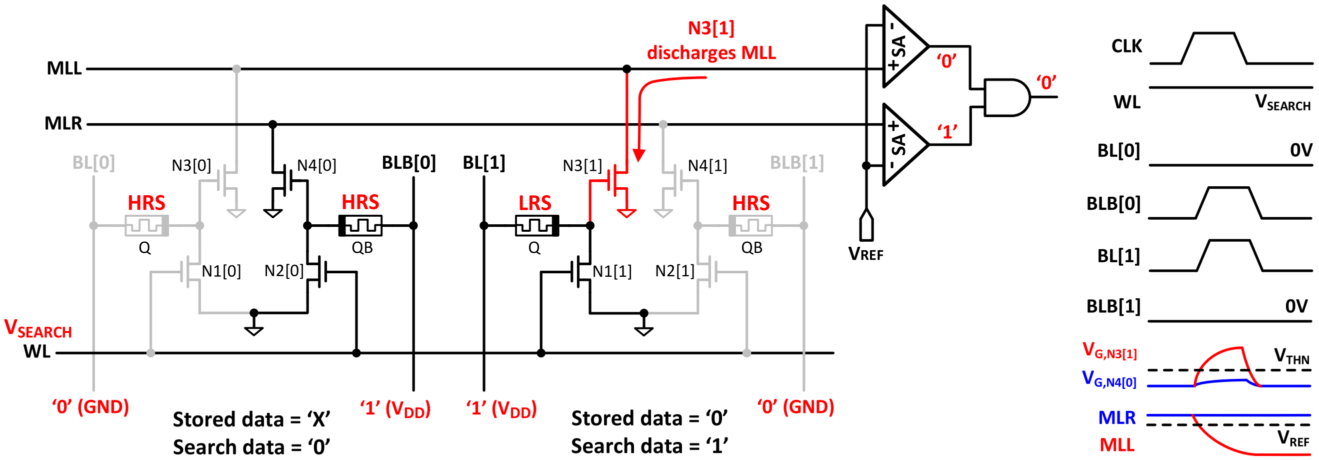


FIGURE 5. TCAM search operations. The left cell is a match while the right cell is a mismatch. MLL discharges below V_{REF}.

to an ML. Since two MLs are sensed separately, the smaller capacitance on each ML gives faster discharge speed. Of course, the 1T1R voltage divider makes the gate voltage of N3 or N4 lower than VDD during a mismatch, and the AND gate after two SAs brings additional delay. Nevertheless, given enough R-ratio (e.g., 100), the proposed 4T2R TCAM can still achieve comparable performance to sTCAM for a long word-length as shown later in Section V.

D. IM-DP OPERATIONS

Deep neural networks (DNNs) are powerful tools to achieve state-of-the-art results for many artificial intelligence applications such as computer vision and natural language processing. However, traditional DNNs incur a high storage overhead due to large network sizes and high bit precisions (e.g., 32-bit floating-point). Recently introduced bitwise neural networks [19] significantly reduce DNNs' storage requirements by restricting the input activation to 0/1 and the weight to $+1/-1$ with marginal accuracy degradation compared with original full-precision DNNs [20]. In [21], it is also reported that extending the binary weight (+1 and −1) to ternary weight (+1, −1 and 0) demonstrated higher classification accuracy than full-precision DNNs in MNIST and CIFAR-10 datasets. The proposed R-CIM system is able to compute in-memory dot products (IM-DP) for a binary-input ternary-weighted (BITW) network [12]. The BITW network combines ternary weights (+1, 0 and −1) and a modified binary input (1 and 0). Compared with binarized neural networks [22] that use binary inputs +1 and −1, the modified binary input only causes 0.17% and 1.49% accuracy degradation on MNIST and CIFAR-10 datasets, respectively [20].

Fig. 6 shows the operation principle of the proposed R-CIM structure in IM-DP mode. SL is grounded and WL is biased to V_{DP}. A ternary weight ((Q, QB) = (HRS, LRS) for '+1', (Q, QB) = (LRS, HRS) for '−1', and (Q, QB) = (HRS, HRS) for '0') is stored in the 4T2R bit-cell. A binary input is applied through BL and BLB. Fig. 6(a)-(e) shows six possible states of the proposed 4T2R bit-cell with binary weight/input combinations. If the input is '0', BL and BLB are grounded. The gate voltages of N3 and N4 are both '0' so there is no discharge in MLL and MLR as shown in Fig. 6(a)-(c). If the input is '1', BL and BLB are biased to VDD. The 1T1R voltage divider that contains an HRS device will make the gate voltage of N3 or N4 close to '0' so no discharge occurs on MLL or MLR. On the other hand, the 1T1R voltage divider that contains an LRS device will make the gate voltage of N3 or N4 higher enough to discharge MLL or MLR. As a result, when the binary weight is '0', no discharge occurs on MLL or MLR; when the binary weight is '+1', N4 will discharge MLR; when the binary weight is '−1', N3 will discharge MLL. The voltage difference between MLL and MLR (+ΔV, −ΔV, 0) represents the product of input and weight as shown in Fig. 6(d)-(f). Table 3 summarizes the relationship between inputs/weights and the resulting outputs in both computed values and signal representations.

Fig. 7 illustrates the accumulation principle of the proposed R-CIM structure in IM-DP mode. Fig. 7(a) shows a row containing 128 bit-cells for a dot product with 128 input-weight pairs. All inputs are applied through BL and BLB to bias the gate voltages of N3 and N4 simultaneously. However, this will cause considerable switching noise as 128 NMOS may be all turned on at the same time to discharge MLL/MLR. Therefore, we use CSL to control the discharge of MLL/MLR since one CSL is shared by the whole row. Fig. 7(b) shows the principle of the accumulation operation on MLL/MLR. During standby mode, MLL/MLR and CSL are biased to VDD to turn off N3 and N4. During the active time, the inputs are applied to BL and BLB slightly earlier to settle the gate voltages of N3 and N4. Then a negative short pulse is applied to CSL to bias it to ground to discharge the capacitance of the match-line (C_{ML}). After the pulse, CSL is biased to VDD again and all BLs and BLBs are grounded. A voltage difference is developed between MLL and MLR to represent the dot

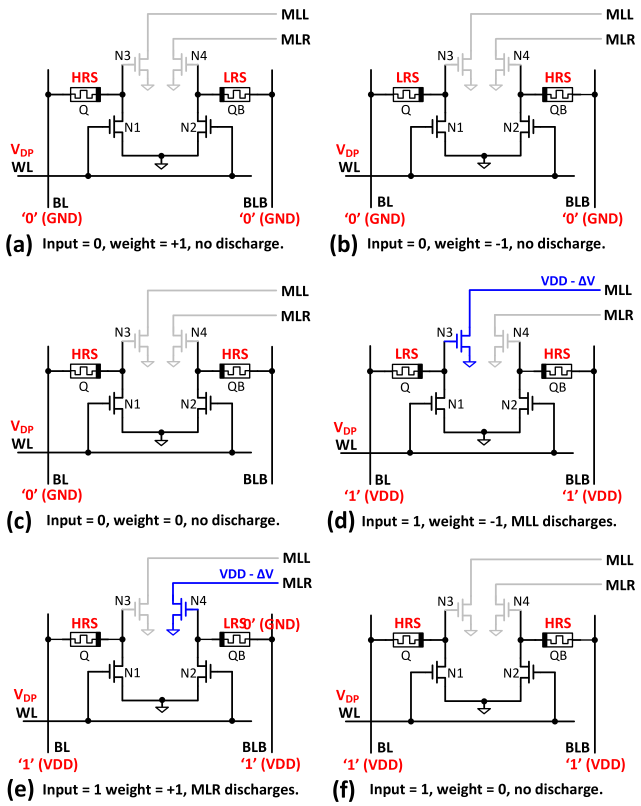

MLL
MLR
N3
N4
HRS
LRS
Q
QB
N1
N2
V_DP
WL
BL
BLB
'0' (GND)
'1' (VDD)
VDD - ΔV
(a) Input = 0, weight = +1, no discharge.
(b) Input = 0, weight = -1, no discharge.
(c) Input = 0, weight = 0, no discharge.
(d) Input = 1, weight = -1, MLL discharges.
(e) Input = 1 weight = +1, MLR discharges.
(f) Input = 1, weight = 0, no discharge.

FIGURE 6. Operation principles of the R-CIM architecture in IM-DP mode. (a)-(e) Six possible weight/input combinations. The 4T2R bit-cell stores a ternary weight (+1, −1 and 0) and a binary input (0 or 1) is applied to BL/BLB.

TABLE 3. Relationship between the computaiton outputs and circuit operation behaviors.

Input (BL, BLB) Weight (Q,QB)	0 (BL=0V, BLB=0V)	1 (BL=V_{DD}, BLB=V_{DD})
+1 (Q=HRS, QB=LRS)	0 (No change)	+1 (+ΔV)
-1 (Q=LRS, QB=HRS)	0 (No change)	-1 (-ΔV)
0 (Q=HRS, QB=HRS)	0 (No change)	0 (No change)

(): Circuit behavior representation

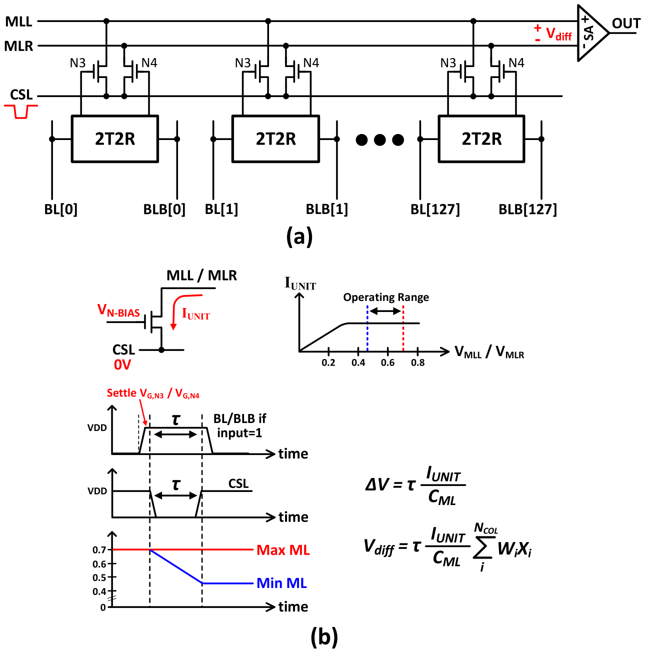


FIGURE 7. Accumulation principle for IM-DP operations: (a) Structure of one row with 128 cells to perform IM-DP operations; (f) principle of the current accumulation on MLL and MLR to represent the dot product result.

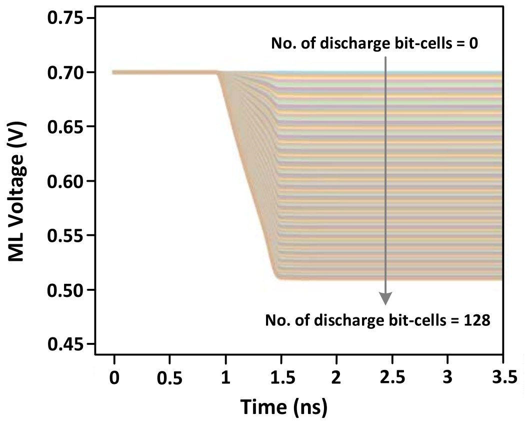


FIGURE 8. Simulation results for MLL/MLR discharge showing all possible 128 voltage levels.

product result, and a binary activation can be performed by a differential SA in each row to get 1-bit outputs.

It should be noted that the unit current (I_{UNIT}) drawn by a single NMOS (N3 or N4) is affected by the match-line voltage. As the match-line voltage drops, the variation in I_{UNIT} increases. Moreover, N3 and N4 of deactivated cells also start to push more leakage current. Both unit current variation and leakage current degrade the accuracy of IM-DP operations. Therefore, the dynamic range of MLL/MLR is selected to be ∼200 mV (0.5-to-0.7 V) to reduce the effects of I_{UNIT} variation and leakage current. Besides, VDD is set to 0.7 V for IM-DP mode to better control the dynamic range of MLL/MLR. Fig. 8 shows the simulation result for all possible voltage levels of a match-line with 128 bit-cells connected in a single row. The pulse width of CSL is $\tau = 500$ ps. The match-line discharges at a rate proportional to the dot product result as mentioned above and finally settles to an intermediate voltage level.

The differential sensing scheme in IM-DP mode only supports 1-bit outputs, which can be accurate enough for simple image classification tasks. For example, the work in [23] achieves 95.1% classification accuracy on MNIST using bitwise neural networks with 1-bit outputs. For more complicated image classification tasks like CIFAR-10, 1-bit outputs will cause a significant accuracy degradation [20]. The proposed architecture can be extended to support for a larger or deeper network by employing an accelerator chip with multiple unit-macros [24] and modifications in the read-out circuits. Multi-bit partial sums in each array can be generated through single-ended sensing with different reference voltages. To achieve this feature, both MLL and

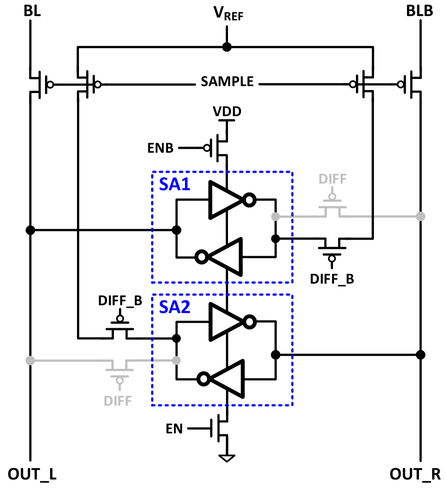


FIGURE 9. Schematic of the reconfigurable sense amplifier employed in [4].

MLR are connected to single-ended SAs as in the TCAM mode. The input is applied to BL, and BLB is always grounded. The corresponding MLL will therefore generate an intermediate voltage level as shown in Fig. 8. The multi-bit output can be obtained by a 1-bit SA with different V_{REF} in successive sensing cycles [23], and 3-bit outputs with 8 quantization levels can provide satisfactory accuracy for the CIFAR-10 dataset [20].

IV. OPTIMIZATIONS OF THE PROPOSED R-CIM

This section presents two optimizations to the proposed 4T2R R-CIM structure. One optimization focuses on the circuit-level design and another particularly addresses the reliability issue of ReRAM devices.

A. RECONFIGURABLE SENSE AMPLIFIER

As described in Section III, TCAM search operations require two single-ended SAs to compare MLL and MLR voltages with a reference voltage V_{REF} separately, while IM-DP operations require differential SAs or single-ended SAs to compare MLL, MLR voltages and V_{REF}. This indicates that three different SAs are needed in each row to support different CIM operations. However, a naïve implementation of three SAs introduces significant peripheral circuits overhead as SAs are usually large-sized to increase the read speed. Fortunately, Jeloka *et al.* [4] already proposed a reconfigurable sense amplifier (RSA) structure that can support both single-ended sensing and differential sensing, as shown in Fig. 9. It consists of two small latch-type SAs. Therefore, the area overhead is negligible compared with a large SA in conventional memory architectures. Additional pass transistors control the behavior of the RSA in different modes. For single-ended sensing. “DIFF” = ‘0’; for differential sensing, “DIFF” = ‘1’. However, the two small latch-type SAs have inputs and outputs directly connected through PMOS. Thus, a different control signal “SAMPLE” is required to isolate the RSA’s outputs and inputs to prevent memory bit-lines (BL and BLB) from regeneration to save power. The “SAMPLE” signal is on before precharge begins and off before the RSA is enabled. Furthermore, since V_{REF} is shared by all RSAs, the switching of “SAMPLE” introduces considerable noise to V_{REF} generator through connected PMOS. Therefore, V_{REF} must be strong enough and an off-chip V_{REF} supply is generally required in this configuration.

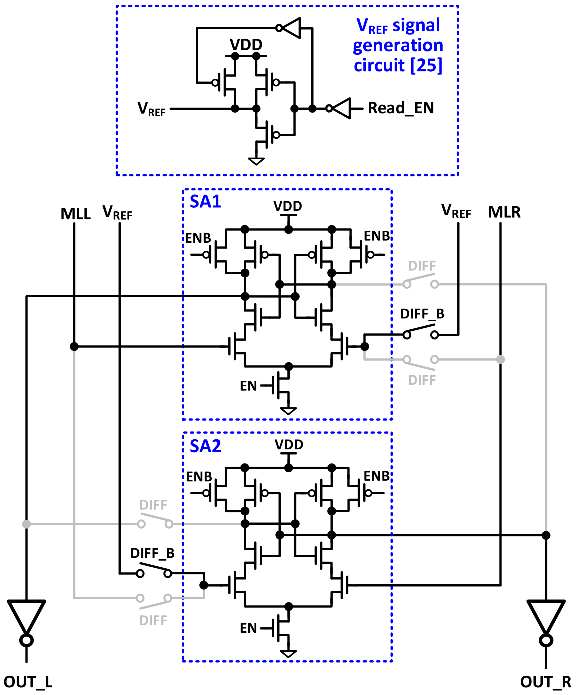


FIGURE 10. Schematic of the proposed reconfigurable sense amplifier with V_{REF} generation circuit.

Instead of the simple latch-type RSA, we propose a modified RSA as shown in Fig. 10. Two ML inputs MLL and MLR are driving high impedance (gates of NMOS transistors) through switches controlled by “DIFF” and “DIFF_B”, and full discharge of match-lines due to timing mismatch is not a concern [25]. Therefore, no additional “SAMPLE” signal is required to isolate MLL/MLR from SA outputs, simplifying the timing design. Moreover, since digital signals “DIFF” and “DIFF_B” are not changing during the sensing period, there is no concern about switching noise introduced to V_{REF} generator. Thus, an on-chip V_{REF} generator is possible. The reference generation circuit employed in this work is adapted from [26], as shown at the top of Fig. 9.

B. RERAM RELIABILITY CONSIDERATION

One big challenge associated with R-CIM systems is that the ReRAM device may incur unwanted disturb due to different biasing conditions when performing CIM operations. Take TCAM as an example, Fig. 11(a) shows the biasing conditions for half of the 4T2R bit-cell containing two NMOS (N1 and N3) and one ReRAM device (Q) when search-1 operation is performed. For Q, the biasing polarity

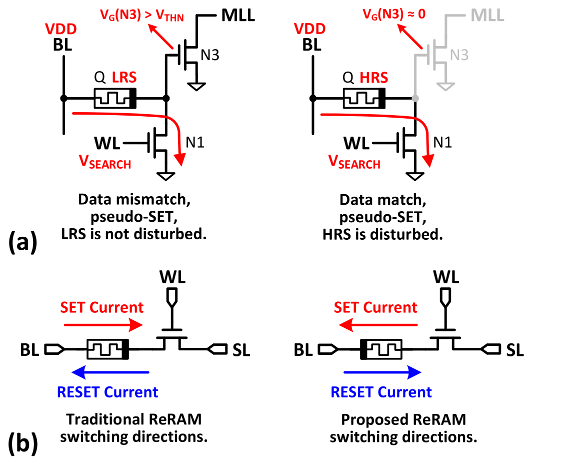


FIGURE 11. (a) Biasing conditions for half of the 4T2R bit-cell when performing search-1 operation in TCAM mode. Traditional ReRAM switching will make HRS significantly disturbed. (b) Switching directions of traditional and proposed ReRAM SET and RESET operations.

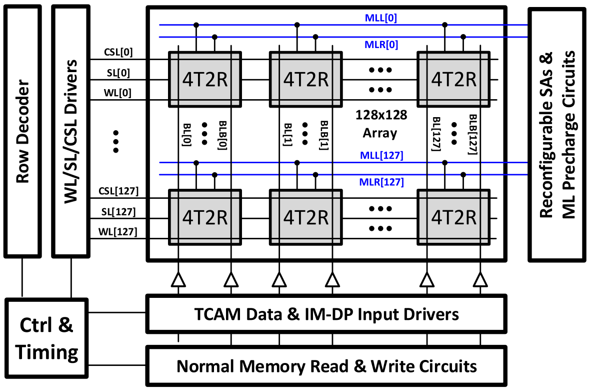


FIGURE 12. Block diagram of the 128x128 array.

across it is the same as that of a SET operation. Thus, the proposed TCAM search introduces a pseudo-SET to ReRAM devices. If search and stored data are mismatched, which means that Q is in LRS state, the voltage across Q won't cause any disturbance. However, if search and stored data are matched, which means that Q is in HRS state and the voltage across it approximately equals VDD, the pseudo-SET may cause significant disturbance to the HRS device and finally change it to LRS after enough search cycles. This will degrade the search endurance and should be carefully addressed to increase the robustness of the R-CIM system. Similarly, the HRS device in IM-DP mode will also incur disturbance when BL and BLB are biased to VDD (i.e., input = '1').

To overcome such an issue, this work proposes to reverse the directions of ReRAM SET and RESET operations, as shown in Fig. 11(b). Note that this technique incurs no overhead at circuit-level. Only magnitudes of V_{BL} and V_{SL} need to be adjusted depending on different ReRAM device characteristics. Besides, the physical layer arrangement of the ReRAM device also needs to be reversed during fabrication. Relevant work has already demonstrated the possibility of reversed SET and RESET directions for HfOx-based ReRAM devices [27].

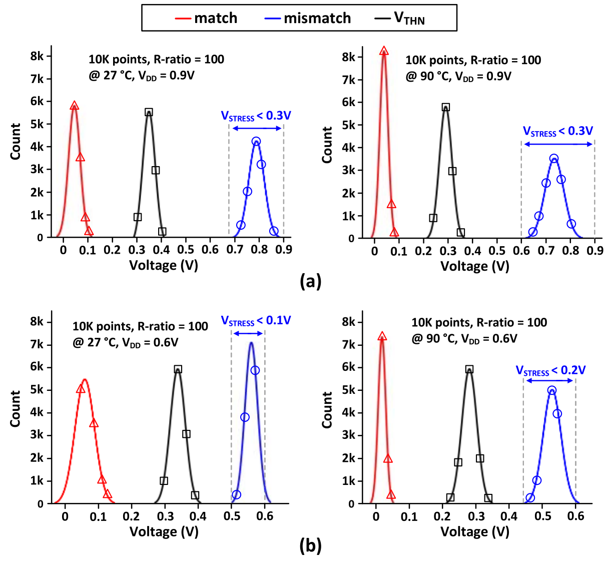


FIGURE 13. Simulated distributions of $V_{G,N3}$ and $V_{G,N4}$ in TCAM mode, at different temperatures when (a) VDD = 0.9 V and (b) VDD = 0.6 V.

By employing reversed switching directions for ReRAM devices, the biasing conditions in Fig. 11(a) become pseudo-RESET and the LRS device may be disturbed when search and stored data are mismatched. However, the biasing is designed to make the gate voltage of N3/N4 larger than the threshold voltage of NMOS during a mismatch. In fact, this gate voltage is close to VDD during a mismatch. Therefore, the voltage across the LRS device has a magnitude close to 0 so the disturbance due to pseudo-RESET is negligible.

V. SIMULATION RESULTS AND COMPARISON

In this section, we evaluate the proposed R-CIM architecture based on a 128×128 array (Fig. 12) designed in 40nm CMOS technology. The ReRAM device is modeled by Verilog-A as described in Section II-A, and the model can add a Gaussian distribution to HRS/LRS with different standard deviations to emulate ReRAM variations in Monte-Carlo simulations. The ReRAM resistance variations in this work are adopted from the HfOx ReRAM in [7] with 20% variation in LRS and 50% variation in HRS.

A. EVALUATION OF TCAM OPERATIONS

For TCAM operations, the search energy and the delay are directly affected by the match-line voltage swing during discharge period. To make a fair comparison with other ReRAM-based TCAM works [6], [8], we set the same match-line swing of 150 mV in the following analysis.

Fig. 13 shows the simulated distributions of gate voltages and threshold voltages of N3/N4 in TCAM mode based on 10K Monte-Carlo runs. Both transistor and ReRAM resistance variations are considered. The simulated conditions with room temperature and high temperature (27°C and 90°C, respectively), and nominal VDD and lower VDD

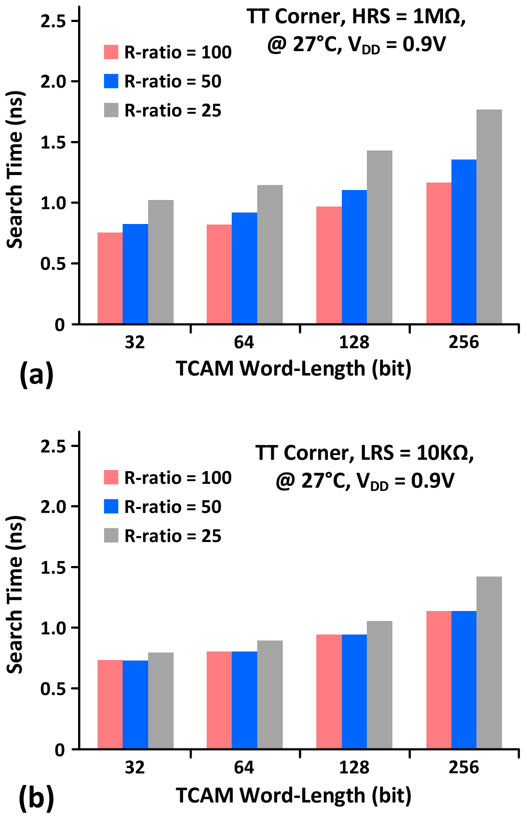


FIGURE 14. Simulated TCAM search delay versus search word-length under different R-ratios with (a) fixed HRS = 1MΩ and (b) fixed LRS = 10kΩ.

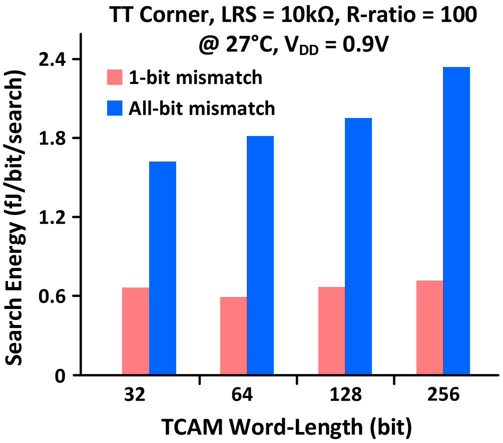


FIGURE 15. Simulated TCAM search energy versus search word-length under (a) 1-bit mismatch case and (b) all-bit mismatch case.

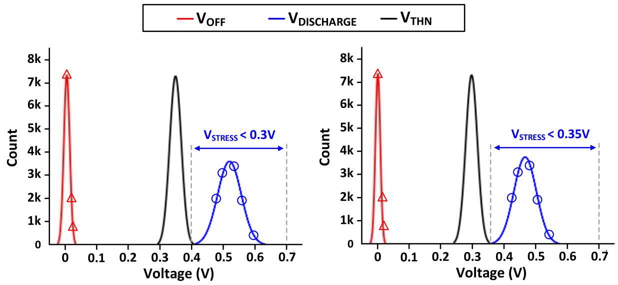


FIGURE 16. Simulated distributions of $V_{G,N3}$ and $V_{G,N4}$ in IM-DP mode at room temperature 27°C and high temperature 90°C.

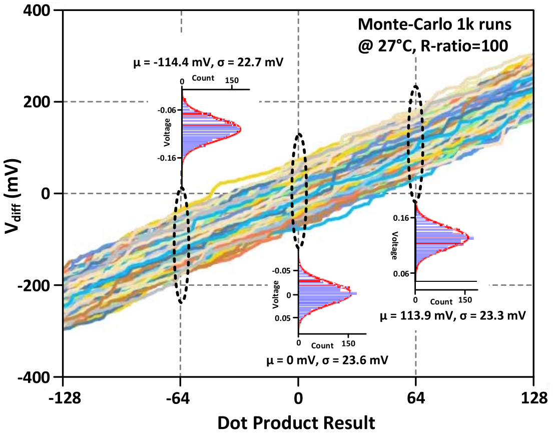


FIGURE 17. Monte-Carlo simulation (1k runs) of accumulation linearity on match-lines in IM-DP mode.

(0.9 V and 0.6 V, respectively) demonstrate a clear separation between the match case and the mismatch case, indicating the robustness of the proposed 4T2R TCAM. Note that by employing reversed SET and RESET operations for ReRAM devices, the LRS device during a mismatch case will incur unwanted disturbance as explained in Section IV-B. However, Monte-Carlo simulations show that even under the worst case, the magnitude of the stress voltage across LRS is less than 0.3 V. Such a stress voltage has negligible disturbance on LRS device since it is much smaller than the magnitude of the reset voltage which is 0.7 V as shown in Table 1.

Fig. 14 shows the simulated search delay versus TCAM word-length under different R-ratios. Fig. 14(a) is based on fixed HRS = 1MΩ and varying LRS while Fig. 14(b) is based on fixed LRS = 10kΩ and varying HRS. The proposed 4T2R TCAM can achieve 0.92 ns search delay for a 128b search word-length at LRS = 10kΩ and R-ratio = 100. A longer search word-length increases the search delay because of larger load capacitance on each MLL/MLR. It can be observed from Fig. 14(a) that LRS value has a larger impact on TCAM search performance since the discharge of MLL/MLR is controlled by LRS during the mismatch case. When LRS value becomes higher, the gate voltages of N3/N4 become lower during a mismatch due to less current in the voltage divider path. Therefore, it is desirable to have a low LRS value to achieve a high-performance search comparable to that of sTCAM [27]. On the other hand, it can be observed in Fig. 14(b) that HRS value has s small effect on TCAM search performance since the gate voltages of N3/N4 are close to '0' during a match case. However, a lower HRS value still increases the gate voltage of N3/N4 and incurs more leakage current to discharge MLL/MLR even during a whole data match. As a result, more time is required for match-lines to develop enough voltage swing to distinguish between match and mismatch cases. Thus, it is desirable to have a high HRS value to reduce leakage current.

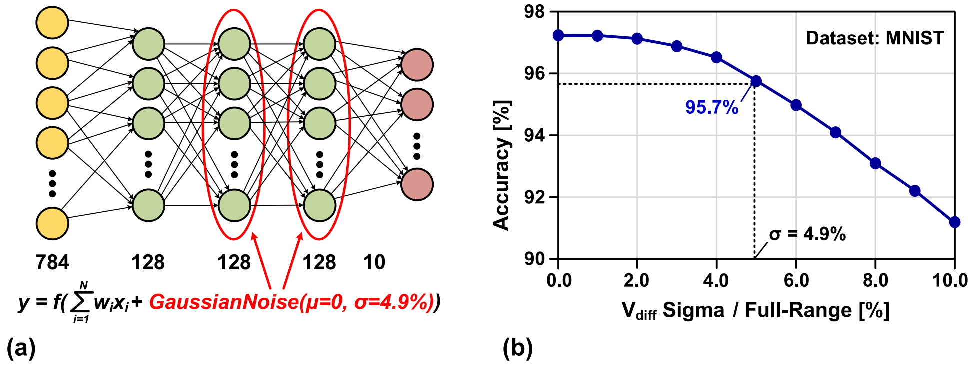


FIGURE 18. (a) Structure of the 4-layer MLP (784-128-128-128-10) for MNIST image classification; (b) simulated classification accuracy of the MNIST dataset versus standard deviation of match-line accumulation.

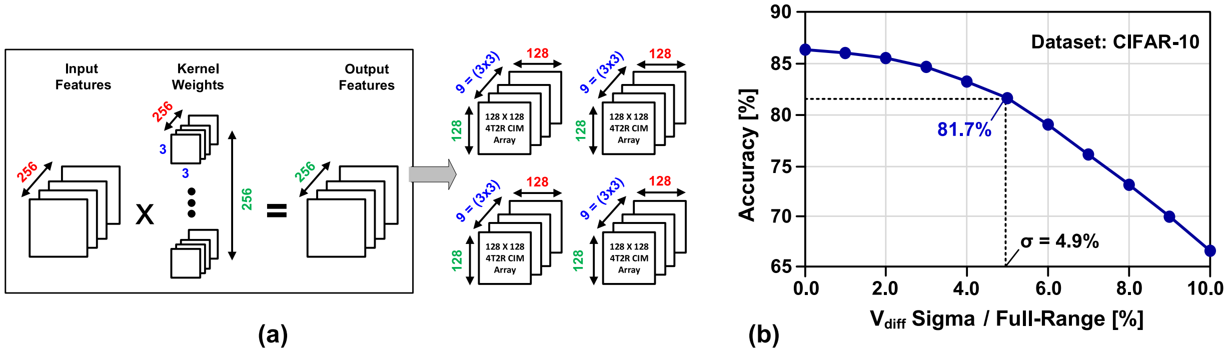


FIGURE 19. (a) Mapping of a large convolution layer (3 × 3 × 256 × 256) onto multiple R-CIM arrays; (b) simulated classification accuracy of the CIFAR-10 dataset versus standard deviation of match-line accumulation.

Fig. 15 shows the simulated search energy versus TCAM word-length under 1-bit mismatch case and all-bit mismatch case. It can be observed that the search energy of the proposed 4T2R TCAM highly depends on the data statistics given a TCAM word-length since the 4T2R bit-cell has one ReRAM device that consumes DC current during the search. The mismatched bit-cell has an LRS device in the voltage divider path and consumes more DC power than the matched bit-cell that has an HRS device in the voltage divider path. The search energy also depends on the TCAM word-length, especially when a large number of cells are mismatched. This is because a longer word-length increases the search delay and thus consumes more DC energy which becomes dominant when many bits are mismatched.

B. EVALUATION OF IM-DP OPERATIONS

Fig. 16 shows the simulated distributions of the gate voltages and the threshold voltages of N3/N4 in the IM-DP mode based on 10K Monte-Carlo runs at room temperature (27°C) and high temperature (90°C). During MLL/MLR discharging in the IM-DP mode, the gate voltage of N3/N4 is designed to be close to the transistor threshold voltage (by controlling the gate voltage of N1/N2) for improving the linearity [10], [11]. This is different from the TCAM mode where the gate voltage of N3/N4 during the mismatch case is much higher than the transistor threshold voltage. By employing reversed SET and RESET operations in the ReRAM devices, the LRS device during a MLL/MLR discharge case can incur unwanted disturbance. Worst case simulation shows that the magnitude of the stress voltage across LRS is less than 0.35 V which is far less than the reset voltage 0.7 V as shown in Table 1.

Fig. 17 shows the simulated linearity of accumulation operation on MLs considering both transistor and ReRAM resistance variations based on 1,000 Monte-Carlo runs. When the dot-product outputs are (−64, 0, and +64), the standard deviations of corresponding ML voltage levels are (22.7 mV, 23.6 mV, and 23.3 mV) while the mean values are (−114.4 mV, 0 mV, and 113.9 mV). The variations in each bit-cell are averaged out during current accumulation on MLs. The worst-case standard deviation of 23.6 mV equals to 4.9% of the mean dynamic range (480 mV, −240 mV to 240 mV) of match-line voltage in IM-DP mode. Note that DNNs inherently can tolerant some errors in computation. For example, a BITW ReRAM-based multilayer perception (MLP) in [29] achieves a 95% classification accuracy on MNIST even with a 20% standard deviation in ReRAM resistances. Thus, the amount of variation in match-line accumulation will not affect too much on the overall classification accuracy.

To further characterize the impact of variations in the linearity of match-line accumulation, we evaluated the image classification accuracy using two common datasets: MNIST and CIFAR-10. For MNIST, a 4-layer BITW MLP (784-128-128-128-10) is implemented in Keras [30]. The structure of the 4-layer MLP is shown in Fig. 18(a). The first hidden layer has floating point image pixel inputs. Therefore, we followed the approach in [9] to start IM-DP from the second hidden layer. The second and third hidden layers have a weight matrix with size 128 × 128 and can be fully mapped to the proposed 128 × 128 R-CIM array. Specifically, each row of the 128 × 128 weight matrix is mapped to a row of the R-CIM array. Therefore, a binary output can be directly obtained in each row using differential sensing as explained in Section III-D. To include the variations of match-line accumulation in Keras, we adopted the approach in [11] to add a Gaussian-Noise layer (available in Keras) with different standard deviations at the hidden layer outputs. Fig. 18(b) shows the MNIST classification accuracy versus the standard deviation of the match-line accumulation. The simulated accuracy based on a 4.9% standard deviation is 95.7%, which is 1.6% lower than the baseline with perfect linearity and no variation.

For CIFAR-10, we used a VGG-like convolutional network [22] with six convolution layers and three fully connected (FCN) layers. Due to the limited storage capacity in each R-CIM array, the network size cannot directly fit into the array. We use a weight-stationary strategy to map the network weights to multiple 128 × 128 arrays. For FCN layer mappings, weights are organized row-wise, and inputs are applied at each column. Mapping a convolution layer can be considered as mapping multiple FCN layers, e.g., mapping a 3 × 3 × 256 × 256 kernel from a convolution layer is the same as mapping nine 256 × 256 FCN layers. As shown in Fig. 19(a), channels are organized in row orientation while different kernels are assigned to different rows. Fig. 19(b) shows the CIFAR-10 classification accuracy versus the standard deviation of match-line accumulation using 3-bit partial-sums (i.e., 8 quantization levels). The simulated accuracy based on a 4.9% standard deviation in match-line accumulation is 81.7%, which is 4.6% lower than the baseline with perfect linearity and no variation.

The variation of the match-line accumulation is primarily affected by the variations in LRS since the discharge of MLL/MLR is controlled by the 1T1R voltage divider containing LRS as explained in Section III-D. Fig. 20 illustrates the impact of the LRS variation on the match-line accumulation, and the resultant MNIST classification accuracy. Larger LRS variations increase the variation in the match-line accumulation and degrades the classification accuracy. Therefore, it is desirable to keep the LRS variations as low as possible for reliable CIM operation. Many ReRAM devices have been reported to have less than 20% variation in LRS [7], [27].

Fig. 21 shows the impact of temperature on match-line accumulation in IM-DP mode, and the resultant MNIST classification accuracy. Under a wide range of temperatures, the variation on match-line accumulation changes only ∼0.5% and has a negligible impact on the classification accuracy.

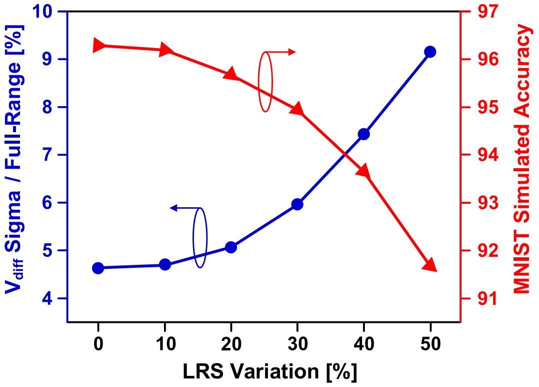


FIGURE 20. Impact of LRS variation on match-line accumulation, and the corresponding MNIST classification accuracy.

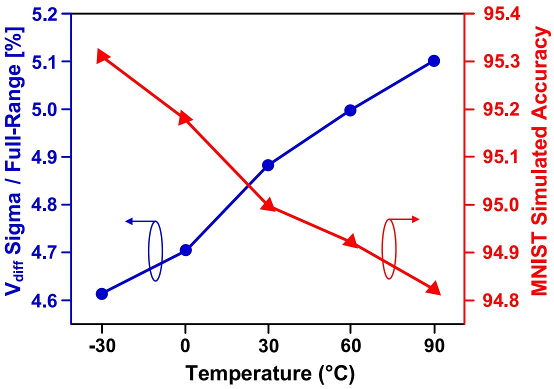


FIGURE 21. Impact of temperature on match-line accumulation, and the corresponding MNIST classification accuracy.

C. COMPARISON WITH PRIOR CIM WORKS

Table 4 compares the proposed R-CIM work and prior R-CIM works. The proposed 4T2R R-CIM structure can support versatile CIM operations in addition to NVM operations.

Compared with ReRAM-based TCAM in [6], the proposed 4T2R TCAM achieves better search performance and energy efficiency. Comparing with ReRAM-based TCAM in [8], the proposed word achieves better search performance and a slightly worse energy efficiency. However, the metrics in [8] is based on a TCAM with 64-bit word-length. If considering the same word-length, the energy efficiency of the proposed TCAM becomes 0.57 fJ/bit/search as shown in Fig. 15. Moreover, the proposed 4T2R TCAM can also operate as a conventional NVM storage system, offering a better flexibility than [6], [8]. When compared with prior ReRAM-based IM-DP for BITW networks [12], the proposed IM-DP offers better energy. This is because that [12] employs current-mode sensing for IM-DP. Therefore, the DC current through ReRAM bit-cells must be present during the entire sensing period. In contrast, the DC current only exists for a short period for MLL/MLR voltage development in the proposed IM-DP system.

TABLE 4. Comparison with prior R-CIM works.

		This Work	VLSI'19 [13]	JSSC'16 [6]	JSSC'17 [8]	ISSCC'18 [12]
Supported Functions		ReRAM access, TCAM, IM-DP	ReRAM access, TCAM, IM-DP	TCAM	TCAM	ReRAM access, IM-DP
Technology		40nm	130nm	0.18μm	90nm	65nm
ReRAM Device		HfOx (Verilog model)	HfOx	HfOx	HfOx	Unipolar CRRAM
Bit-cell Size / Area (μm^2)		4T2R / 0.55	1T1R / 7.32	4T2R / 9.7	3T1R / 1.57	1T1R / 0.25
Array Size		128 × 128	128 × 128	128 × 32	64 × 64	512 × 256
TCAM	Word Length (bit)	128	128	32	64	-
	Supply (V)	0.9	0.65 ~ 1.2	1.8	1	-
	Delay (ns)	0.92	-	1.2	0.96	-
	Energy (fJ/bit/search)	0.69 (1-bit mismatch) 1.97 (all-bit mismatch)	16.42	1.3 (1-bit mismatch) 4.36 (all-bit mismatch)	0.51 (1-bit mismatch)	-
IM-DP	Supply (V)	0.7	0.65 ~ 1.2	-	-	1
	Energy Efficiency (TOPS/W)	223.6	60.95	-	-	57.6
	Network Structure	Binary input, Ternary weight	Binary input, Binary weight	-	-	Binary input, Ternary weight

The R-CIM system in [13] supports the same operations as this work, but it has lower energy efficiency in TCAM and IM-DP modes partly due to an old technology node. However, even technology scaling is performed (assuming 3x better efficiency since the technology in [13] is about 3x of the technology in this work), The proposed R-CIM system still achieves better energy efficiency. This is because that [13] is based on an FPGA-like architecture that requires large hardware overhead for array interconnections, and complicated data mappings for different operations. For example, the NVM storage functionality in [13] requires three 128 x 128 sub-arrays to achieve the storage of one sub-array (other two arrays act as row and column decoders, respectively). On the contrary, the proposed R-CIM system introduces a novel 4T2R cell structure to perform different CIM operations without FPGA-like interconnections and complicated data mappings.

VI. CONCLUSION

In this work, we present a reconfigurable R-CIM structure using a novel 4T2R bit-cell to support NVM, TCAM and IM-DP operations. We perform optimizations in circuit and device levels to enhance the efficiency and robustness of the proposed R-CIM system. Simulation results on a 128 x 128 array show a search delay of 0.92 ns at VDD = 0.9 V which is comparable to sTCAM. The energy efficiency in IM-DP mode is 223.6 TOPS/W. Comprehensive Monte-Carlo simulations of accumulation linearity in IM-DP mode show a standard deviation of 4.9% which corresponds to a classification accuracy of 95.7% on the MNIST dataset and 81.7% on the CIFAR-10 dataset.

REFERENCES

[1] M. Horowitz, "Computing's energy problem (and what we can do about it)," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 10–14.

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

[3] H.-S. P. Wong *et al.*, "Metal–oxide RRAM," *Proc. IEEE*, vol. 100, no. 6, pp. 1951–1970, Jun. 2012.

[4] 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.

[5] A. Do, C. Yin, K. Velayudhan, Z. C. Lee, K. S. Yeo, and T. T.-H. Kim, "0.77 fJ/bit/search content addressable memory using small match line swing and automated background checking scheme for variation tolerance," *IEEE J. Solid-State Circuits*, vol. 49, no. 7, pp. 1487–1498, Jul. 2014.

[6] M.-F. Huang *et al.*, "A ReRAM-based 4T2R nonvolatile TCAM using RC-filtered stress-decoupled scheme for frequent-OFF instant-ON search engines used in IoT and big-data processing," *IEEE J. Solid-State Circuits*, vol. 51, no. 11, pp. 2786–2798, Nov. 2016.

[7] D. R. B. Ly *et al.*, "In-depth characterization of resistive memory-based ternary content addressable memories," in *Proc. IEEE IEDM*, Dec. 2018, pp. 1–4.

[8] M.-F. Chang *et al.*, "A 3T1R nonvolatile TCAM using MLC ReRAM for frequent-off instant-on filters in IoT and big-data processing," *IEEE J. Solid-State Circuits*, vol. 52, no. 6, pp. 1664–1679, Jun. 2017.

[9] S. Yin, Z. Jiang, J.-S. Seo, and M. Seok, "XNOR-SRAM: In-memory computing SRAM macro for binary/ternary deep neural networks," *IEEE J. Solid-State Circuits*, vol. 55, no. 6, pp. 1733–1743, Jun. 2020.

[10] 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.

[11] T. Yoo, H. Kim, Q. Chen, T. T.-H. Kim, and B. Kim, "A logic compatible 4T dual embedded DRAM array for in-memory computation of deep neural networks," in *Proc. IEEE/ACM ISLPED*, Jul. 2019, pp. 1–6.

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

[13] Y. Zha, E. Nowak, and J. Li, "Liquid silicon: A nonvolatile fully programmable processing-in-memory processor with monolithically integrated ReRAM," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2019, pp. 206–207.

[14] M. Bocquet *et al.*, "In-memory and error-immune differential RRAM implementation of binarized deep neural networks," in *Proc. IEEE IEDM*, Dec. 2018, pp. 484–487.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[16] H. Lee *et al.*, "Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO$_2$ based RRAM," in *Proc. IEEE IEDM*, Dec. 2008, pp. 1–4.

[17] R. Waser *et al.*, “Redox-based resistive switching memories-nanoionic mechanisms, prospects, and challenges,” *Adv. Mater.*, vol. 21, nos. 25–26, pp. 2632–2663, Jul. 2009.

[18] S. Yu, *Resistive Random Access Memory (RRAM): From Devices to Array Architectures*. San Rafael, CA, USA: Morgan & Claypool, 2016, pp. 35–38.

[19] M. Kim and P. Smaragdis, “Bitwise neural networks,” in *Proc. Int. Conf. Mach. Learn. Workshop Resource-Efficient Mach. Learn.*, 2015, pp. 1–5.

[20] R. Liu *et al.*, “Parallelizing SRAM arrays with customized bit-cell for binary neural networks,” in *Proc. ACM/IEEE Design Autom. Conf. (DAC)*, Jun. 2018, pp. 1–6.

[21] Z. Lin, M. Courbariaux, R. Memisevic, and Y. Bengio, “Neural networks with few multiplications,” 2016. [Online]. Available: https://arxiv.org/abs/1510.03009.

[22] 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.

[23] X. Si *et al.*, “A dual-split 6T SRAM-based computing-in-memory unit-macro with fully parallel product-sum operation for binarized DNN edge processors,” in *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 11, pp. 4172–4185, Nov. 2019.

[24] R. Guo *et al.*, “A 5.1 pJ/neuron 127.3μs/inference RNN-based speech recognition processor using 16 computing-in-memory SRAM macros in 65 nm CMOS,” in *IEEE Symp. VLSI Circuits Dig. Tech. Papers*, Jun. 2019, pp. 120–121.

[25] B. Mohammad, P. Dadabhoy, K. Lin, and P. Bassett, “Comparative study of current mode and voltage mode sense amplifier used for 28nm SRAM,” in *Proc. IEEE Int. Conf. Microelectron. (ICM)*, Dec. 2012, pp. 1–6.

[26] L. Lu, T. Yoo, V. L. Le, and T. T.-H. Kim, “A 0.506-pJ 16-kb 8T SRAM with vertical read wordlines and selective dual split power lines,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 28, no. 6, pp. 1345–1356, Jun. 2020.

[27] H. Lee *et al.*, “Comprehensively study of read disturb immunity and optimal read scheme for high speed HfOx based RRAM with a Ti layer,” in *Proc. IEEE Int. Symp. VLSI Technol. Syst. Appl. (VLSI-TSA)*, Apr. 2010 pp. 132–133.

[28] A. Grossi *et al.*, “Experimental investigation of 4-kb RRAM arrays programming conditions suitable for TCAM,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 26, no. 12, pp. 2599–2607, Dec. 2018.

[29] Z. Li, P.-Y. Chen, H. Xu, and S. Yu, “Design of ternary neural network with 3-D vertical RRAM array,” *IEEE Trans. Electron Devices*, vol. 64, no. 6, pp. 2721–2727, Jun. 2017.

[30] F. Chollet *et al. Keras: Deep Learning for Humans*. [Online]. Available: https://github.com/keras-team/keras

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

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

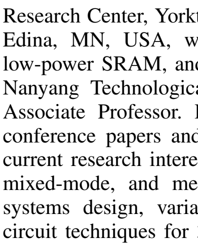
LU LU (Student Member, IEEE) received the B.E. degree from the School of Computer and Information, Hefei University of Technology in 2007, and the M.E. degree from the School of Microelectronics and Solid-State Electronics, Xiamen University, Xiamen, China, in 2010, and the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2019.

She is currently a Research Fellow with Nanyang Technological University. Her research interests include low power SRAM and SRAM based PUF. She 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.

From 2006 to 2010, 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 was as a Postdoctoral Research Fellow with Stanford University, Stanford, CA, USA, for a year. From 2006 to 2010, he was with Samsung Electronics, Yongin, South Korea, where he performed research on clock generators for high-speed serial links and clock generators. He also worked as a Research Intern with Texas Instruments, Dallas, TX, USA, IBM T. J. Watson Research, Yorktown Heights, NY, USA, and Rambus, during his Ph.D., from 2012 to 2014. After working as an assistant professor at Nanyang Technological University in Singapore for three years (from 2017 to 2020), he joins Department of Electrical and Computer Engineering (ECE) at University of California, Santa Barbara, CA, USA. His current research interests include memory-centric computing devices, circuits, and architectures, hardware accelerators, alternative computing, and mixed-signal circuit design techniques and methodologies.

Dr. Kim was a recipient of the Prestigious Doctoral Dissertation Fellowship Award based on his Ph.D. research works, the International Low Power Design Contest Award from ISLPED, and the Intel/IBM/Catalyst Foundation Award from CICC. His research works appeared at top integrated circuit design and automation conference proceedings and journals, including ISSCC, VLSI Symposium, CICC, ESSCIRC, ASSCC, ISLPED, DATE, ICCAD, JSSC, and TVLSI.



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 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 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/coauthored over 160 journal and conference papers 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*.