A 40-nm 118.44-TOPS/W Voltage-Sensing Compute-in-Memory RRAM Macro With Write Verification and Multi-Bit Encoding

Jong-Hyeok Yoon, *Member, IEEE*, Muya Chang, *Member, IEEE*, Win-San Khwa, *Member, IEEE*, Yu-Der Chih, Meng-Fan Chang, *Fellow, IEEE*, and Arijit Raychowdhury, *Fellow, IEEE*

***Abstract*—Computing-in-memory (CIM) architectures have paved the way for energy-efficient artificial intelligence (AI) systems while outperforming von Neumann architectures. In particular, resistive RAM (RRAM)-based CIM has drawn attention due to high cell density, non-volatility, and compatibility with a CMOS process. RRAM also exhibits the feasibility of high-capacity CIM with multi-bit encoding per cell exploiting an appropriate ON/OFF resistance ratio. However, the prior work regarding multi-level RRAM cells mainly focused on achieving higher bit resolution in write without consideration of CIM performance. Thus, the circuit solution to achieve multi-bit encoding per cell dedicated to RRAM-based CIM (RCIM) is of importance to support high-capacity AI systems with reliable CIM performance. This article presents a 256 × 256 CIM multi-level RRAM macro featuring iterative write with verification to achieve reliable multi-bit encoding per cell and the voltage-sensing readout circuit to surmount the underlying logic ambiguity in RCIM architectures. In addition, we also demonstrate the key design space of a fabricated RRAM array in the write operation with extensive experiments. The test chip fabricated in a Taiwan Semiconductor Manufacturing Company (TSMC) 40-nm CMOS and RRAM process achieves a peak energy efficiency of 118.44 TOPS/W in the ternary-weight multiply-and-accumulate (MAC) operation and demonstrates the feasibility of multi-level RCIM with voltage-sensing RCIM.**

***Index Terms*—Computing-in-memory (CIM), convolutional neural network, multi-level cell, multiply-and-accumulate (MAC), processing-in-memory, resistive RAM (RRAM), write verification.**

Manuscript received July 25, 2021; revised October 22, 2021 and December 29, 2021; accepted January 2, 2022. Date of publication January 21, 2022; date of current version February 24, 2022. This article was approved by Associate Editor Farhana Sheikh. This work was supported in part by the Semiconductor Research Corporation through the Center for Brain-Inspired Computing (C-BRIC) under Grant 2777.005 and 2777.006 and in part by the Applications and Systems-Driven Center for Energy-Efficient Integrated Nano Technologies (ASCENT) under Grant 2776.037. *(Corresponding author: Jong-Hyeok Yoon.)*

Jong-Hyeok Yoon was with the School of Electrical and Computing Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA. He is now with the Department of Electrical Engineering and Computer Science, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu 42988, South Korea (e-mail: jonghyeok.yoon@dgist.ac.kr).

Muya Chang and Arijit Raychowdhury are with the School of Electrical and Computing Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA.

Win-San Khwa and Meng-Fan Chang are with TSMC Corporate Research, Hsinchu 30075, Taiwan.

Yu-Der Chih is with TSMC Design Technology, Hsinchu 30075, Taiwan.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSSC.2022.3141370.

Digital Object Identifier 10.1109/JSSC.2022.3141370

I. Introduction

The advent of artificial intelligence (AI) systems and deep neural networks (DNNs) increases the demands of energy-efficient computing systems outperforming von Neumann architecture. In response to the demands, computing-in-memory (CIM) architectures have emerged. CIM architectures exploit the features of on-die memory such as the bitline (BL) structure that inherently supports the multiply-and-accumulate (MAC) operation. Compared to von Neumann architecture, the matrix–vector multiplication is conducted in memory such that massive data transfer between the processing elements and memory is avoidable. However, even in CIM architectures, data transfer between the CIM memory and the weight- and activation-storing memory occurs due to the limited capacity of CIM memory. It undermines the advantage of CIM architectures, thereby hindering the transition from von Neumann architecture to CIM architectures in practical AI systems [1]–[3]. Thus, bit density and memory capacity are of importance in CIM architectures.

The prior works regarding CIM architectures have employed emerging memory in addition to mature memory technology, such as SRAM and embedded DRAM (eDRAM). SRAM-based CIM architectures [4]–[14] have successfully demonstrated the energy-efficient CIM operation. However, the cell density of the standard 6T-SRAM is apparently low such that the complexity of AI systems is limited in the CIM architectures. Furthermore, an SRAM cell cannot contain multi-bit weights, thereby precluding the multi-bit CIM operation. As a solution to multi-bit CIM architectures, 8T-SRAM has drawn attention by employing a 2T-read path that represents the binary weight (i.e., 2^k, $k \in Z_0^+$) per cell [13]. However, the 8T-SRAM exacerbates the low cell density and cannot still achieve multi-bit encoding per cell. The eDRAM-based CIM architecture has recently been proposed as a solution to multi-level cells for the CIM operation [15]. With a 2T gain cell and an additional transistor for the input pulse where the pulsewidth (PW) represents the input value, the memory cell successfully supports the multi-bit CIM operation with multi-bit encoding per cell. However, notwithstanding the multi-bit encoding, the cell size also increases such that it neutralizes the advantage of multi-level cells in the view of bit density to an extent.

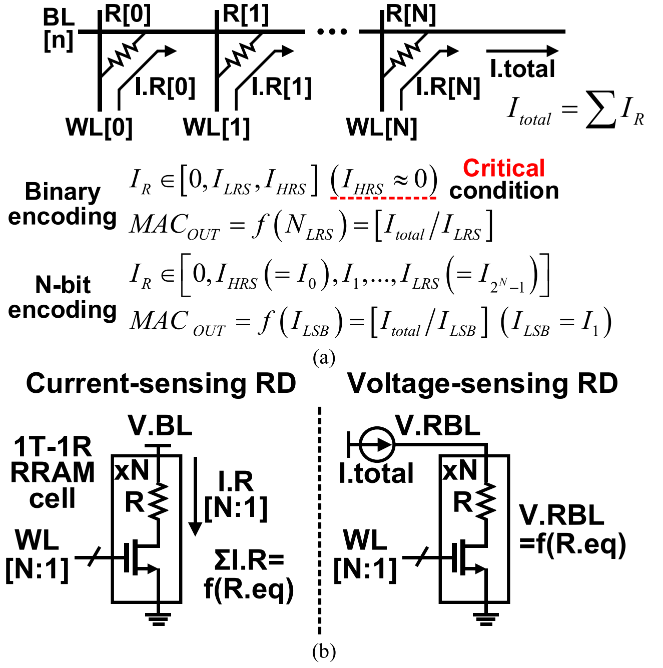


Fig. 1. (a) Current-sensing RCIM at the BL. (b) Simplified structure of current- and voltage-sensing read in RCIM architectures.

Considering superior cell density as well as non-volatility, emerging memory has been in the spotlight. Compared to conventional memory, emerging memory sheds light on the feasibility of high-capacity CIM architectures in practical AI systems, especially for edge devices [16]–[25]. Among emerging memory, resistive RAM (RRAM) accommodates multi-bit encoding per cell exploiting an appropriate ON/OFF resistance ratio. However, there are some obstacles in exploiting multi-level cells in RRAM-based CIM (RCIM) architectures. Fig. 1 shows the current-sensing RCIM at the BL and the simplified structure of current- and voltage-sensing read (RD) in RCIM architectures. In a binary RRAM array, RRAM cells are programmed in a low-resistance state (LRS) or a high-resistance state (HRS) to represent the data such as the weights of DNNs. Since a fixed BL voltage is used in the current-sensing RCIM, the cell current is directly affected by the cell resistance [Fig. 1(b)]. In the current-sensing RCIM, the MAC output is estimated by the ratio of the total cell current to the LRS current under the presumption that the current of HRS cells is negligible. In the case of concurrent accesses to multiple HRS cells, the total HRS current at the BL eventually exceeds the LRS current, thereby incurring logic ambiguity. Even if the ON/OFF ratio is sufficiently high so that the HRS current is virtually negligible, the ratio of the LSB current to the HRS current drastically deteriorates over increasing the bit resolution in multi-level cells. It eventually limits the maximum bit resolution per cell in RCIM architectures due to logic ambiguity.

In order to surmount the aforementioned problems, a voltage-sensing RCIM architecture has piqued our interest. A fixed-current voltage-sensing RCIM architecture suffers from the severely nonlinear readout BL voltage (V.RBL) that is inversely proportional to the parallel resistance of accessed RRAM cells [Fig. 1(b)]. Thus, the prior works tried to mitigate the nonlinearity by using the variable current source despite the remaining nonlinearity [23], [24]. Recently, the voltage-sensing RCIM architecture demonstrated the linear V.RBL with a binary RRAM array [25]. To obtain the high-capacity RCIM architectures without logic ambiguity, a voltage-sensing RCIM architecture with multi-level cells is necessary. In addition, the multi-level cell resistance considering the voltage-sensing RCIM architecture should also be addressed. Due to the appropriate ON/OFF ratio of an RRAM array, the multi-level RRAM cell has been addressed [26]–[32]. The prior works mitigated the variation of RRAM cells such as a different sensitivity to a write (WR) pulse, thereby achieving a tight distribution of cell resistance. However, the prior works have more focused on the device characteristics and the resistance distribution of RRAM cells. It leads to a lack of consideration for the placement of the multi-level resistance optimized for the voltage-sensing RCIM architecture. Thus, the joint optimization considering the device characteristics and the voltage-sensing RCIM should be addressed in a high-capacity RCIM architecture with multi-level cells.

In this article, a voltage-sensing multi-level RCIM architecture [33] is proposed to support multi-bit CIM operation while achieving the joint optimization for the cell characteristics and the RCIM architecture with multi-level cells. The proposed RRAM macro features: 1) the iterative WR with verification (IWR) to achieve reliable multi-level cells and 2) multi-bit voltage-sensing RCIM architectures surmounting the logic ambiguity in the current-sensing RCIM architecture that is much severe with multi-level cells. *In situ* IWR achieves a tight resistance distribution of multi-level cells with two thresholds for target resistance while adjusting the WR pulse amplitude. An intermediate-resistance state (IRS) is determined to achieve the linear V.RBL in the voltage-sensing RCIM architecture with multi-level cells. The voltage-sensing RCIM incorporates the input-aware (IA) BL current control with an active feedback amplifier [25] to linearize the V.RBL, thereby attaining reliable CIM operation with multi-level cells. Compared to the prior work regarding binary RCIM [25], the test chip is reconfigured and features a multi-level RCIM architecture performing reliable MAC operation with multi-level cells for AI systems with an energy efficiency of 118.44 TOPS/W. To the best of the authors' knowledge, this work is the first RCIM architecture with multi-level cells fabricated in the standard monolithic RRAM and CMOS process. In addition, we provide the measured data of the resistance distribution over the WR operation and the key design space of various inter-dependent WR parameters, such as pulse configuration, target resistance, and WL/BL voltages (V.WL/V.BL), thereby helping develop statistical models of RRAM and the corresponding design techniques.

The rest of this article is organized as follows. Section II describes the architecture of the proposed multi-level RRAM macro. Section III discusses the detailed implementation of the voltage-sensing RD for multi-level RCIM. Section IV delineates the IWR in the proposed RRAM macro. Section V describes the measured device characteristics of the fabricated

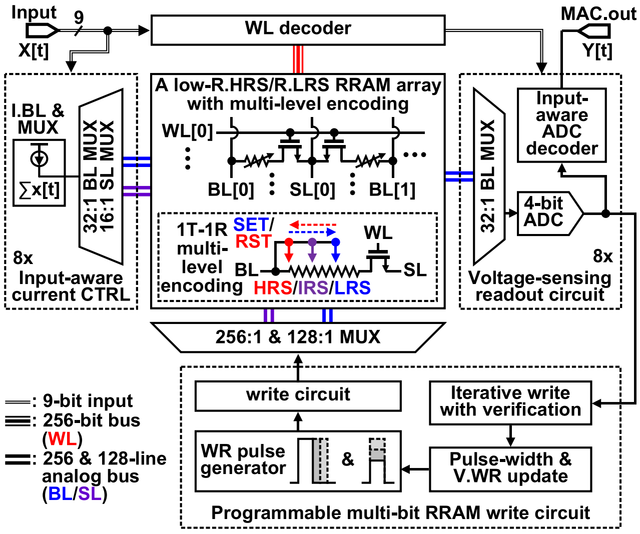


Fig. 2. Top block diagram of the proposed CIM multi-level RRAM macro.

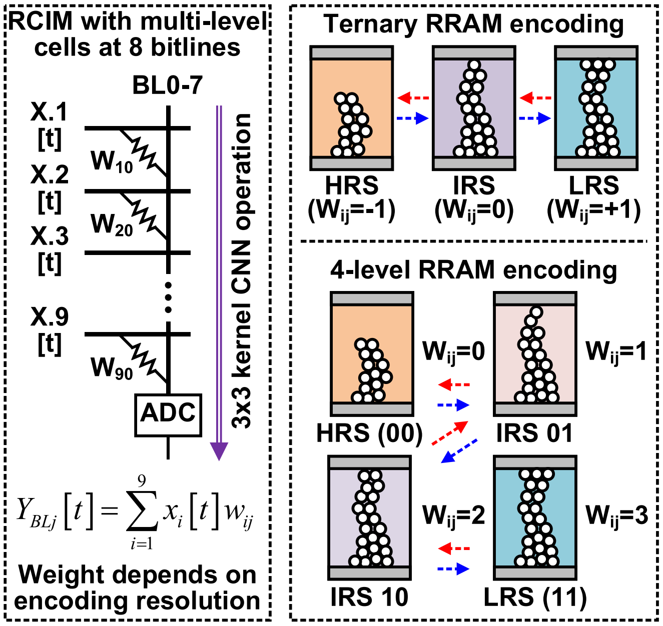


Fig. 3. Ternary-weight CIM operation and the ternary encoding of the proposed RRAM macro.

RRAM array. Section VI presents the measurement results of the proposed RCIM architectures. Section VII presents the conclusions drawn from this study.

II. Proposed Multi-Level RRAM Macro

As the solution to high-capacity RCIM architectures, the proposed multi-level RRAM macro supports RCIM with multi-level cells. The voltage-sensing RCIM has been proposed while demonstrating the linearized V.RBL over the combinations of accessed binary RRAM resistance [25]. This work expands the application of the voltage-sensing RCIM to that with multi-level cells, thereby enabling the high-capacity RCIM architecture. Regarding multi-bit encoding per cell, RRAM has technical challenges, such as the tradeoff between the programmability and the encoding margin [34]. A high ON/OFF ratio is preferred to obtain multi-level cells since it provides a sufficient encoding margin. However, a back-to-back WR with a high ON/OFF ratio leads to the drastic formation and rupture of conductive filaments in an RRAM cell. It gradually degrades the programmability of RRAM cells. Thus, a circuit solution tightening the resistance distribution of RRAM cells under an appropriate ON/OFF ratio is necessary for reliable RCIM architectures with multi-level cells.

Fig. 2 shows the top block diagram of the proposed RRAM macro utilizing multi-level RRAM cells. The proposed RRAM macro consists of a 256 × 256 multi-level 1T-1R RRAM array (101.4 kb in ternary encoding), the IA BL current control with a feedback amplifier, a 4-b flash analog-to-digital converter (ADC) with an IA ADC decoder, and the IWR. The proposed RRAM macro supports eight-BL RD accessing up to nine WLs simultaneously to render 3 × 3 convolutions. The 3 × 3 filter size in the proposed RRAM macro is determined to support the scalability of the AI systems. A 3 × 3 filter is a primary filter in convolutional layers where the odd-sized filter is preferred considering symmetry at the output. Multiple convolutional layers of 3 × 3 filters are equivalent to a single layer of larger odd-sized filters so that versatile CNNs, such as MobileNet, have utilized 3 × 3 filters. The multi-level CIM operation and the ternary and unsigned four-level encoding per cell of the proposed RRAM macro are shown in Fig. 3.

In the CIM operation, the binary input is fed to the WL decoder to access nine WLs. Then, the designated RRAM cells are selected via the eight-BL/four-sourceline (SL) MUX for simultaneous CIM operation. Due to the two-BL/one-SL RRAM array, the SL MUX has half the size of the BL MUX, thereby attaining area efficiency in the RRAM access. The accessed RRAM cells are connected with the BL such that the V.RBL represents the CIM output. The IA BL current control is used to provide the current proportional to the number of accessed RRAM cells (N.RRAM), thereby mitigating the drastic decrease of the V.RBL over the parallel resistances of accessed RRAM cells. However, the remaining nonlinearity over the combinations of resistance states, including the IRS, exacerbates a narrow sampling margin at the ADC in the readout circuits. Thus, active feedback control at BLs is employed to control the current source, thereby linearizing the sampling levels in the proposed macro. The linearized V.RBL is applied to the 4-b ADC. The ADC threshold is uniformly distributed over the range of the V.RBL. The IA ADC decoder readouts the CIM output with the logic thresholds considering the N.RRAM.

To program multi-levels in an RRAM array, the IWR is employed in the proposed RRAM macro. The challenges in RRAM technology, such as reliability, necessitate an iterative WR process called write-verify. The prior works regarding write-verify successfully achieved a tightened distribution of multi-level cell resistance. However, the prior works focused on the feasibility of multi-level cells itself, not the optimization for the RCIM performance. On the contrary, the IWR in the proposed RRAM macro conducts multi-level encoding per cell while achieving the joint optimization for the resistance distribution and the voltage-sensing RCIM architecture. In particular, since multi-level RRAM cells are accessed, the number of cases in the combination of cell resistances is much higher than binary RRAM cells. Thus, the IRS resistance (R.IRS) is set to maintain the linear V.RBL over the various combinations of accessed cell resistances. Due to the relation between the linearized V.RBL and the cell resistance, the cell resistance can be indirectly measured by the V.RBL in single-cell access. Thus, the IWR estimates whether the cell resistance is placed within the target range by using the 4-b ADC. In case the resistance is out of the target range, another WR iteration is initiated while adjusting the WR pulse amplitude and width. It eventually achieves a tightened distribution of cell resistances in multi-bit encoding per cell.

It is noteworthy that the voltage-sensing RCIM architecture outperforms the current-sensing RCIM suffering from the logic ambiguity problem that is exacerbated over multi-bit encoding per cell. Besides, the proposed voltage-sensing RRAM macro provides the linearized V.RBL that is essential to achieve reliable multi-bit CIM. The prior work achieves the CIM operation with the nonlinear ADC to read out nonlinear V.RBLs [23], [24]. The nonlinear voltage-sensing RCIM can support CIM with a low N.RRAM. However, the density of the ADC thresholds exponentially increases over higher N.RRAMs such that the ADC cannot read out the CIM results appropriately due to the sensitivity of the ADC comparators. It even deteriorates in employing multi-level cells. On the contrary, in the case of the linearized V.RBL, the sampling margin of the V.RBL gradually decreases over increasing the N.RRAM and the encoding bits. Thus, the maximum N.RRAM where the spacing of ADC thresholds exceeds the sensitivity of the ADC comparators is higher than the nonlinear voltage-sensing RCIM architecture.

The test chip of the proposed multi-level RRAM macro supports only positive inputs. However, considering a ReLU activation function where the output is always positive, this is not a hindrance to implement AI systems with the proposed RRAM macro [35]. Furthermore, negative inputs can be easily supported by using two RRAM arrays as the prior works supporting both positive and negative inputs.

III. Voltage-Sensing Read in Multi-Level RCIM

A high-resolution readout is of importance in RCIM with multi-level cells due to the increasing number of combinations of input–weight pairs. As a solution to the logic ambiguity problem incurred in current-sensing RCIM, the linearized voltage-sensing RCIM has been demonstrated with a binary RRAM array [25]. The proposed multi-level RRAM macro exploits the advantage of the voltage-sensing RD featuring the IA BL current control with a feedback amplifier and the following ADC-based readout circuits while expanding the application to multi-bit RRAM arrays.

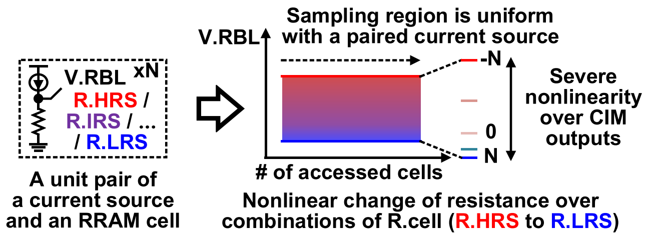


Fig. 4. Remaining nonlinearity of the readout BL voltage with the IA BL current control.

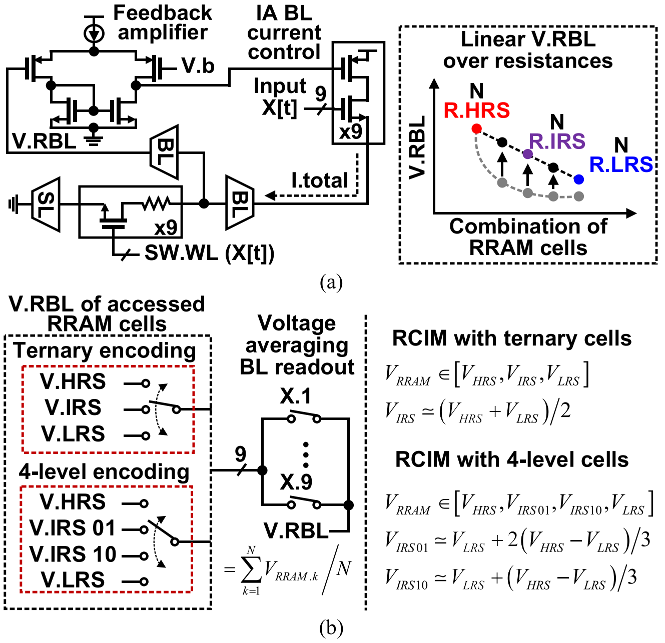
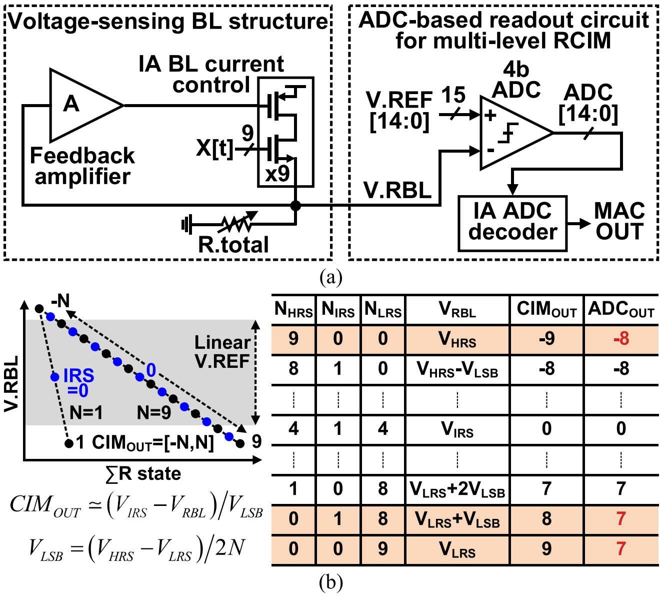


Fig. 5. (a) Structure of the proposed voltage-sensing BL in multi-level RCIM and (b) voltage averaging BL readout model.

A. IA BL Current Control With a Feedback Amplifier for Multi-Level RCIM

To read out the CIM outputs from the V.RBL with multi-level cells appropriately, the factors to introduce the nonlinearity to the V.RBL should be addressed. The V.RBL is directly affected by the total BL current and the parallel resistance of accessed RRAM cells. Since the parallel resistance drastically decreases over increasing the N.RRAM, the IA BL current control provides the BL current proportional to the N.RRAM to neutralize the nonlinearity due to the N.RRAM. Fig. 4 shows the remaining nonlinearity over the combinations of the accessed cell resistances with the IA BL current control. Since the parallel resistance is based on the harmonic mean of accessed multi-level cells, the current proportional to the N.RRAM cannot fully compensate for the nonlinearity introduced by the parallel resistance. Thus, the remaining nonlinearity due to the combination of accessed multi-level cell resistances is suppressed by employing an active feedback amplifier.

Fig. 5 shows the voltage-sensing BL structure of the proposed multi-level RRAM macro and the simplified model for the BL readout. Since the IA BL current control eliminates



$$CIM_{OUT} \simeq (V_{IRS} - V_{RBL})/V_{LSB}$$

$$V_{LSB} = (V_{HRS} - V_{LRS})/2N$$

N_{HRS}	N_{IRS}	N_{LRS}	V_{RBL}	CIM_{OUT}	ADC_{OUT}
9	0	0	V_{HRS}	-9	-8
8	1	0	V_{HRS}-V_{LSB}	-8	-8
⋮	⋮	⋮	⋮	⋮	⋮
4	1	4	V_{IRS}	0	0
⋮	⋮	⋮	⋮	⋮	⋮
1	0	8	V_{LRS}+2V_{LSB}	7	7
0	1	8	V_{LRS}+V_{LSB}	8	7
0	0	9	V_{LRS}	9	7

Fig. 6. (a) Block diagram of the ADC-based readout circuits for the multi-level RCIM. (b) Uniform distribution of the reference voltages and the logic saturation under non-sparse ternary-weight CIM (N.RRAM = 9).

the nonlinearity due to the N.RRAM and the feedback amplifier linearizes the V.RBL over the combinations of the accessed multi-level cell resistances, the proposed BL structure can be modeled as a voltage-averaging circuit with the V.RBL in the case of a single HRS/IRS/LRS cell access (V.HRS/V.IRS/V.LRS). Thus, the V.RBL represents the normalized CIM output over the N.RRAM and is fed to the ADC-based readout circuits. The comprehensive analysis regarding the effectiveness of the IA BL current control and feedback amplifier has been conducted in the prior work [25].

It is worth noting that the proposed voltage-sensing RD can readily support larger filters in CNNs. In the case of the current-sensing RD, the logic ambiguity due to the HRS current limits the maximum N.RRAM, thereby limiting the scalability of RCIM. On the contrary, the voltage-sensing RD can increase the N.RRAM. In the case of a larger filter size, the overhead is only the number of the unit current source in the IA BL current control that is set to the filter size.

B. ADC-Based Readout Circuits

The ADC-based readout circuits read out the CIM output from the linearized V.RBL. Fig. 6 shows the block diagram of the ADC-based readout circuits, the distribution of the ADC references (V.REFs), and the cases of logic saturation with non-sparse inputs and weights in the proposed RRAM macro. The V.RBL represents the normalized CIM output that has a constant range from V.HRS to V.LRS where the CIM output is from $-N$.RRAM to $+N$.RRAM. Thus, the readout circuits should consider the N.RRAM in addition to the V.RBL to read out the CIM output. The linearized V.RBL is scanned by a 4-bit flash ADC. The reference voltages are linearly distributed and the IA ADC decoder determines the CIM output with the multi-level RRAM macro considering the ADC output and the N.RRAM. In case that the N.RRAM > 7 and all the accessed RRAM cells are in the HRS (−1) or LRS (+1), the V.RBL exceeds the dynamic range of the 4-b ADC. Thus, the CIM output is eventually saturated in the IA ADC decoder. Considering the sparsity of inputs and weights in CNNs, the probability that it occurs is sufficiently low and the saturation has a negligible impact on the performance of CNNs. Thus, we use a 4-b ADC instead of a 5-b ADC to achieve energy efficiency without loss of accuracy. The detailed schematics of the ADC-based readout circuits are shown in the prior work [25].

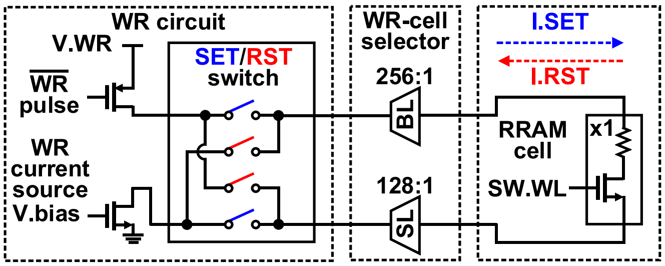


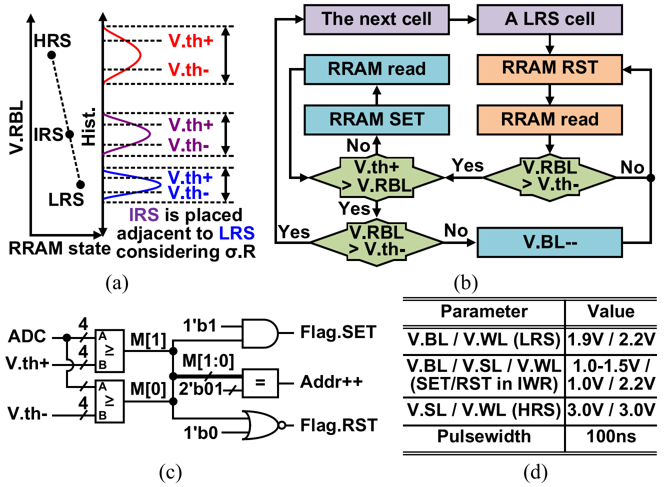
Fig. 7. Schematics of the write circuit.

IV. Iterative Write With Verification

RRAM has technical challenges such as the reliability of cell resistance. RRAM does not have a complete set or reset state since the conductive filaments in an RRAM cell cannot be fully formed and ruptured. In addition, RRAM cells have different sensitivities to a WR pulse. Thus, RRAM suffers from a wide distribution of cell resistance over the WR operation. It eventually leads to erroneous CIM operation in the readout.

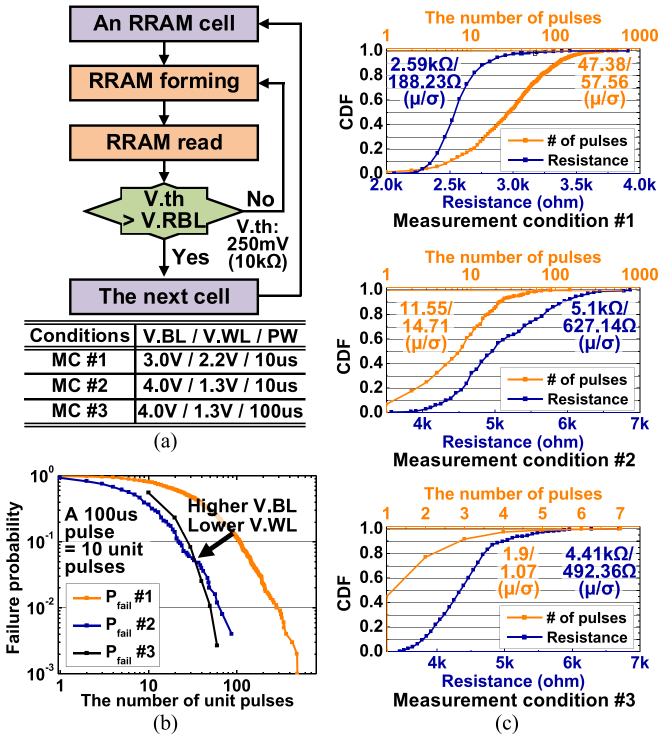
To tighten the distribution of the cell resistance, prior works successfully conducted write-verify that is a WR process with iterations. While exploiting write-verify, the prior works also demonstrated multi-level RRAM cells with arbitrary resistances [26]–[32]. It could shed light on the feasibility of high-capacity RRAM. However, RCIM architectures require not only the tightened resistance distribution but also the cell resistance optimized for the reliable CIM. Thus, we address these challenges by employing the IWR with multi-bit encoding per cell optimized for the voltage-sensing RCIM.

Fig. 7 shows the WR circuit of the proposed RRAM macro. The WR circuit supports the cell-by-cell WR operation. The WR MUX selects an RRAM cell to be programmed. By using the set/reset selector, the direction of the WR current is controlled to set/reset the selected RRAM cell. Then, the WR pulse is applied to the WR circuit, thereby conducting the WR operation. Fig. 8 shows the placement of the cell resistances in ternary encoding, the flowchart of the IWR, the schematics of the resistance verification, and the pulse configuration used in the IWR. To tighten the resistance distribution and enable multi-bit encoding per cell under an appropriate ON/OFF ratio, the IWR is employed in the proposed RRAM macro. The IWR consists of a WR-pulse injection and resistance verification. The WR PW is 100 ns with the pulse configuration (V.BL/V.SL/V.WL) designated for the LRS/IRS/HRS encoding. After every WR pulse, the



Parameter	Value
V.BL / V.WL (LRS)	1.9V / 2.2V
V.BL / V.SL / V.WL (SET/RST in IWR)	1.0-1.5V / 1.0V / 2.2V
V.SL / V.WL (HRS)	3.0V / 3.0V
Pulsewidth	100ns

Fig. 8. (a) Placement of the cell resistance in ternary encoding and its threshold in the iterative write with verification, (b) flowchart in programming the IRS, (c) schematics of the resistance verification, and (d) pulse configuration used in the iterative write with verification.



Conditions	V.BL / V.WL / PW
MC #1	3.0V / 2.2V / 10us
MC #2	4.0V / 1.3V / 10us
MC #3	4.0V / 1.3V / 100us

Fig. 9. (a) Forming process and the measurement conditions, (b) measured failure probability in the forming process, and (c) measured resistance and the number of forming pulses over various pulse configurations.

readout circuit detects whether the resistance of an RRAM cell reaches the target resistance by estimating the resistance based on the V.RBL. The upper and lower thresholds for the target resistance are set to 4-b digital signals that are compared with the output of the 4-b ADC in the resistance verification.

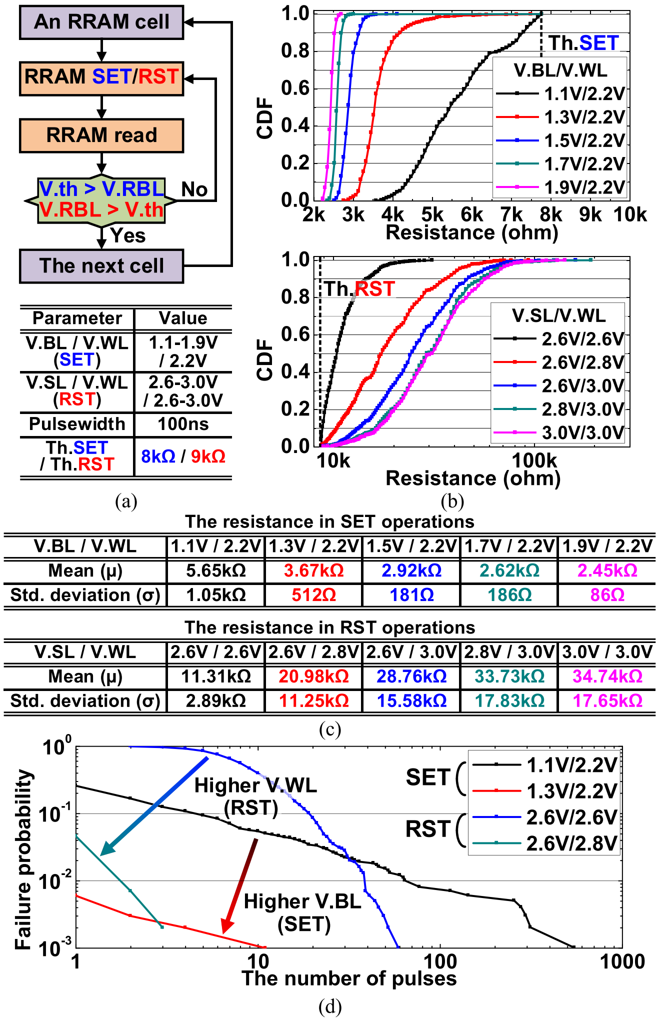
In the IRS encoding, the WR operation is started from the LRS cell for the narrow distribution of initial resistances. A reset pulse is applied until the RRAM resistance exceeds the lower threshold of the target R.IRS. If the resistance is over the upper threshold, a set pulse is applied to the RRAM cell. In case the RRAM resistance fluctuates over the thresholds, another iteration is initiated with a lower BL voltage until the target resistance is achieved. For the LRS/HRS encoding, a single threshold with a fixed pulse configuration can be employed. Since an LRS/HRS cell has the inherent lower/upper limit of resistance due to the device characteristics, an upper/lower threshold for the LRS/HRS in the IWR is sufficient to achieve a desirable distribution of the cell resistance. Besides, the optimized WR pulse amplitude for the LRS/HRS is exploited without adjusting the WR voltages, thereby reducing the complexity of WR process. Compared to the narrow distribution of the LRS resistance (R.LRS), the HRS resistance (R.HRS) can have a wide upper distribution after the WR process with a single threshold. However, due to the tolerance to the distribution of the R.HRS in the proposed voltage-sensing RD, the reliable RCIM operation is achieved with a single threshold in the WR process.

In the multi-level encoding, the R.IRS is placed adjacent to the R.LRS since the resistance distribution of RRAM cells is narrow near the LRS regime. In addition, the linearized V.RBL is also considered in determining the R.IRS. Without consideration of the voltage-sensing RD, arbitrary encoding only considering the space of encoding resistances introduces a severe nonlinearity to the V.RBL, thereby hindering the reliable RCIM operation. Thus, the proposed IWR achieves the multi-level cells with the consideration of the RCIM performance.

V. Characteristics of the Fabricated RRAM Array

While utilizing the proposed voltage-sensing RRAM macro with the IWR, we extensively characterize the resistance distribution on the fabricated RRAM macro and demonstrate key post-silicon inter-dependent WR parameters, such as PW, target resistance, voltages, and the number of pulses. In addition to obtaining the optimized WR configuration for the proposed RRAM macro, these results will act as foundations to further develop statistical models of filamentary memory devices as well as enable design techniques and characterization methodologies of RRAM array macros.

The forming process and the measured results are shown in Fig. 9. The forming pulse is applied until the V.RBL is below the threshold voltage (V.th). Since the forming resistance lies in the LRS regime, V.th is set to 250 mV where the corresponding resistance is 10 kΩ. In the measurement condition (MC) 1 using 3 V of the V.BL and 2.2 V of the V.WL, the average formed resistance is 2.59 kΩ. The average number of pulses required for the MC1 is 47.38. In the MC2 and MC3 using 4 V of the V.BL and 1.3 V of the V.WL where the PW is 10 and 100 μ s, respectively, the forming resistance increases to 4.5–5 kΩ. However, the number of required unit pulses is reduced to 24%–40%. In the forming process, we observe empirically that the V.BL affects the number of pulses required in forming the RRAM cells and the V.WL determines the formed resistance. The failure probability (P.fail) of the MC1-3 is shown in Fig. 9. Here, P.fail does not represent cells where forming is not possible,



Parameter	Value
V.BL / V.WL (SET)	1.1-1.9V / 2.2V
V.SL / V.WL (RST)	2.6-3.0V / 2.6-3.0V
Pulsewidth	100ns
Th.SET / Th.RST	8kΩ / 9kΩ

(a)

(b)

The resistance in SET operations

V.BL / V.WL	1.1V / 2.2V	1.3V / 2.2V	1.5V / 2.2V	1.7V / 2.2V	1.9V / 2.2V
Mean (μ)	5.65kΩ	3.67kΩ	2.92kΩ	2.62kΩ	2.45kΩ
Std. deviation (σ)	1.05kΩ	512Ω	181Ω	186Ω	86Ω

The resistance in RST operations

V.SL / V.WL	2.6V / 2.6V	2.6V / 2.8V	2.6V / 3.0V	2.8V / 3.0V	3.0V / 3.0V
Mean (μ)	11.31kΩ	20.98kΩ	28.76kΩ	33.73kΩ	34.74kΩ
Std. deviation (σ)	2.89kΩ	11.25kΩ	15.58kΩ	17.83kΩ	17.65kΩ

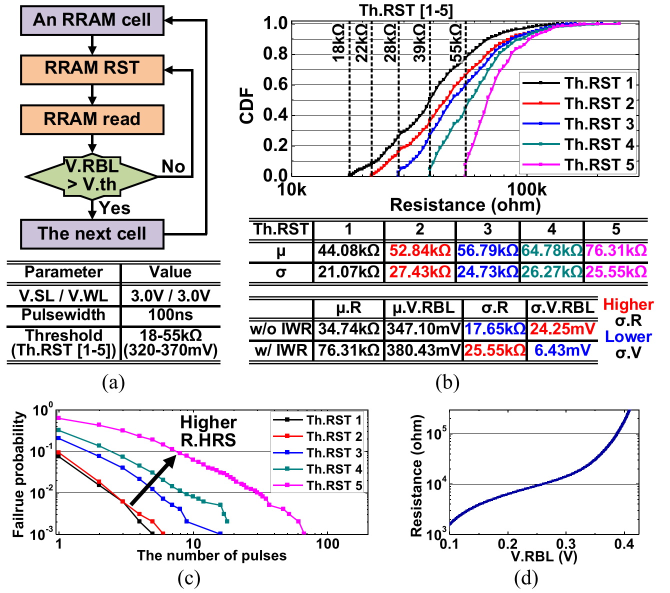
(c)

(d)

Fig. 10. (a) Write operations and the measurement setups, (b) measured resistance in the set/reset operation over various pulse configurations, (c) measured statistics of measured resistances, and (d) measured failure probability in the write operation.

but rather the cells that require additional pulses to complete the forming process. We note that the P.fail of the MC2 and MC3 exhibits near-identical profiles and we conclude that the forming process is affected by the total forming time, not a unit PW.

In Fig. 10, the WR operation and the measured results are shown. In the set operation, 1.1–1.9 V of the V.BL s are used with 2.2 V of the V.WL. In the reset operation, 2.6–3.0 V of the V.WLs and V.SLs are used. A different V.th is used during the set/reset operation to secure a target readout margin between the R.LRS and R.HRS. During WR, the pulse is applied until the RRAM cell has changed to the target state. During set, the average R.LRS decreases to 2.45 kΩ and achieves a lower standard deviation of the resistance (σ. R) of 1.05 kΩ. The R.LRS does not have the tail distribution close to the threshold when V.BL = 1.9 V. During reset, the average R.HRS increases to 34.74 kΩ for higher V.WLs and V.SLs. However, the tail distribution of R.HRS occurs even for higher voltages. The reset of RRAM cells with the V.WL < 2.4 V cannot be completed even with higher V.SLs considering the body effect at the NMOS switch in the 1T-1R structure. We observe that during WR, the V.WL is a critical condition for reset.



Parameter	Value
V.SL / V.WL	3.0V / 3.0V
Pulsewidth	100ns
Threshold (Th.RST [1-5])	18-55kΩ (320-370mV)

(a)

Th.RST	1	2	3	4	5
μ	44.08kΩ	52.84kΩ	56.79kΩ	64.78kΩ	76.31kΩ
σ	21.07kΩ	27.43kΩ	24.73kΩ	26.27kΩ	25.55kΩ

	μ.R	μ.V.RBL	σ.R	σ.V.RBL
w/o IWR	34.74kΩ	347.10mV	17.65kΩ	24.25mV
w/ IWR	76.31kΩ	380.43mV	25.55kΩ	6.43mV

(b)

(c)

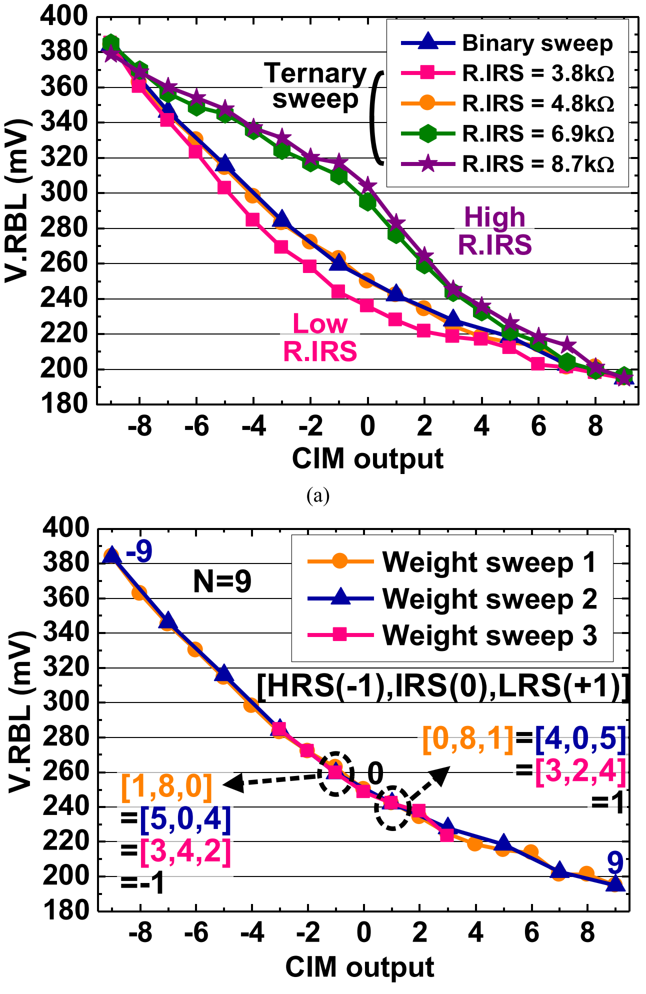
(d)

Fig. 11. (a) Iterative write with verification and the measurement setup; (b) measured resistance over various reset thresholds, statistics of measured resistance, and the comparison of the distribution of resistance and readout voltage with and without iterative write with verification; (c) measured failure probability over various reset thresholds; and (d) simulated resistance-readout voltage characteristics in the proposed RRAM macro.

To shorten the tail of R.HRS, various reset thresholds are employed with the optimized reset pulse configuration (V.SL = V.WL = 3.0 V). The measurement flowchart and measured results of the IWR are shown in Fig. 11. The reset thresholds are set to remove the tail of the R.HRS distribution. With the IWR, the average R.HRS increases to 76.31 kΩ and σ.R is ~25 kΩ, which is higher than σ.R without the IWR. However, the standard deviation of the V.RBL is reduced to 26.5% due to the insensitivity to resistance changes in the HRS regime [Fig. 11(d)] such that the IWR achieves a higher margin in a voltage-based readout or MAC logic. Furthermore, the insensitivity helps attain the tolerance to random telegraph noise in HRS cells [36]. As expected, P.fail increases for higher thresholds.

VI. Measurement Results

The proposed multi-level RRAM macro is fabricated in a standard monolithic 40-nm CMOS and RRAM process exploiting multi-level cells in RCIM architectures. The test chip demonstrates voltage-sensing multi-level RCIM. Fig. 12 shows the measured V.RBL of the proposed RRAM macro in ternary encoding. The measured V.RBL represents the CIM outputs determined by the inputs and weights. In this measurement, all the input is set to high (i.e., N.RRAM = 9) to show the wide V.RBL distribution over various combinations of the cell resistance. The curvature of the V.RBL over the CIM outputs is affected by the bias voltage of the



(a)

(b)

Fig. 12. (a) Measured readout voltages over various intermediate resistances. (b) Measured readout voltages over various weight sweeps.

feedback amplifier and the R.IRS in the multi-level RCIM. The bias voltage is set to secure the linearity of the V.RBL while considering the cell resistance (R.LRS and R.HRS), the dynamic range, and the worst case sampling margin of the V.RBL [25]. The R.IRS is, in turn, set to attain consistency of the V.RBL regardless of binary and multi-level weights. The measured results show that the V.RBL in ternary sweep when the R.IRS = 4.8 kΩ is exactly consistent with the V.RBL in binary encoding. Thus, the proposed RRAM macro employs the R.IRS optimized for the voltage-sensing RD with ternary RRAM cells.

With the optimized R.IRS, the V.RBL over various weight sweeps is also measured. The CIM output can have various combinations of resistance states. In particular, the IRS (W = 0) in ternary encoding per cell provides more degrees of freedom in composing the CIM output. Thus, the measurement of the CIM output should consider various weight sweeps with ternary resistance states. The first weight sweep is conducted by cell-by-cell transitions from the HRS→IRS→LRS. Nine HRS cells are accessed as an initial state (CIM output = −9), and then, the number of the accessed IRS cells increases while decreasing the number of the accessed HRS cells in turn. Once all the accessed RRAM cells are in the IRS (CIM output = 0), the number of the accessed LRS cells starts to increase until the CIM output is set to 9. The first weight sweep shows that the V.RBL is sufficiently linearized over the CIM output. To demonstrate the consistency of the V.RBL over various combinations of the cell resistances, the second and the third weight sweep are also conducted. The second weight sweep is conducted with nine cells from the HRS to the LRS while bypassing the IRS compared to the first weight sweep. The third weight sweep consists of three cells fixed to the HRS and six cells with transitions from the IRS to LRS in turn. The maximum difference of V.RBL over the weight sweeps is 4.75 mV. These results show that a stable and repeatable CIM readout is obtained as the cells are written from any of ternary states to another. Even if the zero point is placed slightly lower, multi-level RCIM can be conducted due to the linearized V.RBL.

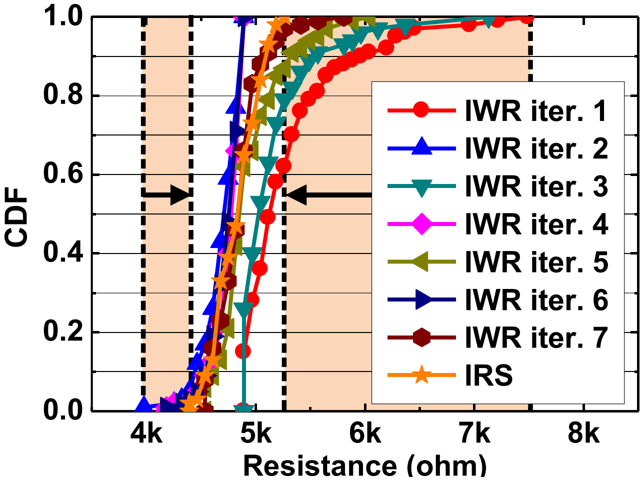


Fig. 13. Measured resistance distribution over write iterations in the IRS encoding.

Fig. 13 shows the distribution of the R.IRS over WR iterations. For each resistance state in RRAM cells, the forming processes and WRs are accordingly conducted with the IWR. The IRS cell has a wide distribution of cell resistances in a single WR operation. Over WR iterations, the peak-to-peak R.IRS decreases from 2.6 to 0.87 kΩ. The ternary encoding in the proposed RRAM macro is shown in Fig. 14. The IWR enables a tight IRS distribution to prevent the overlap of LRS or HRS distributions. The measured mean and standard deviation of the R.IRS is 4.85 kΩ and 204.90 Ω, respectively, over 100 RRAM cells.

Considering the aforementioned relationship between the bias voltage and the R.IRS and the fact that the standard deviation of the cell resistance exhibits the tendency proportional to the mean of the cell resistance, the target resistance of IRS cells in ternary encoding is set to 2.4×R.LRS but sufficiently outside the 3σ-window between the R.LRS and R.IRS. It is worth noting that the 3σ-window between the IRS and HRS appears not to be secured. However, the R.HRS has a strict lower limit since a single lower threshold is employed in the

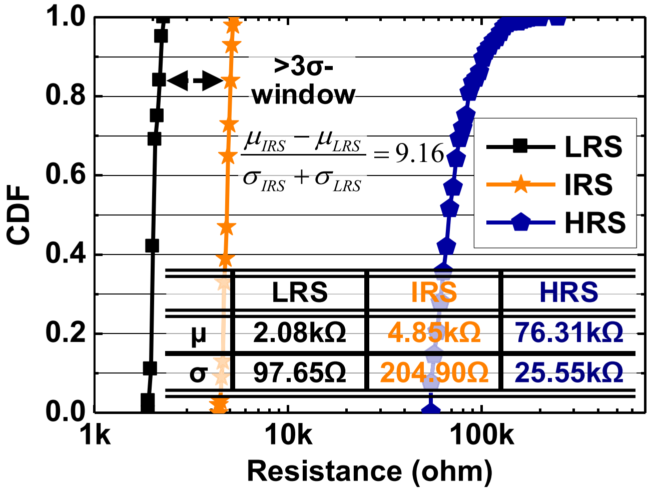


Fig. 14. Measured resistance distribution of ternary RRAM cells in the proposed RRAM macro.

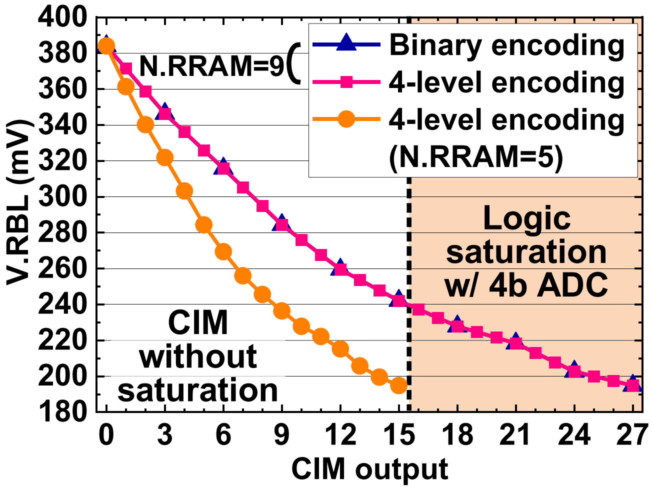


Fig. 16. Estimated readout voltages in voltage-sensing RCIM with four-level RRAM cells.

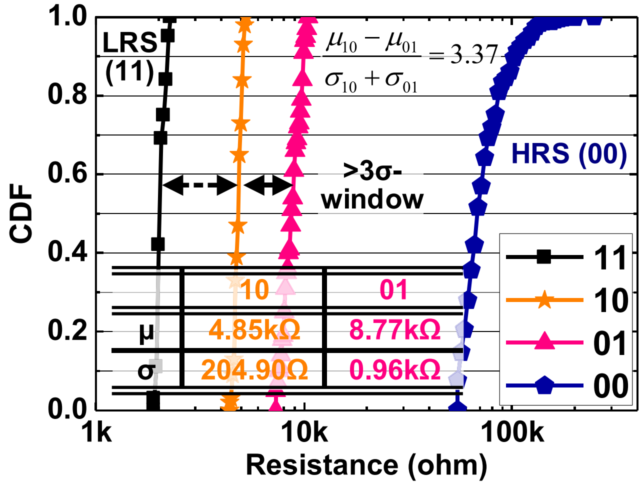


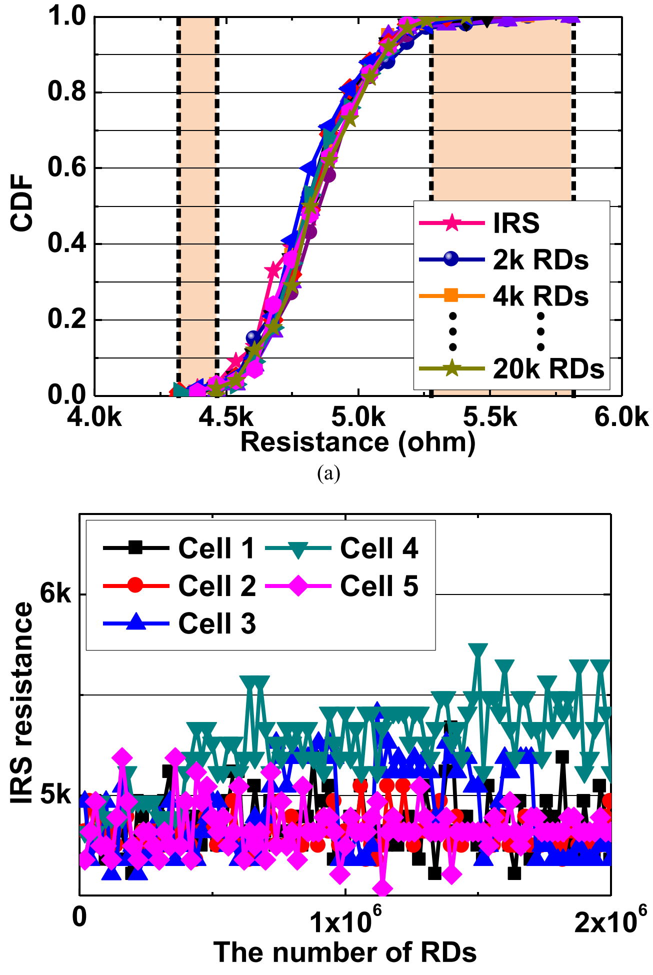
Fig. 15. Measured resistance distribution of four-level RRAM cells.

HRS encoding. Thus, the standard deviation of the R.HRS is dominantly determined by the upper side, which is not invasive to the IRS regime. Furthermore, due to the characteristics of the voltage-sensing RD (Figs. 11 and 12), it does not have an impact on the CIM performance while lessening the number of WR iterations.

Fig. 15 shows the measured resistance distribution in four-level encoding with the IWR. A new resistance state (01 in Fig. 15) is added to the ternary encoding shown in Fig. 14. The cell resistance of the 01 state is set adjacent to the IRS in ternary encoding (10 in Fig. 15) to demonstrate the dense placement of the cell resistance, thereby exhibiting the feasibility of the high resolution of multi-bit encoding per cell while securing the 3σ-window. It is worth noting that the cell resistance of the 01 and 10 states can be adjusted with the optimization for the voltage-sensing RD with four-level RRAM cells. Fig. 16 shows the estimated V.RBL with four-level RRAM cells. Since the proposed RRAM macro successfully demonstrates that the V.RBL with multi-level cells is exactly consistent with that in binary encoding (Fig. 12), the V.RBL with four-level RRAM cells can be estimated as shown in Fig. 16, thereby showing the possibility of the four-level RCIM for further scalability to advanced RCIM with multi-level cells with the estimation. In four-level encoding per cell, the CIM output exhibits positive values with unsigned inputs and weights. The 4-b ADC in the proposed RRAM macro also exploits logic saturation in four-level RCIM. The CIM output higher than 4-b ADC resolution will be saturated to 15 in four-level RCIM. In case the N.RRAM is less than 6, logic saturation does not occur even in the worst weight combination where all the accessed RRAM cells are in the LRS.

Fig. 17 shows the distribution of the R.IRS over RDs. Since the IRS cell is susceptible to read disturb due to the absence of an upper or lower bound of resistances, the tolerance of the IRS cells for read disturb is measured with 100 RRAM cells under 20k RDs and five RRAM cells under 2-million RDs. The IRS cells successfully retain the resistance with variations of 1 kΩ toward the HRS regime under 20k RDs. The measured R.IRS under 2-million RDs demonstrates that the R.IRS does not invade the LRS or HRS regime under an extreme repetitive RD scenario. In addition, we have observed that the drift toward the HRS regime appears to be bounded. It can be explained by the conflict of the drift toward the HRS regime and read disturb and it eventually prevents the unbound drift of the R.IRS.

Fig. 18 shows the estimated inference accuracy over tasks and network architectures. The estimation is conducted by applying the measured worst case error rate of the ternary CIM output, which is 13%, to the MAC operation of AI systems. In addition, the logic saturation in the ADC-based readout circuit is also considered. The inference accuracy in CIFAR-10 and CIFAR-100 is estimated with VGG-11, VGG-16, and ResNet-18 architectures. The estimated inference accuracy in four-level CIM is also shown considering the logic saturation. The simulated error rate for four-level CIM is similar to ternary CIM since the error rate of the test chip is dominantly



(a)

(b)

Fig. 17. (a) Measured resistance distribution of 100 IRS cells over 20k reads. (b) Measured distribution of five IRS cells over 2-million reads.

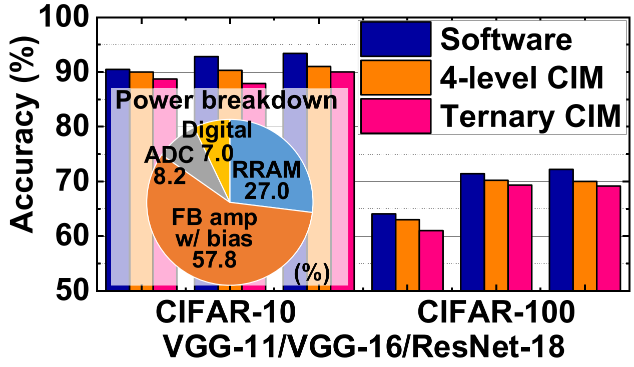


Fig. 18. Estimated inference accuracy over tasks and network architectures and the power breakdown of the test chip.

TABLE I
System Summary and Comparison

	ISSCC 2018 [16]	JSSC 2020 [17]	ISSCC 2020 [18]	ISSCC 2020 [19]	ISSCC 2019 [11]	SSC-L 2020 [33]	ISSCC 2021 [25]	This work
Technology	65 nm	55 nm	22 nm	130 nm	28 nm	40 nm	40 nm	40 nm
Memory	RRAM	RRAM	RRAM	RRAM	SRAM	RRAM	RRAM	RRAM
Supply	1.0 V	1.0 V	0.7-0.9 V	1.8 V	0.6-1.1 V	0.9 V	0.9 V	0.9 V
Array size	128Kb	128Kb	256Kb	64Kb	128Kb	8Kb	64Kb	64Kb
Sensing mode	Current	Current	Current	I&F	N/A	Voltage	Voltage	Voltage
Tolerance for low R-ratios	No	No	No	No	N/A	No	Yes	Yes
Multi-level cell	No	No	No	No	N/A	Yes (2-bit)	No	Yes (ternary)
Cell access per BL	9	9	9	256	9	64	9	9
Resolution (Input / weight / output)	Not specified (output : 1-3 bits)	1-2 bit / 3 bits / 3 bits	1-4 bits / 2-4 bits / 6-11bits	1 bit / analog / 1 bit	Integer & floating point	1 bit / 2 bits / 1 bit	1-8 bits /1-8 bits / 20 bits	1 bit / 1.58 bits / 4 bits
Peak energy efficiency (TOPS/W)	19.2	53.17	121.38	148	0.55	51.4	56.67	118.44 w/ ternary weights

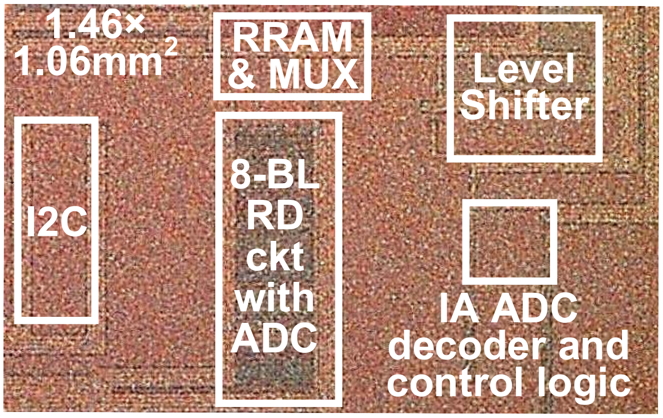


Fig. 19. Microphotograph of the test chip.

determined by non-Gaussian random telegraph noise and the readout circuit attains error-free CIM outputs when external voltages are applied instead of the V.RBL. It is worth noting that logic saturation even in four-level encoding has less impact on the accuracy under the sparsity of inputs and weights. In the case of non-sparse inputs and weights, linear or nonlinear quantization can be employed in the proposed RRAM macro due to the flexibility of ADC references. The proposed techniques help a multi-level RRAM macro achieve high algorithm-level accuracy across AI benchmarks with less than 5% loss of accuracy.

For CIM, a peak energy efficiency of 118.44 TOPS/W is measured with a ternary RRAM array, which is limited by the ON/OFF ratio of the current process. The peak energy efficiency is measured when the 9-bit input has the sparsest vector ($N = 1$) and the weight is randomly distributed. The energy efficiency using the randomized and densest input vector ($N = 9$) is 6.89 and 4.24 TOPS/W, respectively. In the estimation of the energy efficiencies, the power consumption of the V.REF generators [25] is excluded since it is negligible in high-parallelized RCIM architectures in further applications. The average power consumption per BL is 0.183 mW, including that of all peripheral circuits, where the dominant power consumption is incurred by the V.REF generators. The power breakdown of the test chip is shown in Fig. 18. The improvement of the energy efficiency compared to the binary RCIM [25] is achieved due to the power management in the digital blocks and the higher cell resistance. The presence of the R.IRS incurs the increase of the power consumption at the BL compared to the binary RCIM under the same condition. However, considering the application for RCIM with multi-level cells, a 2× higher cell resistance compared to [25] and an appropriate ON/OFF ratio is employed in this work such that the resultant energy efficiency increases significantly. Table I shows competitive metrics while addressing key challenges essential to multi-bit CIM RRAM macro with the state-of-the-art CIM architectures. The die photograph is shown in Fig. 19.

VII. Conclusion

This article presents a voltage-sensing CIM multi-level RRAM macro for reliable, high-capacity CIM architectures. RCIM architectures are of importance to achieve energy-efficient computing systems for AI systems considering the inherent MAC-friendly BL structure, high cell density, and non-volatility. However, the limited capacity of on-chip memory hinders RCIM architectures from supporting advanced AI systems. To increase the bit density by employing multi-bit encoding per cell, some challenges should be addressed in the RCIM applications. Widespread current-sensing RCIM architectures suffer from logic ambiguity incurred by the non-negligible HRS current under a low ON/OFF ratio, and it even worsens over increasing the bit resolution. Besides, the encoding level optimized for the readout circuits is imperative to achieve reliable multi-bit RCIM. Thus, the voltage-sensing multi-level RCIM is proposed to achieve reliable RCIM with multi-level cells. The proposed RRAM macro features the IWR and the voltage-sensing RD utilizing the IA BL current control with a feedback amplifier. The IWR conducts the WR operation considering the optimized encoding level to attain reliable RCIM. The extensive experiment regarding the device characteristics is also conducted to obtain the optimized WR configuration while providing comprehensive understanding of device characteristics of a fabricated RRAM array. The proposed voltage-sensing BL structure successfully achieves multi-level RCIM without logic ambiguity while exploiting the linearized V.RBL. To the best of the authors' knowledge, the test chip is the first IC supporting RCIM with multi-level cells fabricated in a standard monolithic RRAM and CMOS process. The test chip with a 101.4-kb ternary-weight RRAM array exhibits a peak energy efficiency of 118.44 TOPS/W.

Acknowledgment

The authors would like to thank Taiwan Semiconductor Manufacturing Company (TSMC) for technical discussions and for providing chip fabrication.

References

[1] A. Amaravati, S. B. Nasir, J. Ting, I. Yoon, and A. Raychowdhury, "A 55-nm, 1.0–0.4 V, 1.25-pJ/MAC time-domain mixed-signal neuromorphic accelerator with stochastic synapses for reinforcement learning in autonomous mobile robots," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 75–87, Jan. 2019.

[2] J.-H. Yoon and A. Raychowdhury, "NeuroSLAM: A 65-nm 7.25-to-8.79-TOPS/W mixed-signal oscillator-based SLAM accelerator for edge robotics," *IEEE J. Solid-State Circuits*, vol. 56, no. 1, pp. 66–78, Jan. 2021.

[3] N. Cao, M. Chang, and A. Raychowdhury, "A 65-nm 8-to-3-b 1.0–0.36-V 9.1–1.1-TOPS/W hybrid-digital-mixed-signal computing platform for accelerating swarm robotics," *IEEE J. Solid-State Circuits*, vol. 55, no. 1, pp. 49–59, Jan. 2020.

[4] J. Zhang, Z. Wang, and N. Verma, "In-memory computation of a machine-learning classifier in a standard 6T SRAM array," *IEEE J. Solid-State Circuits*, vol. 52, no. 4, pp. 915–924, Apr. 2017.

[5] 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," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 11, pp. 4172–4185, Nov. 2019.

[6] 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.

[7] D. Bankman, L. Yang, B. Moons, M. Verhelst, and B. Murmann, "An always-on 3.8 μ J/86% CIFAR-10 mixed-signal binary CNN processor with all memory on chip in 28-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 158–172, Jan. 2019.

[8] M. Kang, M.-S. Keel, N. R. Shanbhag, S. Eilert, and K. Curewitz, "An energy-efficient VLSI architecture for pattern recognition via deep embedding of computation in SRAM," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 8326–8330.

[9] M. Kang, S. K. Gonugondla, A. Patil, and N. R. Shanbhag, "A multi-functional in-memory inference processor using a standard 6T SRAM array," *IEEE J. Solid-State Circuits*, vol. 53, no. 2, pp. 642–655, Feb. 2018.

[10] M. Kang, S. K. Gonugondla, S. Lim, and N. R. Shanbhag, "A 19.4-nJ/decision, 364-K decisions/s, in-memory random forest multi-class inference accelerator," *IEEE J. Solid-State Circuits*, vol. 53, no. 7, pp. 2126–2135, May 2018.

[11] J. Wang *et al.*, "A compute SRAM with bit-serial integer/floating-point operations for programmable in-memory vector acceleration," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2019, pp. 224–226.

[12] N. Cao *et al.*, "A 65 nm image processing SoC supporting multiple DNN models and real-time computation-communication trade-off via actor-critical neuro-controller," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2020, pp. 1–2.

[13] A. Agrawal, A. Jaiswal, C. Lee, and K. Roy, "X-SRAM: Enabling in-memory Boolean computations in CMOS static random access memories," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 12, pp. 4219–4232, Dec. 2018.

[14] X. Si *et al.*, "A twin-8T SRAM computation-in-memory macro for multiple-bit CNN-based machine learning," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2019, pp. 396–398.

[15] Z. Chen, X. Chen, and J. Gu, "A 65 nm 3T dynamic analog RAM-based computing-in-memory macro and CNN accelerator with retention enhancement, adaptive analog sparsity and 44 TOPS/W system energy efficiency," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2021, pp. 240–242.

[16] 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. 494–496.

[17] C.-X. Xue *et al.*, "Embedded 1-Mb ReRAM-based computing-in-memory macro with multibit input and weight for CNN-based AI edge processors," *IEEE J. Solid-State Circuits*, vol. 55, no. 1, pp. 203–215, Jan. 2020.

[18] C.-X. Xue *et al.*, "A 22 nm 2 Mb ReRAM compute-in-memory macro with 121-28TOPS/W for multibit MAC computing for tiny AI edge devices," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2020, pp. 244–246.

[19] 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. 498–500.

[20] R. Mochida *et al.*, "A 4 M synapses integrated analog ReRam based 66.5 TOPS/W neural-network processor with cell current controlled writing and flexible network architecture," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2018, pp. 175–176.

[21] B. Chen, F. Cai, J. Zhou, W. Ma, P. Sheridan, and W. D. Lu, "Efficient in-memory computing architecture based on crossbar arrays," in *IEDM Tech. Dig.*, Dec. 2015, pp. 1–4.

[22] T. F. Wu *et al.*, "Brain-inspired computing exploiting carbon nanotube FETs and resistive RAM: Hyperdimensional computing case study," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2018, pp. 492–494.

[23] W. Li, S. Huang, X. Sun, H. Jiang, and S. Yu, "Secure-RRAM: A 40 m 16 b compute-in-memory macro with reconfigurability, sparsity control, and embedded security," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2021, pp. 1–2.

[24] S. Yin, X. Sun, S. Yu, and J.-S. Seo, "High-throughput in-memory computing for binary deep neural networks with monolithically integrated RRAM and 90-nm CMOS," *IEEE Trans. Electron Devices*, vol. 67, no. 10, pp. 4185–4192, Oct. 2020.

[25] J.-H. Yoon, M. Chang, W.-S. Khwa, Y.-D. Chih, M.-F. Chang, and A. Raychowdhury, "A 40-nm, 64-kb, 56.67 TOPS/W voltage-sensing computing-in-memory/digital RRAM macro supporting iterative write with verification and online read-disturb detection," *IEEE J. Solid-State Circuits*, vol. 57, no. 1, pp. 68–79, Jan. 2022.

[26] W. Shim, J.-S. Seo, and S. Yu, "Two-step write–verify scheme and impact of the read noise in multilevel RRAM-based inference engine," *Semicond. Sci. Technol.*, vol. 35, no. 11, Oct. 2020, Art. no. 115026.

[27] W. Shim, Y. Luo, J.-S. Seo, and S. Yu, "Impact of read disturb on multilevel RRAM based inference engine: Experiments and model prediction," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Apr. 2020, pp. 1–5.

[28] V. Milo *et al.*, "Multilevel HfO_2-based RRAM devices for low-power neuromorphic networks," *APL Mater.*, vol. 7, no. 8, Aug. 2019, Art. no. 081120.

[29] E. Perez, C. Zambelli, M. K. Mahadevaiah, P. Olivo, and C. Wenger, "Toward reliable multi-level operation in RRAM arrays: Improving post-algorithm stability and assessing endurance/data retention," *IEEE J. Electron Devices Soc.*, vol. 7, pp. 740–747, 2019.

[30] M. A. Lastras-Montaño, O. D. Pozo-Zamudio, L. Glebsky, M. Zhao, H. Wu, and K.-T. Cheng, "Ratio-based multi-level resistive memory cells," *Sci. Rep.*, vol. 11, no. 1, Jan. 2021, Art. no. 1351.

[31] W. He *et al.*, "Characterization and mitigation of relaxation effects on multi-level RRAM based in-memory computing," in *Proc. IEEE Int. Rel. Phys. Symp. (IRPS)*, Mar. 2021, pp. 1–7.

[32] W. He *et al.*, "2-bit-per-cell RRAM-based in-memory computing for area-/energy-efficient deep learning," *IEEE Solid-State Circuits Lett.*, vol. 3, pp. 194–197, 2020.

[33] J.-H. Yoon, M. Chang, W.-S. Khwa, Y.-D. Chih, M.-F. Chang, and A. Raychowdhury, "A 40 nm 100 Kb 118.44 TOPS/W ternary-weight computein-memory RRAM macro with voltage-sensing read and write verification for reliable multi-bit RRAM operation," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2021, pp. 1–4.

[34] C. Nail *et al.*, "Understanding RRAM endurance, retention and window margin trade-off using experimental results and simulations," in *IEDM Tech. Dig.*, Dec. 2016, pp. 1–4.

[35] B. Crafton, S. Spetalnick, Y. Fang, and A. Raychowdhury, "Merged logic and memory fabrics for accelerating machine learning workloads," *IEEE Design Test*, vol. 38, no. 1, pp. 39–68, Feb. 2021.

[36] Z. Chai *et al*, "Impact of RTN on pattern recognition accuracy of RRAM-based synaptic neural network," *IEEE Electron Device Lett.*, vol. 39, no. 11, pp. 1652–1655, Nov. 2018.

Jong-Hyeok Yoon (Member, IEEE) received B.S., M.S., and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2012, 2014, and 2018, respectively.

From 2018 to 2020, he was a Post-Doctoral Fellow at the Georgia Institute of Technology, Atlanta, GA, USA. In 2021, he joined the Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, South Korea, where he is currently an Assistant Professor with the Department of Electrical Engineering and Computer Science. His research interests include non-volatile memory (NVM)-based processing-in-memory architectures for deep learning, neuromorphic circuits for edge intelligence, high-speed wireline communications, and mixed-signal circuit designs.

Dr. Yoon was a recipient of the Best Regular Paper Award at the IEEE Custom Integrated Circuits Conference (CICC) in 2021.

Muya Chang (Member, IEEE) received the M.S. degree in computer science and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2020.

He is currently a Post-Doctoral Fellow with the Integrated Circuits and Systems Research Laboratory, Georgia Institute of Technology, and is advised by Prof. Arijit Raychowdhury. His research interest includes energy-efficient hardware design for distributed optimizations.

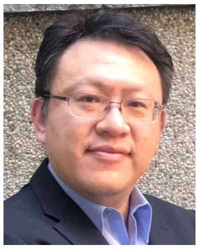
Win-San Khwa (Member, IEEE) received the B.S. degree from the University of California at Los Angeles, Los Angeles, CA, USA, in 2007, the M.S. degree from the University of Michigan, Ann Arbor, MI, USA, in 2010, and the Ph.D. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2017.

In 2012, he joined Macronix International (MXIC), Hsinchu. He is currently a Technical Manager with the Corporate Research Design Solution Department, Taiwan Semiconductor Manufacturing Company, Hsinchu, Taiwan, on emerging memory path finding and IP development. His research interests include circuit-device optimization designs of emerging memories for artificial intelligence applications.

Dr. Khwa also serves as the Digital Circuits Subcommittee Member for CICC 2021.

Yu-Der Chih received the B.S. degree in physics from National Taiwan University, Taipei, Taiwan, in 1988, and the M.S. degree in electronics engineering from National Tsing Hua University, Hsinchu, Taiwan, in 1992.

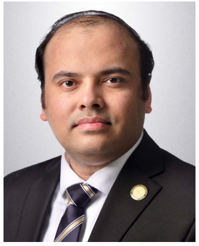
From 1992 to 1997, he was a Design Engineer of Ethernet transceiver circuits for data communication with Macronix, Hsinchu, and a Circuit Design Engineer of SDRAM with Powerchip, Hsinchu. In 1997, he joined Taiwan Semiconductor Manufacturing Company (TSMC), Hsinchu, where he was involved in the development of embedded nonvolatile memory IP, including embedded flash, OTP, MTP, and emerging memory. He is currently a TSMC Academician and the Director of the Memory Solution Division, Embedded Nonvolatile Memory Library Department.



Meng-Fan Chang (Fellow, IEEE) received the M.S. degree from Pennsylvania State University, State College, PA, USA, and the Ph.D. degree from National Chiao Tung University, Hsinchu, Taiwan.

In 2001, he co-founded IPLib, Hsinchu, where he developed embedded SRAM and ROM compilers, Flash macros, and flat-cell ROM products until 2006. He is currently a Distinguished Professor at National Tsing Hua University (NTHU), Hsinchu, and the Director of Corporate Research at Taiwan Semiconductor Manufacturing Company (TSMC), Hsinchu. Prior to 2006, he worked in industry for over ten years, including the design of memory compilers at Mentor Graphics from 1996 to 1997 and the design of embedded SRAM and Flash macros at the Design Service Division, TSMC, from 1997 to 2001. His research interests include circuit design for volatile and nonvolatile memory, ultra-low-voltage systems, 3-D memory, circuit-device interactions, cryogenic CMOS circuits, spintronic circuits, memristor logics for neuromorphic computing, and computing-in-memory for artificial intelligence.

Dr. Chang was a recipient of several prestigious national-level awards in Taiwan, including the Outstanding Research Award of MOST Taiwan, the Outstanding Electrical Engineering Professor Award, the Academia Sinica Junior Research Investigator Award, and the Ta-You Wu Memorial Award. He has been serving as an Associate Editor for IEEE JOURNAL OF SOLID-STATE CIRCUITS, IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS}, and IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS and a Guest Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS and IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS. He has also been serving on the Executive Committee for IEDM and the Subcommittee Chair for ISSCC, IEDM, DAC, and ISCAS. He was a Distinguished Lecturer of the IEEE Solid-State Circuits Society (SSCS) and the Circuits and Systems Society (CASS), the Chair of the Nano-Giga Technical Committee of IEEE CASS, and an Administrative Committee (AdCom) Member of the IEEE Nanotechnology Council. He has been serving as the Program Director for the Micro-Electronics Program at the Ministry of Science and Technology in Taiwan, the Chair for the IEEE Taipei Section, and an Associate Executive Director for Taiwan's National Program of Intelligent Electronics (NPIE) and NPIE bridge program.



Arijit Raychowdhury (Fellow, IEEE) received the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, USA, in 2007.

He joined Georgia Tech in January 2013. From 2013 to July 2019, he was an Associate Professor and held the ON Semiconductor Junior Professorship in the department. Prior to joining Georgia Tech, he held research positions at Intel Corporation for six years and Texas Instruments for one and a half years. He is currently the Steve W. Chaddick Chair and Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. He holds more than 27 U.S. and international patents and has published over 250 articles in journals and refereed conferences. His research interests include low-power digital and mixed-signal circuit design, design of power converters, signal processors, and exploring interactions of circuits with device technologies.

Dr. Raychowdhury is a mentor of IEEE Young Professionals and IEEE Women in Circuits and a Distinguished Lecturer of the IEEE Solid-State Circuits Society (SSCS). He serves on the Technical Program Committee of key circuits and design conferences, including ISSCC, VLSI Symposium, DAC, and CICC. He is the winner of several prestigious awards, including the SRC Technical Excellence Award in 2021, the Qualcomm Faculty Award in 2020, the IEEE/ACM Innovator under 40 Award, the NSF CISE Research Initiation Initiative Award (CRII) in 2015, the Intel Labs Technical Contribution Award in 2011, the Dimitris N. Chorafas Award for outstanding doctoral research and best thesis in 2007, and several fellowships. He and his students have won 14 best paper awards over the years.