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, 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.

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., $2^k$, $k \in \mathbb{Z}_0$) 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 pulselwidth (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.

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
Fig. 1. (a) Current-sensing RCIM at the BL. (b) Simplified structure of current- and voltage-sensing read (RD) 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
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 $3 \times 3$ convolutions. The $3 \times 3$ filter size in the proposed RRAM macro is determined to support the scalability of the AI systems. A $3 \times 3$ filter is a primary filter in convolutional layers where the odd-sized filter is preferred considering symmetry at the output. Multiple convolutional layers of $3 \times 3$ filters are equivalent to a single layer of larger odd-sized filters so that versatile CNNs, such as MobileNet, have utilized $3 \times 3$ 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-source-line (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 $N_{RRAM}$.

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 eliminates 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
ADC output and the CIM output with the multi-level RRAM macro considering the is scanned by a 4-bit flash ADC. The reference voltages are the $V_{\text{readout circuits should consider the}}$. The CIM output is from that has a constant range from $V_{\text{macro}}$. The with non-sparse inputs and weights in the proposed RRAM from the linearized $V_{\text{BL}}$. The overhead is only the number of the unit current source in can increase the $N_{\text{RRAM}}$ that the current-sensing RD, the logic ambiguity due to the HRS can readily support larger filters in CNNs. In the case of feedback amplifier has been conducted in the prior work [25].

It is worth noting that the proposed voltage-sensing RD can support larger filters in CNNs. In the case of the current-sensing RD, the logic ambiguity due to the HRS current limits the maximum $N_{\text{RRAM}}$, thereby limiting the scalability of RCIM. On the contrary, the voltage-sensing RD can increase the $N_{\text{RRAM}}$. 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 $V_{\text{RBL}}$. Fig. 6 shows the block diagram of the ADC-based readout circuits, the distribution of the ADC references ($V_{\text{REFs}}$), and the cases of logic saturation under non-sparse ternary-weight CIM ($N_{\text{RRAM}} = 9$).

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

Fig. 7. Schematics of the write circuit.
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 k\( \Omega \). 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 k\( \Omega \). 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 \( \mu \)s, respectively, the forming resistance increases to 4.5–5 k\( \Omega \). 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 (\( P_{\text{fail}} \)) of the MC1-3 is shown in Fig. 9. Here, \( P_{\text{fail}} \) does not represent cells where forming is not possible,
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 $P_{\text{fail}}$ 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 $V_{\text{BL, WL}}$ are used with 2.2 V of the $V_{\text{SL, WL}}$. In the reset operation, 2.6–3.0 V of the $V_{\text{SL, WL}}$ and $V_{\text{SL, WL}}$ are used. A different $V_{\text{SL}}$ is used during the set/reset operation to secure a target readout margin between the $R_{\text{LRS}}$ and $R_{\text{HRS}}$. During WR, the pulse is applied until the RRAM cell has changed to the target state. During set, the average $R_{\text{LRS}}$ decreases to 2.45 kΩ and achieves a lower standard deviation of the resistance ($\sigma_{R}$) of 1.05 kΩ. The $R_{\text{LRS}}$ does not have the tail distribution close to the threshold when $V_{\text{BL}} = 1.9$ V. During reset, the average $R_{\text{HRS}}$ increases to 34.74 kΩ for higher $V_{\text{SL}}$ and $V_{\text{WL}}$. However, the tail distribution of $R_{\text{HRS}}$ occurs even for higher voltages. The reset of RRAM cells with the $V_{\text{WL}} < 2.4$ V cannot be completed even with higher $V_{\text{SL}}$ considering the body effect at the NMOS switch in the 1T-1R structure. We observe that during WR, the $V_{\text{WL}}$ is a critical condition for reset.

To shorten the tail of $R_{\text{HRS}}$, various reset thresholds are employed with the optimized reset pulse configuration ($V_{\text{SL}} = V_{\text{WL}} = 3.0$ 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 $R_{\text{HRS}}$ distribution. With the IWR, the average $R_{\text{HRS}}$ increases to 76.31 kΩ and $\sigma_{R}$ is $\sim$25 kΩ, which is higher than $\sigma_{R}$ without the IWR. However, the standard deviation of the $V_{\text{RBL}}$ 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, $P_{\text{fail}}$ 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 $V_{\text{RBL}}$ of the proposed RRAM macro in ternary encoding. The measured $V_{\text{RBL}}$ represents the CIM outputs determined by the inputs and weights. In this measurement, all the input is set to high (i.e., $N_{\text{RRAM}} = 9$) to show the wide $V_{\text{RBL}}$ distribution over various combinations of the cell resistance. The curvature of the $V_{\text{RBL}}$ over the CIM outputs is affected by the bias voltage of the $V_{\text{WL}}$. 

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on May 07,2022 at 20:07:43 UTC from IEEE Xplore. Restrictions apply.
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 $R_{IRS} = 4.8 \, k\Omega$ 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 $\rightarrow$ IRS $\rightarrow$ 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 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 k$\Omega$. 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 k$\Omega$ and 204.90 $\Omega$, 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 $\times$ $R_{LRS}$ but sufficiently outside the $3 \sigma$-window between the $R_{LRS}$ and $R_{IRS}$. It is worth noting that the $3 \sigma$-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
Yoon et al.: 40-nm 118.44-TOPS/W VOLTAGE-SENSING COMPUTE-IN-MEMORY RRAM MACRO

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

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

Fig. 16. Estimated readout voltages in voltage-sensing RCIM with 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 $3\sigma$-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. 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 IRS.

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 IRS 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

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on May 07,2022 at 20:07:43 UTC from IEEE Xplore. Restrictions apply.
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.

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 $2 \times$ higher cell resistance compared to [25] and

<table>
<thead>
<tr>
<th>System Summary and Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
</tr>
<tr>
<td>Memory</td>
</tr>
<tr>
<td>Supply</td>
</tr>
<tr>
<td>Array size</td>
</tr>
<tr>
<td>Sensing mode</td>
</tr>
<tr>
<td>Tolerance for low bits</td>
</tr>
<tr>
<td>Multi-level cell</td>
</tr>
<tr>
<td>Cell access per BL</td>
</tr>
<tr>
<td>Resolution</td>
</tr>
<tr>
<td>Output (input / weight)</td>
</tr>
<tr>
<td>Power energy efficiency (TOPS/W)</td>
</tr>
</tbody>
</table>

The microphotograph of the test chip is shown in Fig. 19.
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 RAM 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


Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on May 07, 2022 at 20:07:43 UTC from IEEE Xplore. Restrictions apply.


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 sen 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 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.