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

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, *Senior Member, IEEE*

***Abstract*— Computing-in-memory (CIM) architectures have gained importance in achieving high-throughput energy-efficient artificial intelligence (AI) systems. Resistive RAM (RRAM) is a promising candidate for CIM architectures due to a multiply-and-accumulate (MAC)-friendly structure, high bit density, compatibility with a CMOS process, and nonvolatility. Notwithstanding the advancement of RRAM technology, the reliability of an RRAM array hinders the spread of RRAM applications such that a circuit-technology joint approach is necessary to attain reliable RRAM-based CIM architectures. This article presents a 64-kb hybrid CIM/digital RRAM macro supporting: 1) active-feedback-based voltage-sensing read (RD) to enable 1–8-b programmable vector-matrix multiplication under a low-resistance ratio of the high-resistance state to the low-resistance state in an RRAM array; 2) iterative write with verification to secure a tight resistance distribution; and 3) online RD-disturb detection in the background during CIM. The test chip fabricated in a 40-nm CMOS and RRAM process achieves a peak energy efficiency of 56.67 TOPS/W while demonstrating the eight-bitline hybrid CIM/digital MAC operation with 1–8-b inputs and weights and 20-b outputs without quantization.**

***Index Terms*— Computing-in-memory (CIM), convolutional neural network (CNN), multiply-and-accumulate (MAC), processing-in-memory, read (RD) disturb, resistive RAM (RRAM), write (WR) verification.**

Manuscript received April 16, 2021; revised June 28, 2021 and July 27, 2021; accepted July 27, 2021. This article was approved by Associate Editor Tanay Karnik. 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 Grant 2777.006, in part by the Applications and Systems-Driven Center for Energy-Efficient Integrated Nano Technologies (ASCENT) under Grant 2776.037, and in part by TSMC with technical discussions and chip fabrication. *(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 Information and Communication Engineering, 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.2021.3101209.

Digital Object Identifier 10.1109/JSSC.2021.3101209

I. Introduction

THE ever-increasing demands on energy-efficient computing systems in artificial intelligence (AI), including edge intelligence and its applications, have piqued our interest in recent years. The von Neumann architecture is widespread to support various tasks using processing elements (PEs), control units, and memory. Since the advent of AI systems and deep neural networks (DNNs), the von Neumann architecture has struggled to accommodate DNNs. DNNs in AI systems have a significant depth of layers and require a huge amount of parallel multiply-and-accumulate (MAC) operation. During the MAC operation, the inevitable data transfer of numerous weights and intermediate outputs between PEs and memory incurs prohibitive power dissipation and latency, thereby precluding certain AI applications such as battery-powered edge devices [1]–[3]. Thus, computing-in-memory (CIM) architecture has emerged to perform the energy-efficient parallel MAC operation by concurrently accessing multiple cells at a bitline (BL) of on-die memory. SRAM-based CIM architectures [4]–[14] shed light on the feasibility of CIM with appropriate energy efficiency while outperforming von Neumann architectures. However, SRAM has a large cell size (>100 F^2) and it even worsens in 8T-SRAM dedicated to CIM architectures [13], [14]. The limited capacity of on-die memory restricts the complexity of AI. Thus, emerging memory using resistances, such as resistive RAM (RRAM), magnetoresistive RAM (MRAM), and phase-change RAM (PCRAM), has been in the spotlight due to inherent MAC functionality, high bit density, and nonvolatility. PCRAM provides a moderate ON/OFF current ratio, thereby facilitating reliable CIM operation [15]. However, PCRAM requires longer read (RD) time such that dynamic power consumption is higher than other emerging memory [16]. Compared to PCRAM, MRAM features short RD pulses and the resultant energy-efficient RD [17]. On the other hand, the ON/OFF ratio of MRAM, called a tunneling magnetoresistance ratio, is extremely lower than the other emerging memory such that reliable CIM architectures cannot be achieved. On the contrary, RRAM features the aforementioned advantages of MRAM and PCRAM, such as energy-efficient RD and an appropriate ON/OFF ratio. Due to the virtues, RRAM has been used in CIM

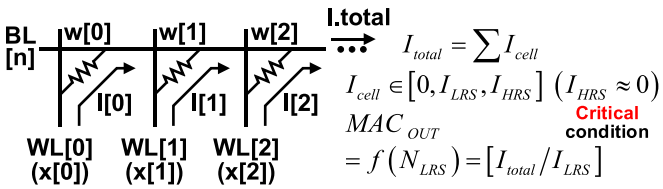


Fig. 1. Current-sensing RRAM-based CIM at the BL.

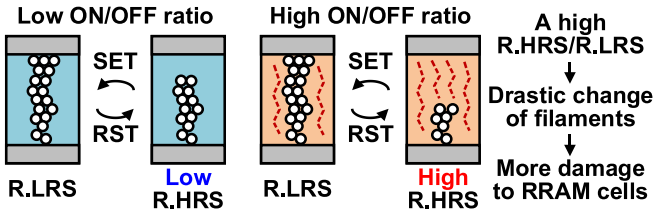


Fig. 2. Conductive filaments in RRAM cells and the tradeoff between the ON/OFF ratio and the damage to RRAM cells over WR operations.

architectures [18]–[26]. However, there are some challenges of RRAM-based CIM (RCIM) architectures. Fig. 1 shows the current-sensing RCIM at the BL. The current-sensing CIM is widespread in RCIM architectures. Each RRAM cell is programmed in a low-resistance state (LRS) or a high-resistance state (HRS) to represent the weights of DNNs. The output of the current-sensing RCIM is determined by the ratio of the total current at the BL to the LRS current. The HRS current is neglected on the premise that an RRAM array has a sufficiently high ON/OFF ratio. In the case of a low ON/OFF ratio, the aggregate current from accessed HRS cells is prone to exceed the amount of the LRS current, thereby incurring logic ambiguity. Thus, in the current-sensing RCIM, a high ON/OFF ratio should be guaranteed to support error-free MAC functionality. However, the drawback to a high ON/OFF ratio should also be considered in view of the device characteristics of RRAM [27]. Fig. 2 shows the conductive filament in RRAM cells and the tradeoff between the ON/OFF ratio and the damage to RRAM cells over write (WR) operation. The formation and rupture of conductive filaments over WR operation gradually damage RRAM cells. In particular, a high ON/OFF ratio is prone to introduce the defect to the conductive filaments over WR operation, thereby restricting the advanced AI systems with frequent weight updates such as online learning. Thus, the desirable RCIM architecture should support reliable MAC performance under a low ON/OFF ratio considering both inferences and weight updates in AI systems. Besides the tradeoff regarding the ON/OFF ratio, the reliability of the resistances of RRAM cells is another challenge in RCIM architectures [27]. Fig. 3 shows the resistance variations of RRAM cells over WR and RD operations. In the WR operation, RRAM has a different sensitivity to a WR pulse over RRAM cells. Furthermore, RRAM has no complete set or reset state in contrast with DRAM that can be fully charged or discharged. It leads to a wide distribution of resistances across RRAM cells, thereby entailing a low readout margin and the resultant error probability during CIM. In addition, RRAM may suffer from resistance variations due to temperature and

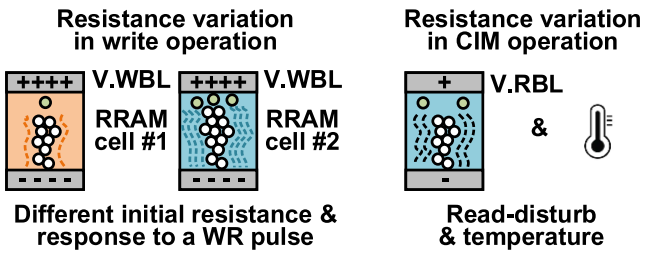


Fig. 3. Resistance variations of RRAM cells over WR and RD operations.

RD-disturb, which is the resistance drift over RD operation. Thus, the feature to secure a tight distribution of resistances in WR and RD operations should also be addressed in RCIM architectures.

In this article, a hybrid voltage-sensing CIM/digital RRAM macro [28] is proposed to support reliable CIM operation under a low ON/OFF ratio of RRAM cells for both inferences and reliable weight updates. The proposed RRAM macrofeatures hybrid CIM/digital post-MAC operation to enable 1–8-b programmable vector-matrix multiplication for versatile AI systems. Voltage-sensing RD employing the input-aware (IA) BL current control and an active feedback amplifier renders the linearized readout BL voltage (V.RBL) representing the CIM result, thereby surmounting the logic ambiguity that precludes the RCIM architectures under a low ON/OFF ratio. *In situ* iterative WR with verification (IWR) achieves a tight resistance distribution of RRAM cells with two thresholds for a target resistance state in WR operation. An online RD-disturb detector is employed to monitor RD-disturb in the background during CIM, thereby maintaining a target resistance with the restoration of resistances. We demonstrate a test chip with a 64-kb RCIM architecture performing the programmable hybrid CIM/digital MAC operation for AI systems with an energy efficiency of 56.67 TOPS/W.

The rest of this article is organized as follows. Section II describes the architecture of the proposed hybrid CIM/digital RRAM macro. Section III discusses the detailed implementation of the voltage-sensing RD enabling reliable CIM operation under a low ON/OFF ratio. Section IV delineates the IWR in the proposed RRAM macro. Section V describes the online RD-disturb detector. Section VI presents the measurement results. Section VII presents the conclusions drawn from this study.

II. Proposed Hybrid Computing-In-Memory and Digital RRAM Macro

As a circuit-technology joint approach, the proposed RRAM macro provides circuit solutions to the following challenges in RRAM technology regarding the ON/OFF ratio and the reliability in the WR and RD operations. When RRAM cells are programmed with a high ON/OFF ratio for lower RD-failure, the defect is introduced to RRAM cells over WR operation [27], which necessitates the use of circuit techniques that provide a high RD margin under a low ON/OFF ratio. Regarding the reliability, a single WR pulse creates a wide resistance distribution of RRAM cells, thereby affecting the

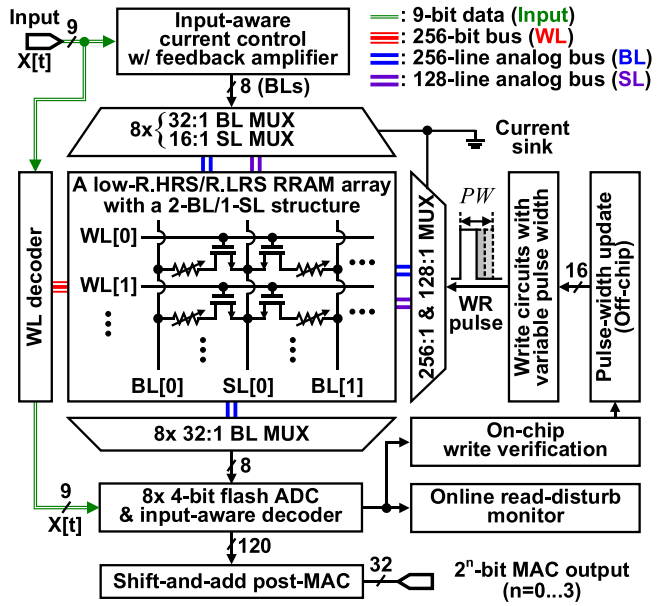


Fig. 4. Top block diagram of the proposed voltage-sensing hybrid CIM/digital RRAM macro.

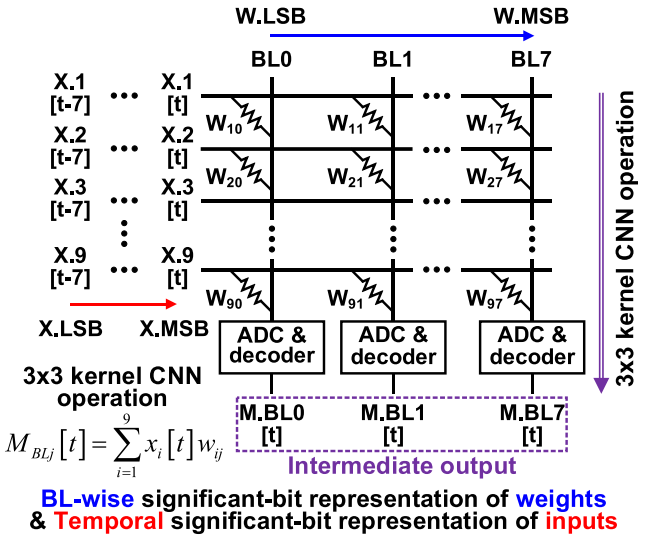


Fig. 5. Multibit CIM at the BLs of the proposed RRAM macro.

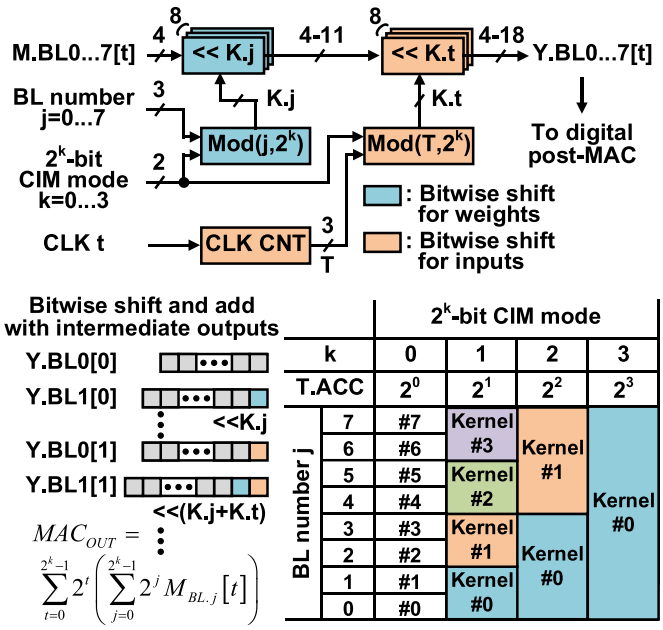


Fig. 6. Bitwise-shift-based binary weighting for the intermediate outputs in the digital post-MAC block and the structure of the final MAC output across eight BLs over multi-bit CIM modes.

accuracy and resolution in CIM. Besides, consecutive RDs in CIM operation lowers the HRS resistance and can eventually cause data corruption and RD-disturb. Thus, a circuit solution tightening and retaining the resistance distribution of RRAM cells is desirable for reliable RCIM architectures.

Fig. 4 shows the top block diagram of the proposed voltage-sensing hybrid CIM/digital RRAM macro. The proposed RRAM macro consists of a 64-kb 1T-1R RRAM array, the IA BL current control with a feedback amplifier, a 4-b flash ADC with an IA ADC decoder, a digital post-MAC block, the IWR, and the online RD-disturb detector. As the filter size in convolutional neural networks (CNNs) such as MobileNet for versatile AI systems is 3×3, the proposed RRAM macro features concurrent nine-wordline (WL) accesses. In the CIM operation, the input is fed to the WL decoder to access the nine WLs. Then, through the eight-BL MUX and the four-sourceline (SL) MUX, designated RRAM cells are selected for concurrent CIM operation. Due to the two-BL/one-SL structure of the RRAM array [29], the size of the SL MUX is reduced by 2 compared to the BL MUX. The IA BL current control activates unit BL current sources over the number of accessed RRAM cells (N.RRAM) to mitigate the nonlinearity of the V.RBL. Considering the combinations of accessed RRAM resistances, the active feedback amplifier suppresses the remaining nonlinearity of the V.RBL, thereby attaining the linearized V.RBL that represents the CIM result. The V.RBL is applied to the 4-b flash ADC. The IA ADC decoder sets the logic thresholds considering the N.RRAM to generate the 4-b CIM output, which is the intermediate output of the proposed RRAM macro. The intermediate output is fed to the digital post-MAC block. Considering the MAC configurations, the final MAC output is obtained in the post-MAC block.

In the proposed RRAM macro, 1–8-b programmable MAC is achieved with CIM and digital post-MAC operation across eight BLs. Fig. 5 shows the multi-bit CIM at the BLs of the proposed RRAM macro. To support the multi-bit input and weight with binary RRAM cells, the binary weight of the input and weight is represented in a temporal and BL-wise manner, respectively. In the binary CIM mode, each BL demonstrates single-cycle 3×3 kernel CNN operation using the binary input fed to nine WLs and the binary weights programmed at accessed RRAM cells. The resultant V.RBL is converted to the 4-b CIM output by the ADC-based readout circuit. The CIM output is directly set to the final binary MAC output of each BL without quantization. In the 2^k-bit CIM mode (when $k = 1$–3), each significant bit of the weight is distributed over 2^k BLs. The multi-bit input is fed to nine

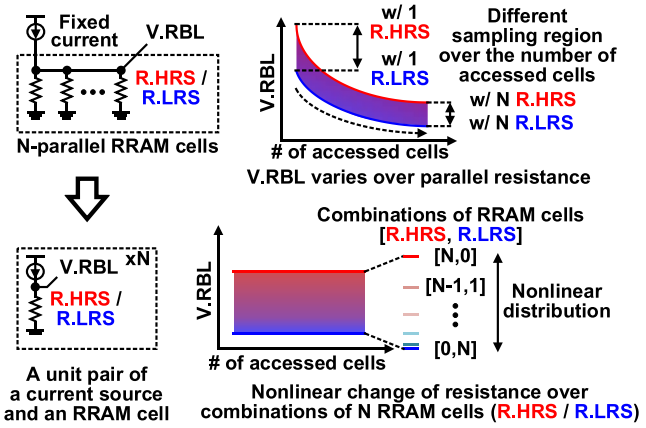


Fig. 7. Nonlinearity of the readout BL voltage with a fixed amount of the BL current, the structure of the IA BL current control, and the remaining nonlinearity over combinations of accessed RRAM resistances.

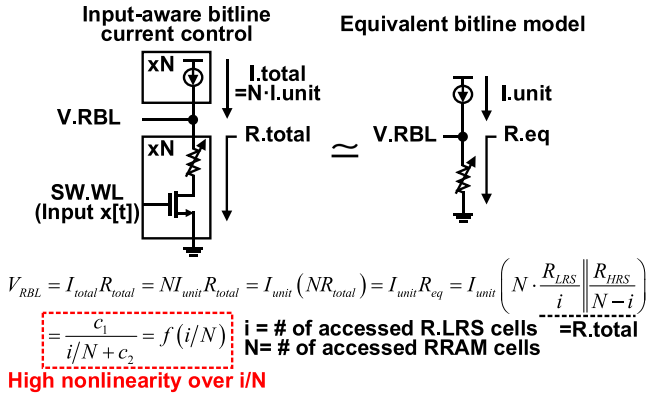


Fig. 8. Equivalent BL model under the IA BL current control and the remaining nonlinearity over the number of accessed LRS cells.

WLs from the LSB in each clock cycle. The 4-b CIM output with each significant bit of the input is the intermediate output without the consideration of the binary weight of the input and weight. Fig. 6 shows the bitwise-shift-based binary weighting for the intermediate outputs in the digital post-MAC block and the structure of the final MAC output across eight BLs over 2^k-bit CIM modes. The digital post-MAC block conducts bitwise-shift-and-add computation to convert the nonweighted intermediate output to the binary-weighted output. The binary weight of the input and weight at the intermediate output is considered with the clock cycle and the position of BLs, respectively. The unsigned 3-bit clock counter and the modulo operator extract the binary weight of the input considering the 2^k-bit CIM configuration. Then, the digital post-MAC block aggregates the weighted outputs, thereby obtaining the final 1–20 b MAC output over 2^k-clock cycles without quantization.

It is worth noting that current-sensing RCIM architectures under a low ON/OFF ratio cannot obtain accurate MAC outputs even if quantization is employed due to the aggregated HRS current incurring the aforementioned logic ambiguity. The prior art regarding the serial-input parallel-weight RCIM architecture successfully performs the RCIM operation by using current- and voltage-sensing CIMs [19], [20]. The current-sensing RCIM architecture [19] demonstrates the current-sensing CIM and MAC operation with the 2-b input and weight by using the analog post-MAC block, such as a time-interleaving binary-weighted current mirror. However, apart from the logic ambiguity, the prior work suffers from the lack of scalability to various AI systems due to the analog post-MAC structure designated to a certain resolution. Furthermore, the aggressive timing margin in the time-interleaving operation for the input limits the maximum bit resolution of the input. The prior work regarding the voltage-sensing RCIM architecture [20] achieves the flexibility in the bit resolution of CIM. However, it suffers from a lack of sampling margin over the nonlinear readout voltages. Thus, the proposed hybrid CIM/digital RRAM macro outperforms the prior arts in view of reliability and flexibility of the multi-bit MAC operation in RCIM architectures while achieving higher energy efficiency. Since a single RRAM array is employed, the proposed RRAM macro supports only positive weights compared to the prior designs supporting negative weights with two RRAM arrays. However, this is not a critical obstacle in AI applications where weight normalization and ReLU activation functions can be used to maintain positive operands only [30]. In addition, due to the hybrid architecture for CIM and digital MAC, the feature to support negative weights can be readily achieved by employing an additional RRAM array for negative weights.

While providing the aforementioned circuit solution under a low ON/OFF ratio, the remaining challenge in RRAM technology, the reliability of RRAM cells, is also addressed in the proposed RRAM macro. To tighten the resistance distribution of RRAM cells, resistance monitoring is conducted over WR and RD operations. Due to the linearized V.RBL in the proposed RRAM macro, the V.RBL in single-cell access indirectly represents the resistance of the accessed RRAM cell. Thus, during the initial forming process and WR operation, the IWR estimates whether an RRAM cell is programmed within the target range of resistances by using the 4-b ADC. If the resistance is out of the target range, the WR iteration is conducted while adjusting the WR pulsewidth (PW) until the resistance is placed within the target range. It eventually attains a tight resistance distribution of RRAM cells. In the RD operation, the online RD-disturb detector monitors the resistance of RRAM cells when a single RRAM cell is accessed and detects the drift of the resistances in the background without hindering CIM operation. In case the drift is detected, the RRAM cell is programmed again to prevent RD-disturb.

III. Active-Feedback-Based Voltage-Sensing Read

A key RCIM requirement is a high-resolution, quantization-free readout for all input-weight combinations. In current-sensing RD, the maximum number of concurrent accesses to RRAM cells is restricted even using quantization due to the logic ambiguity. The proposed RRAM macro features voltage-sensing RD to surmount the logic ambiguity under a low ON/OFF ratio, thereby achieving the aforementioned virtues such as scalability to reliable high-resolution MAC. The proposed voltage-sensing RD is comprised of the IA BL current control with an active

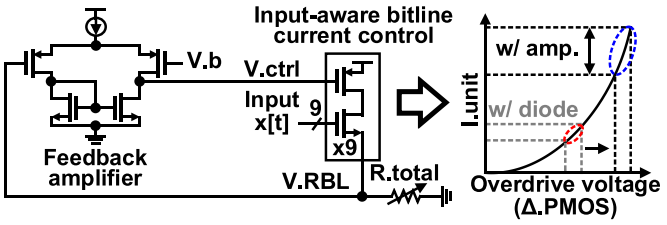


Fig. 9. Simplified BL structure of the proposed RRAM macro employing the IA BL current control with a feedback amplifier.

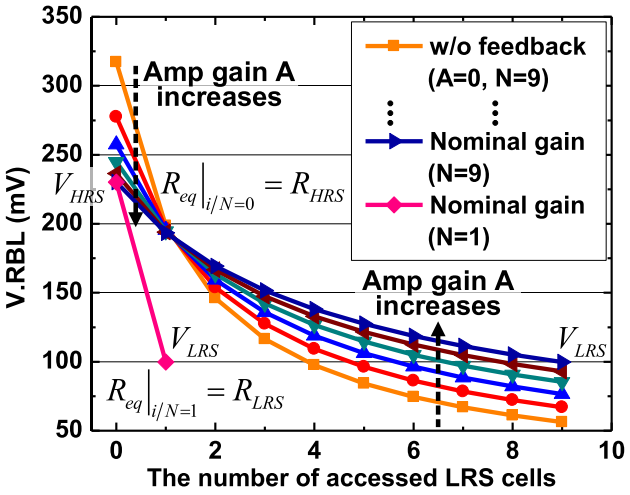


Fig. 10. Simulated readout voltages with a feedback amplifier over the number of accessed RRAM cells.

feedback amplifier to linearize the V.RBL over the N.RRAM in addition to the ADC-based readout circuits to obtain the digital CIM output.

A. Input-Aware Bitline Current Control With a Feedback Amplifier

To obtain the CIM outputs in a voltage-sensing RRAM macro, the nonlinearity of the V.RBL should be addressed. Fig. 7 shows the nonlinearity of the V.RBL in traditional voltage-sensing RD with a fixed amount of the BL current and the structure of IA BL current control. In the case of the voltage-sensing RD with a fixed current, the V.RBL drastically decreases over the N.RRAM due to the parallel resistance, thereby exhibiting an extremely narrow sampling margin at the V.RBL as more LRS cells are accessed in parallel. It eventually limits the accuracy of the CIM result even employing the ADC with nonlinear references. In addition, the V.RBL lies on different regions over the N.RRAM such that the readout circuits should support a wider sampling region that requires excessive resolution. We address this challenge by rendering the BL current proportional to the N.RRAM. The IA BL current control suppresses the aforementioned nonlinearity and attains a constant region of the V.RBL which lessens the burden in resolution. However, the nonlinearity of the V.RBL still remains over combinations of accessed RRAM resistances (see Fig. 7). Fig. 8 shows the equivalent BL model under the IA BL current control and the remaining nonlinearity over the accessed resistances. The remaining nonlinearity

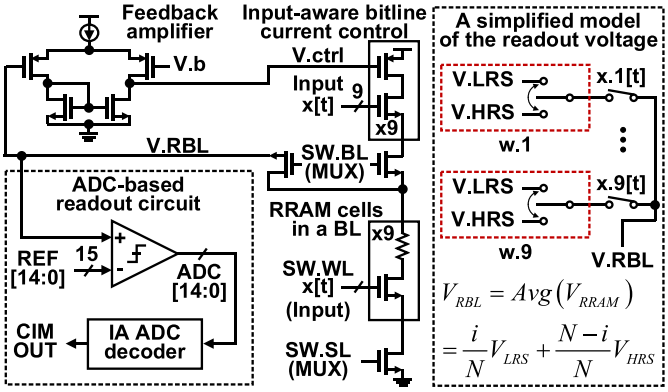


Fig. 11. Structure of the proposed voltage-sensing BL and the simplified model of the readout voltage.

can be estimated by using the equivalent model considering the combinations of accessed RRAM resistances. The model assumes a unit current source and the equivalent resistance, which is the parallel resistance multiplied by the N.RRAM. Since the V.RBL is inversely proportional to the ratio of the number of accessed LRS cells to the N.RRAM, the V.RBL suffers from high nonlinearity (see Fig. 8). Thus, a feedback amplifier is employed to suppress the remaining nonlinearity in the proposed RRAM macro. Fig. 9 shows the simplified BL structure of the proposed RRAM macro employing the IA BL current control with the feedback amplifier. Since RRAM cells retain weights in RCIM architectures, the change of RRAM resistance cannot be conducted to suppress the remaining nonlinearity. Thus, intuitively speaking, to attain the linear V.RBL, another nonlinearity should be introduced at the BL current, which neutralizes the nonlinearity incurred by the resistance (see Fig. 8). A diode-connected current source used in the current-sensing RD can provide the nonlinear BL current even in voltage-sensing RD [31]. However, the dynamic range of the BL current is still limited in suppressing the nonlinearity entailed by the resistances. In the proposed RRAM macro, the feedback amplifier shifts the bias voltage of the current source (V.ctrl) to provide a wider dynamic range of the nonlinear BL current, thereby linearizing the V.RBL. Fig. 10 shows the simulated V.RBL with a feedback amplifier over the N.RRAM. With a nominal gain of the feedback amplifier, the V.RBL is sufficiently linearized over the combinations of resistances. Even if the range of the V.RBL is decreased, the worst case sampling margin increases such that the feedback amplifier helps achieve reliable voltage-sensing RCIM architectures. Besides, the V.RBL is distributed between the V.RBL with an LRS cell (V.LRS) and that with an HRS cell (V.HRS) regardless of the N.RRAM. Thus, the V.RBL linearly represents the ratio of the number of accessed LRS cells to the N.RRAM compared to the current-sensing RD where the BL current directly represents the number of accessed LRS cells. It is noteworthy that the gain of the feedback amplifier should be carefully set to compensate for the nonlinearity of resistances. In case the gain is excessively high, the V.RBL demonstrates a fixed value regardless of the resistances, which is desirable in the current-sensing RD, not in the voltage-sensing RD.

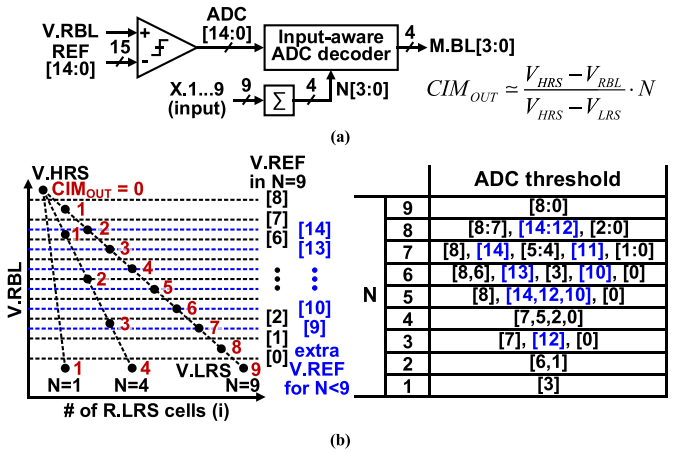


Fig. 12. (a) Block diagram of the ADC-based readout circuits and (b) distribution of the reference voltages and the logic threshold in the IA ADC decoder over the number of accessed RRAM cells.

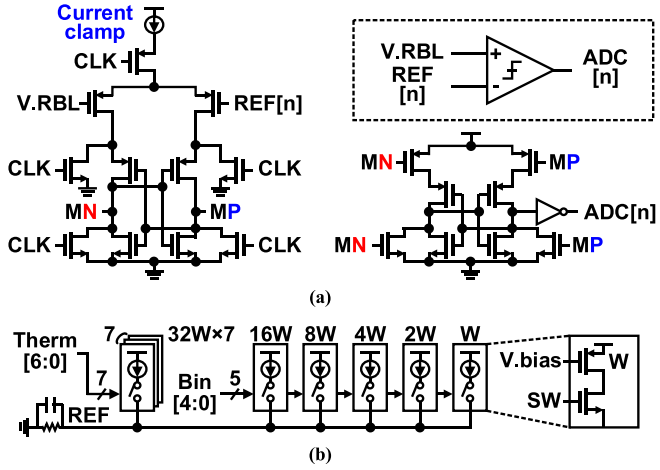


Fig. 13. (a) Schematics of the ADC comparator and (b) reference voltage generator.

The structure of the proposed voltage-sensing BL and the simplified model of the V.RBL is shown in Fig. 11. The input concurrently accesses nine RRAM cells via the BL and SL MUX to conduct CIM. The MUX using IO-voltage devices isolates the IA BL current control and the feedback amplifier, which consists of core-voltage devices from high voltages used in the WR operation, thereby protecting the BL peripheral circuits. Each BL MUX is placed to separately connect the IA BL current control and the feedback amplifier with the accessed RRAM cells. The V.RBL is affected by the total resistance comprising the accessed RRAM resistance and the parasitic resistance of the MUX switch where the BL current flows. Thus, the V.RBL is fed to the feedback amplifier and the following readout circuit via another MUX switch apart from the path of the BL current. It helps the V.RBL to be dominantly determined by the accessed RRAM resistance. Due to the IA BL current control with the feedback amplifier, the proposed BL structure can be modeled as a voltage-averaging circuit with V.HRS and V.LRS. The V.RBL represents the normalized CIM output over the N.RRAM.

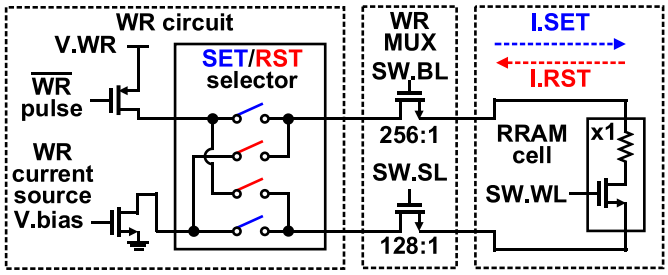


Fig. 14. Schematics of the WR circuit of the proposed RRAM macro.

It is worth noting that the proposed voltage-sensing RD can increase the N.RRAM under a low ON/OFF ratio. In the case of the current-sensing RD, the logic ambiguity due to the HRS current limits the maximum N.RRAM, which is the size of kernels in CNNs, thereby restricting RCIM architectures from supporting advanced AI systems. On the contrary, the voltage-sensing RD can support the scalability to larger filters in CNNs. In the case of increasing the filter size, the number of the current source in the IA BL current control is set to the filter size that has a negligible overhead compared to the readout circuits. Furthermore, even if the sensitivity of the readout circuits is insufficient, the proposed voltage-sensing RD can support the CIM operation by employing quantization, whereas the current-sensing RD cannot due to the logic ambiguity. Thus, the proposed RRAM macro demonstrates the feasibility of advanced AI systems in RCIM architectures.

B. ADC-Based Readout Circuits

To convert the V.RBL to the CIM output, the ADC-based readout circuits are employed in the proposed RRAM macro. Fig. 12 shows the block diagram of the ADC-based readout circuits, the distribution of the reference voltages (V.REFs), and the logic threshold in the IA ADC decoder over the N.RRAM. In the proposed RRAM macro, the V.RBL represents the normalized CIM output, which has a constant region from V.HRS to V.LRS where the CIM output is from 0 to the N.RRAM. Thus, to obtain the CIM output from the V.RBL, the readout circuit should consider not only the V.RBL but also the N.RRAM. The V.RBL exhibits various voltage levels over combinations of the number of accessed LRS cells and the N.RRAM such that a nine-level ADC is insufficient to sample the V.RBL when a fewer number of RRAM cells are accessed. Thus, the extra V.REF is employed to secure the sampling margin of the readout circuits over the N.RRAM. The V.REF consists of the nine-level V.REF for the case that the maximum number of RRAM cells is accessed (when N.RRAM = 9) and the additional six-level V.REF for the other cases. The extra V.REFs are placed between nine-level V.REFs except for the highest and lowest V.REFs (see Fig. 12). The logic threshold in the IA ADC decoder is composed of the sampling margin for various voltage levels of the V.RBL. Then, according to the N.RRAM, the IA ADC decoder obtains the CIM output considering the output of the ADC comparators.

Fig. 13 shows the schematics of the ADC comparator and the V.REF generator. The strong-arm comparator is used to

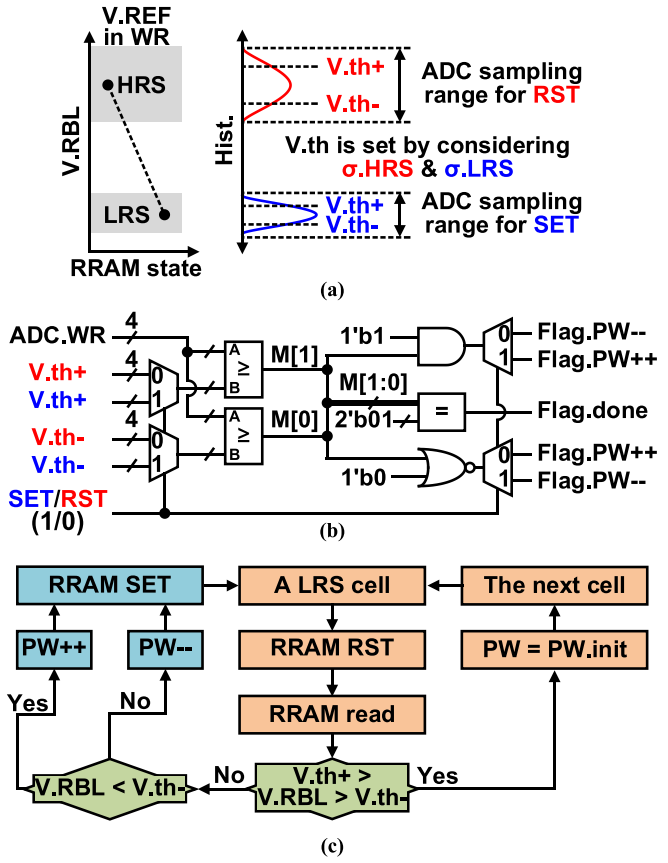


Fig. 15. (a) Distribution of reference voltages in the WR operation and the threshold in iterative WR with verification, (b) schematics of the PW adjustment, and (c) flowchart of iterative WR with verification in reset operation.

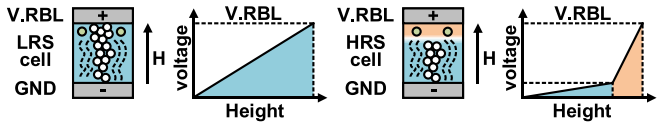


Fig. 16. RRAM cells under the readout voltage.

achieve high sensitivity in the ADC. The current clamp is employed in the strong-arm comparator to mitigate the variation of the input offset over the V.RBL. The voltage-sensing RD has a wide-range V.RBL from V.HRS to V.LRS such that the current flowing at the input stage of the comparator varies over the V.RBL during sampling. The varying current leads to the wide-range input offset depending on the V.RBL and V.REF. By limiting the current, the ADC comparator suppresses the input offset over the V.RBL. Considering the range of V.REF, the 8-b V.REF generator is employed with 3-b thermometer and 5-b binary codes achieving monotonicity of the V.REF.

The proposed voltage-sensing RD facilitates the quantization-free CIM operation with a less number of ADC references. Compared to the current-sensing RD where the range of ADC references significantly varies over the N.RRAM [18], the proposed voltage-sensing RD with only six-level extra V.REFs supports the CIM operation without quantization and logic ambiguity.

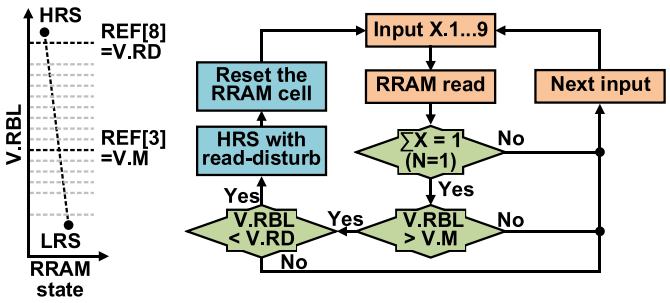


Fig. 17. Threshold and flowchart of the RD-disturb detection.

IV. Iterative Write With Verification

An open-loop single-cycle WR operation creates a wide distribution of RRAM resistance that depends on the PVT conditions of RRAM cells. In addition, RRAM has no complete set or reset state since the conductive filament cannot be completely ruptured or formed at the RRAM cells. Thus, the memory effect that the RRAM resistance in the WR operation is affected by the initial resistance exacerbates the wide distribution of RRAM resistance. This leads to errors at the CIM output. We address this challenge using *in situ* IWR with negligible overhead. Fig. 14 shows the WR circuit of the proposed RRAM macro. The WR circuit and WR MUX use IO-voltage devices to support high WR voltages, including the forming voltage of 4.0 V. The WR MUX selects an RRAM cell to be programmed. Based on the set/reset selector, the direction of the WR current is determined to set/reset the selected RRAM cell. The WR current is set to 200 μA considering the device characteristics of RRAM. The active-low WR pulse from the IWR is level-shifted and injected into the PMOS switch to apply the WR current to the RRAM cell.

Fig. 15 shows the IWR, the schematics of the PW adjustment, and the flowchart of IWR in the reset operation. In the WR operation, the resistance of an RRAM cell changes over WR pulses. After every WR pulse, the ADC-based readout circuit can estimate the resistance of the RRAM cell since the V.RBL indirectly represents the resistance of an accessed RRAM cell. The V.REFs are uniformly distributed in the LRS or HRS regime for the set or reset operation, respectively. The resistance distribution of HRS cells is much wider than that of LRS cells such that the V.REF in the reset operation has a wider range than in the set operation. The resistance threshold of the IWR is set by a digital code, which is compared with the output of the readout circuit. In the WR operation, the ADC decoder works in a normal 4-b ADC mode. The flowchart shows the reset operation with the IWR. From an LRS cell, the WR pulse is applied, and the resistance of the RRAM cell is estimated by the readout circuit. In case the V.RBL of a programmed RRAM cell is below the lower threshold in the HRS regime, the PW is increased since it is not sufficient to reset the RRAM cell. If the V.RBL is above the upper threshold in the HRS regime, the PW is decreased. The initial PW is 100 ns and it is updated in the unit of 10 ns. The initial PW and the amount of the PW update are programmable considering the WR voltages. In the set operation, the WL voltage (V.WL) and BL WR voltage (V.WBL) of 2.2 V are

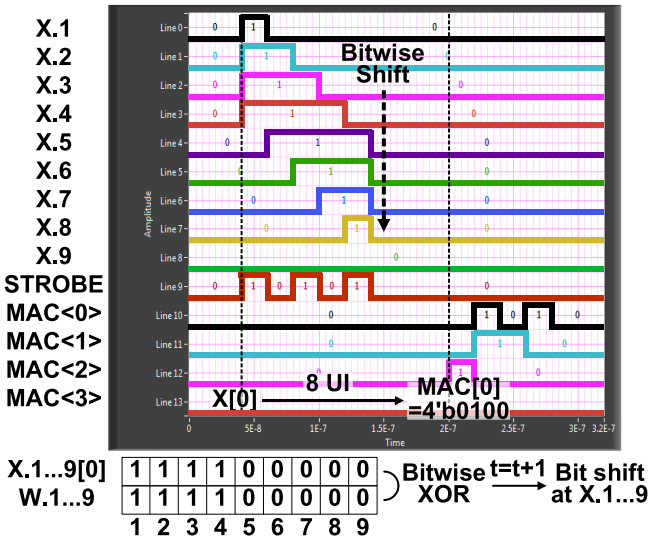


Fig. 18. Measured CIM operations over bitwise shifts of the input.

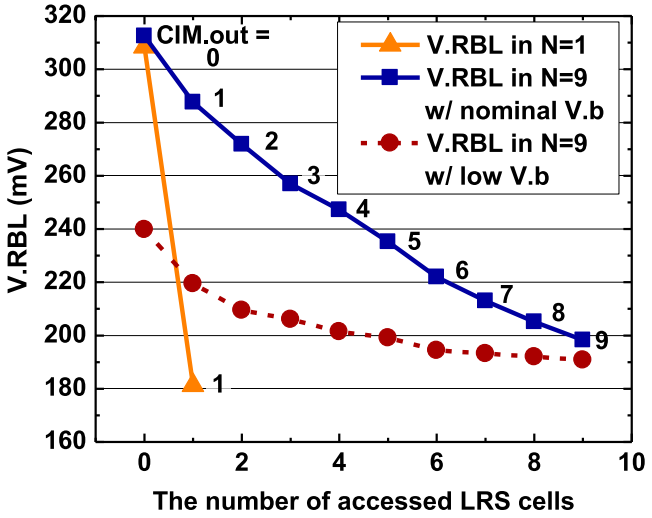


Fig. 19. Measured linear readout voltages over the number of accessed LRS cells in the proposed RRAM macro.

used. The V.WL of 3.0 V and the SL voltage (V.SL) of 2.8 V are used during reset. A higher V.WL is employed in the reset operation considering the body effect at the NMOS switch in the 1T-1R structure (see Fig. 11). After the PW update, to mitigate the aforementioned memory effect regarding the initial resistance, the RRAM cell is set to LRS, which has a narrower resistance distribution than HRS. Then, another reset process is initiated with the updated WR pulse. Finally, the tightened distribution of RRAM resistances is achieved. Once all the RRAM cells are programmed within the target range of resistance, the proposed RRAM macro starts the CIM operation with the designated V.REF distribution, as shown in Fig. 12. The prior work regarding the WR-verify achieves narrow resistance distribution without the return to the initial resistance [32]. The proposed IWR may need more iterations at the first WR process due to the reinitialization. However, the IWR can obtain the optimal WR PW that resets the RRAM

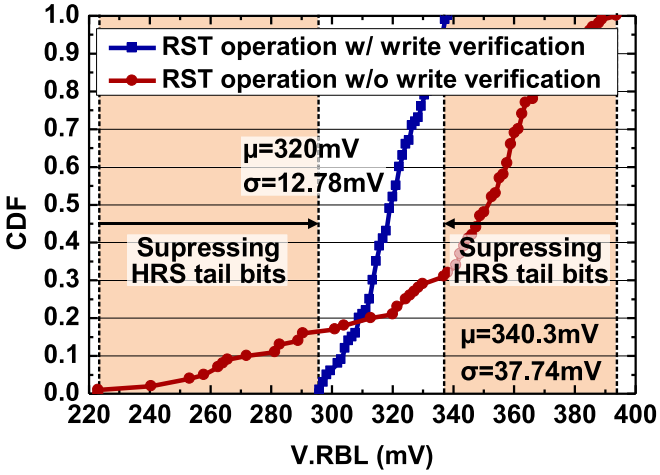


Fig. 20. Measured V.HRS distribution with and without the iterative WR with verification.

cell with a single WR pulse. Thus, the IWR enables fewer iterations over further back-to-back WRs.

It is worth noting that there is a tradeoff between the range of thresholds in the IWR and the number of WR iterations. In case a wider distribution of resistance is allowed, the number of iterations is decreased. In addition, since the IWR uses the 4-b ADC in sensing the RRAM resistance, the proposed IWR can employ a variable PW update considering the distance from the target resistance, thereby decreasing the number of iterations in the IWR.

V. Online Read-Disturb Detector

The nonvolatility of RRAM helps avoid a frequent data refresh process in RCIM architectures. In addition to data retention in RRAM technology, RCIM architectures should monitor whether the RRAM resistance is placed within the target range to secure reliable error-free CIM operation as a solution to long-term functionality in a circuit-domain approach. Since RRAM exhibits appropriate data retention, the RD-disturb detection in the background is desirable. Fig. 16 shows the RD-disturb at LRS and HRS cells. The V.RBL during CIM gradually lowers the RRAM resistances since the RD operation is the same as the set operation with a low V.WBL. In particular, HRS cells more suffer from RD-disturb since a higher voltage per unit length is applied to the insulator. Thus, the online RD-disturb detector is employed to maintain HRS resistances over RD operation.

Fig. 17 shows the threshold and the flowchart of the RD-disturb detection. The distribution of the V.REFs is the same as the CIM mode since the online RD-disturb detector operates in the background without hindering the CIM operation. The highest V.REF is employed as the threshold in the RD-disturb detection. In case a single RRAM cell is accessed (when N.RRAM = 1) during CIM, the online RD-disturb detector checks whether the accessed RRAM cell is in HRS or not by using the logic threshold (see Fig. 12). In case a single HRS cell is accessed, the decrease of HRS resistances is monitored by using the threshold of the online RD-disturb detector. In case the V.RBL of the HRS cell is lower than

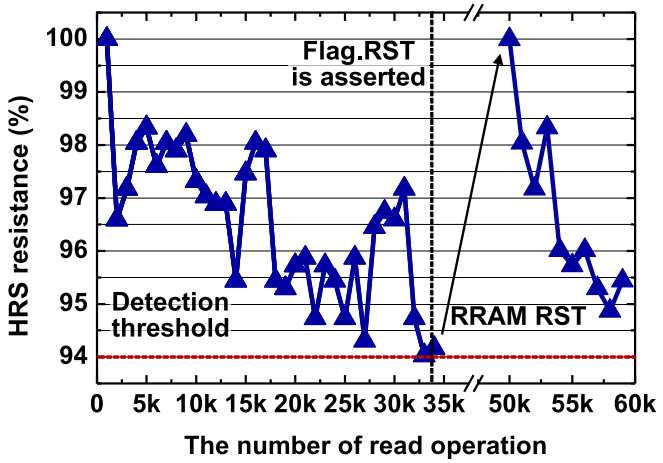


Fig. 21. Measured drift of the normalized HRS resistance over RD operation and the reset of the HRS cell owing to the online RD-disturb detector.

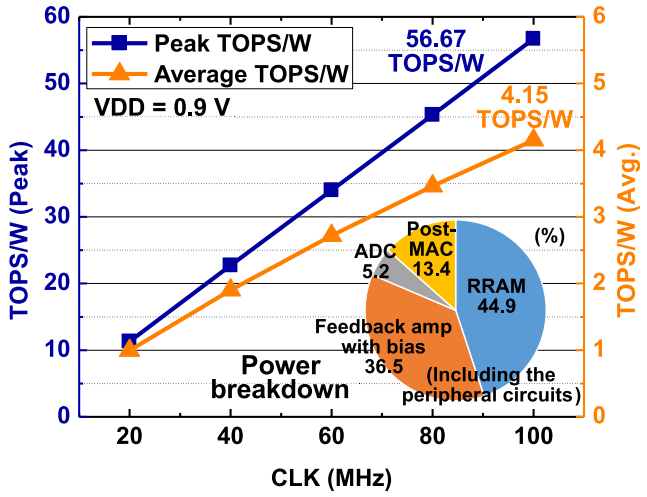


Fig. 22. Measured peak and average energy efficiency and the simulated power breakdown of the proposed RRAM macro.

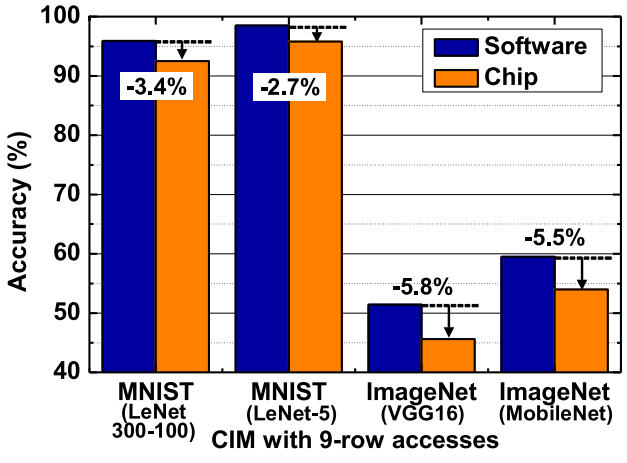


Fig. 23. Estimated inference accuracy over tasks and network architectures.

the threshold, the reset operation is initiated for the accessed HRS cell. A single WR pulse is sufficient to reset the HRS cell that suffers from RD-disturb since the resistance of the RRAM cell is still close to the HRS resistance. An HRS cell

TABLE I

SYSTEM SUMMARY AND COMPARISON

	ISSCC 2018 [18]	JSSC 2020 [19]	ISSCC 2020 [21]	ISSCC 2020 [22]	ISSCC 2019 [11]	This work
Technology	65 nm	55 nm	22 nm	130 nm	28 nm	40 nm
Memory	RRAM	RRAM	RRAM	RRAM	SRAM	RRAM
Supply	1.0 V	1.0 V	0.7-0.9 V	1.8 V	0.6-1.1 V	0.9 V
Sensing mode	Current	Current	Current	I&F	N/A	Voltage
A low R-ratio CIM architecture	No	No	No	No	N/A	Yes
Iterative write w/ verification	No	No	No	No	N/A	Yes
Online read-disturb detection	No	No	No	No	N/A	Yes
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	1-8 bits / 1-8 bits / 20 bits
Energy efficiency	19.2 TOPS/W	53.17 TOPS/W	121.38 TOPS/W	148 TOPS/W	0.55 TOPS/W	56.67 TOPS/W

*The energy efficiencies are measured for 1-bit operations.

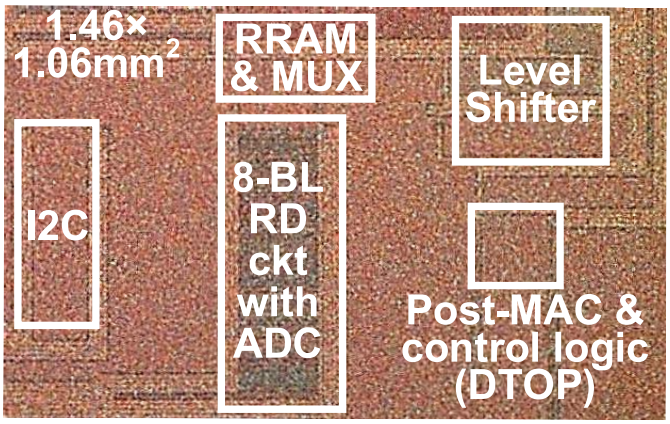


Fig. 24. Microphotograph of the test chip.

is intermittently monitored on the premise that the input bit is independent across WLs. Considering nine-WL accesses in the proposed RRAM macro, the probability to access a single RRAM cell is 1/512. This rate of the RD-disturb detection is sufficient to provide RD-disturb-free operation with no performance or power penalty due to the excellent data retention in RRAM. The online RD-disturb detector is also expected to address other time- and temperature-dependent drifts common in RRAM.

VI. MEASUREMENT RESULTS

The proposed hybrid CIM/digital RRAM macro is fabricated in a 40-nm CMOS and RRAM process and assembled in a QFN48 package. The test chip supports voltage-sensing programmable multi-bit CIM and MAC operation while achieving the reliability of RRAM cells. The measured 1-b CIM operation in the RRAM macro is shown in Fig. 18 where the logic waveform is enhanced for visibility. During the measurement, the weight is comprised of four consecutive LRS cells and five HRS cells, and the input with four-cell accesses (when N.RRAM = 4) is bitwise shifted. The resultant CIM output is generated with the latency of eight unit intervals. The data and data strobe of 50 Mb/s are applied to

the proposed RRAM macro and the system clock frequency is 100 MHz in the digital post-MAC block. The latency is dominantly incurred by the digital post-MAC block supporting the programmability of multi-bit CIM operation. Fig. 19 shows the measured linear V.RBL over the number of accessed LRS cells in the proposed RRAM macro. We indirectly measure the V.RBL for varying N.RRAMs and show the measured results. The linearized V.RBL is achieved due to the IA BL current control with the feedback amplifier. Regardless of the N.RRAM, the V.RBL is placed in a constant region such that the ADC-based readout circuits can obtain the CIM output. In addition, the proposed voltage-sensing RD provides more sampling margin in the case of lower N.RRAMs, thereby attaining additional reliability under sparse inputs compared to the current-sensing RD. The discrepancy of the lowest V.RBL when the N.RRAM is 1 and 9 is incurred by the parasitic resistance of an SL MUX switch (see Fig. 11). This discrepancy can be mitigated by adjusting the combinations of the ADC threshold in the readout circuit (Fig. 12), thereby obtaining the reliable CIM output. With a low bias voltage (V.b in Fig. 11) of the feedback amplifier, the V.RBL demonstrates semi-constant voltages that are desirable in the current-sensing RD. Thus, the bias voltage is carefully set to demonstrate the linearized V.RBL for the voltage-sensing RD. Fig. 20 shows the measured V.RBL distribution of HRS cells with and without the IWR. The V.RBL indirectly represents the resistance of an accessed RRAM cell. Without the IWR, a single reset operation leads to a wide distribution of the HRS tail bits that are the outlier in the HRS distribution. By employing the IWR, the distribution of HRS resistances is tightened over WR operation. Thus, the tail bit is suppressed. Due to the IWR, the standard deviation of V.RBL of HRS cells is decreased from 37.74 to 12.78 mV. The average number of iterations is 5.07. Fig. 21 shows the measured drift of the normalized HRS resistance over RD operation and the restoration of the HRS resistances due to the online RD-disturb detector. In this measurement, the extreme condition where the V.WL of 1.5 V and the VDD of 1.1 V for the BL peripheral circuits are applied in conjunction with the external heat of 85 °C is used to accelerate the RD-disturb. Under the condition, the HRS resistance decreases over RD operation. Then, once the HRS resistance reaches the threshold of the RD-disturb detector, the flag to reset the HRS cell is asserted. Finally, the HRS cell is reset to restore a high resistance. The online RD-disturb detector successfully detects the decrease of HRS resistances without hindering CIM operation. Fig. 22 shows the measured energy efficiency and the simulated power breakdown of the proposed RRAM macro. For CIM, the proposed RRAM macro achieves the average (peak) energy efficiency of 4.15 (56.67) TOPS/W for 1-bit operations. The energy efficiency is limited by low RRAM resistances providing appropriate data retention in the current process. The peak energy efficiency is measured when the 9-bit input has the sparsest vector and the weight is randomly distributed. The measured peak energy efficiency in the 2^k-bit CIM modes ($k = 1$, 2, and 3) are 28.1, 14.1, and 7.0 TOPS/W, respectively, since the 2^k-clock cycles are required in the multi-bit CIM operation. The simulated power breakdown is shown in the case of the average energy efficiency, which has the 50% activity of the inputs and weights. The corresponding power consumption per BL is 0.205 mW. The power consumption of V.REF generators is excluded since it is negligible in high-parallelized RCIM architectures sharing V.REFs across BLs. Fig. 23 shows the estimated inference accuracy over tasks and network architectures. The estimation is conducted by applying the worst case error rate of the CIM output to the MAC operation of AI systems. The measured worst case error rate is 13% when the number of accessed LRS cells and N.RRAM is 9. It is worth noting that the error rate is dominantly determined by the noise of RRAM cells since the readout circuit attains error-free CIM outputs when external voltages are applied instead of the V.RBL. The inference accuracy in MNIST and ImageNet is estimated with LeNet, VGG, and MobileNet architectures. The proposed RRAM macro sheds light on the feasibility of supporting high algorithm-level accuracy across a suite of AI benchmarks with less than a 6% loss of accuracy. The accuracy shown in Fig. 23 is derived from simpler DNNs. The state-of-the-art accuracy on more complex networks is higher and can be analyzed as a part of future work. Table I summarizes and compares the state-of-the-art CIM architectures. Compared to the prior arts, the voltage-sensing CIM architecture enables the reliable CIM and MAC operation under a low ON/OFF ratio while monitoring the RRAM resistance in WR and RD operations. In addition, the proposed RRAM macro supports the programmable 1–8 bit MAC operation. The die photograph is shown in Fig. 24. The test chip is fabricated in a TSMC 40-nm CMOS and RRAM process.

VII. Conclusion

This article presents a voltage-sensing hybrid CIM/digital RRAM macro for the reliable CIM and programmable MAC operation. RCIM architectures are of importance in achieving energy-efficient computing systems due to an inherent MAC-friendly structure, high bit density, and nonvolatility of RRAM. However, some challenges of RRAM technology, such as the tradeoff between the ON/OFF ratio and the damage to RRAM cells, should be addressed in a circuit-domain approach. In particular, under a low ON/OFF ratio, widespread current-sensing RD cannot provide reliable CIM due to the logic ambiguity incurred by the high HRS current. Thus, the design presented in this article surmounts the drawback of a low ON/OFF ratio by incorporating voltage-sensing RD with the resistance calibration in WR and RD operations. The proposed RRAM macro enables voltage-sensing RD due to the IA BL current control with the feedback amplifier. The proposed BL structure attains the linearized V.RBL representing the CIM output without suffering from the logic ambiguity. Furthermore, the resistance distribution of RRAM cells is tightened by the IWR in the WR operation and retained by the online RD-disturb detector without hindering CIM operation. The digital post-MAC block renders programmable 1–8-b MAC operation to support both versatile AI systems. The proposed RRAM macro with a 64-kb RRAM array demonstrates the correct CIM and MAC operation. The test chip fabricated in a 40-nm CMOS and RRAM process exhibits a peak energy efficiency of 56.67 TOPS/W.

REFERENCES

[1] A. Amaravati, S. B. Nasir, J. Ting, I. Yoon, and A. Raychowdhury, "A 55-nm, 1.0–0.4V, 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.*, "14.2 A compute SRAM with bit-serial integer/floating-point operations for programmable in-memory vector acceleration," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 224–226.

[12] N. Cao, B. Chatterjee, M. Gong, M. Chang, S. Sen, and A. Raychowdhury, "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.*, "24.5 A twin-8T SRAM computation-in-memory macro for multiple-bit CNN-based machine learning," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 396–398.

[15] M. L. Gallo and A. Sebastian, "An overview of phase-change memory device physics," *J. Phys. D, Appl. Phys.*, vol. 53, Mar. 2020, Art. no. 213002.

[16] S. Mittal, J. S. Vetter, and D. Li, "A survey of architectural approaches for managing embedded DRAM and non-volatile on-chip caches," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 6, pp. 1524–1537, Jun. 2015.

[17] O. Golonzka *et al.*, "MRAM as embedded non-volatile memory solution for 22FFL FinFET technology," in *IEDM Tech. Dig.*, Dec. 2018, pp. 18.1.1–18.1.4.

[18] W.-H. Chen *et al.*, "A 65 nm 1Mb nonvolatile computing-in-memory ReRAM macro with sub-16ns multiply-and-accumulate for binary DNN AI edge processors," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 494–496.

[19] 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.

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

[21] C.-X. Xue *et al.*, "15.4 A 22 nm 2Mb ReRAM compute-in-memory macro with 121-28TOPS/W for multibit MAC computing for tiny AI edge devices," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 244–246.

[22] W. Wan *et al.*, "33.1 A 74 TMACS/W CMOS-RRAM neurosynaptic core with dynamically reconfigurable dataflow and *in-situ* transposable weights for probabilistic graphical models," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 498–500.

[23] R. Mochida *et al.*, "A 4M 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.

[24] 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. 17.5.1–17.5.4.

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

[26] J.-H. Yoon, M. Chang, W.-S. Khwa, Y.-D. Chih, M.-F. Chang, and A. Raychowdhury, "A 40 nm 100Kb 118.44TOPS/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.

[27] 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. 4.5.1–4.5.4.

[28] J.-H. Yoon, M. Chang, W.-S. Khwa, Y.-D. Chih, M.-F. Chang, and A. Raychowdhury, "A 40 nm 64Kb 56.67TOPS/W read-disturb-tolerant compute-in-memory/digital RRAM macro with active-feedback-based read and *in-situ* write verification," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 64, Feb. 2021, pp. 404–406.

[29] C.-C. Chou *et al.*, "An N40 256K×44 embedded RRAM macro with SL-precharge SA and low-voltage current limiter to improve read and write performance," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 478–480.

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

[31] 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.

[32] 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," *Semiconductor Sci. Technol.*, vol. 35, no. 11, Oct. 2020, Art. no. 115026.

Jong-Hyeok Yoon (Member, IEEE) received the 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 with 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 Information and Communication Engineering. 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 ECE Professor Arijit Raychowdhury. His research interest includes energy-efficient hardware design for distributed optimizations.

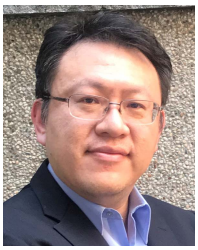
Win-San Khwa (Member, IEEE) received the B.S. degree from the University of California at Los Angeles, Los Angeles, CA, USA, in 2007, the M.S. degree from the University of Michigan, Ann Arbor, MI, USA, in 2010, and the Ph.D. degree in electrical engineering from National Tsing Hua University, Hsinchu,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

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

Dr. Khwa 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 non-volatile memory IP, including embedded flash, OTP, MTP, and emerging memory. He is a TSMC Academician and is currently a Director of the Embedded Nonvolatile Memory Library Department in the Memory Solution Division.



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

Prior to 2006, he worked in industry for over ten years. This included the design of memory compilers (Mentor Graphics, Wilsonville, OR, USA; from 1996 to 1997) and the design of embedded SRAM and flash macros (Design Service Division, TSMC, Hsinchu; from 1997 to 2001). 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 with National Tsing Hua University (NTHU), Hsinchu, and the Director of Corporate Research, TSMC. His research interests include circuit design for volatile and nonvolatile memory, ultra-low-voltage systems, 3-D memory, circuit–device interactions, 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 the IEEE TVLSI, IEEE TCAS-I, and IEEE TCAD. He is serving on the Executive Committee for IEDM. He is serving as the Subcommittee Chair for ISSCC, IEDM, DAC, ISCAS, VLSI-DAT, and ASP-DAC. He was a Distinguished Lecturer of the IEEE Solid-State Circuits Society (SSCS) and Circuits and Systems Society (CASS), the Chair of the Nano-Giga Technical Committee of CASS, and an Administrative Committee (AdCom) Member of the IEEE Nanotechnology Council. He is 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 the Associate Executive Director for Taiwan's National Program of Intelligent Electronics (NPIE) and the NPIE Bridge Program.

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

His industry experience includes five years as a Staff Scientist with the Circuits Research Laboratory, Intel Corporation, Portland, OR, USA, and a year as an Analog Circuit Researcher with Texas Instruments Inc., Bengaluru, India. He joined the Georgia Institute of Technology, Atlanta, GA, USA, in 2013, where he is currently an Associate Professor with the School of Electrical and Computer Engineering and. He holds more than 25 U.S. and international patents and has published over 100 articles in journals and refereed conferences. His research interests include low-power digital- and mixed-signal circuit design, device–circuit interactions, and novel computing models and hardware realizations.

Dr. Raychowdhury was a recipient of the Intel Early Faculty Award in 2015, 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 in 2007, the Best Thesis Award from the College of Engineering, Purdue University, in 2007, and multiple best paper awards and fellowships. He holds an ON Semiconductor Junior Professorship.