Applied Physics Letters

Dual-configuration in-memory computing bitcells using SiO_x RRAM for binary neural networks

Cite as: Appl. Phys. Lett. **120**, 034102 (2022); https://doi.org/10.1063/5.0073284
Submitted: 29 September 2021 • Accepted: 04 January 2022 • Published Online: 18 January 2022

Sandeep Kaur Kingra, Vivek Parmar, Shubham Negi, et al.

Applied Physics Letters ARTICLE scitation.org/journal/apl

Dual-configuration in-memory computing bitcells using SiO_x RRAM for binary neural networks

Cite as: Appl. Phys. Lett. **120**, 034102 (2022); doi: 10.1063/5.0073284
Submitted: 29 September 2021 · Accepted: 4 January 2022 ·
Published Online: 18 January 2022

Sandeep Kaur Kingra,[1] Vivek Parmar,[1] Shubham Negi,[1] Alessandro Bricalli,[2] Giuseppe Piccolboni,[2] Amir Regev,[2] Jean-François Nodin,[3] Gabriel Molas,[3] and Manan Suri[1,a)]

AFFILIATIONS

[1]Indian Institute of Technology Delhi, Hauz Khas, New Delhi, India
[2]Weebit Nano, Hod Hasharon, Israel
[3]CEA-Leti, Grenoble, France

Note: This paper is part of the APL Special Collection on Neuromorphic Computing: From Quantum Materials to Emergent Connectivity.
[a)]**Author to whom correspondence should be addressed:** manansuri@ee.iitd.ac.in

ABSTRACT

Conventional DNN (deep neural network) implementations rely on networks with sizes in the order of MBs (megabytes) and computation complexity of the order of Tera FLOPs (floating point operations per second). However, implementing such networks in the context of edge-AI (artificial intelligence) poses limitations due to the requirement of high precision computation blocks, large memory requirement, and memory wall. To address this, low-precision DNN implementations based on IMC (in-memory computing) approaches utilizing NVM (non-volatile memory) devices have been explored recently. In this work, we experimentally demonstrate a dual-configuration XNOR (exclusive NOR) IMC bitcell. The bitcell is realized using fabricated 1T-1R SiO_x RRAM (resistive random access memory) arrays. We have analyzed the trade-off in terms of circuit-overhead, energy, and latency for both IMC bitcell configurations. Furthermore, we demonstrate the functionality of the proposed IMC bitcells with mobilenet architecture based BNNs (binarized neural networks). The network is trained on VWW (visual wake words) and CIFAR-10 datasets, leading to an inference accuracy of ≈80.3% and ≈84.9%, respectively. Additionally, the impact of simulated BER (bit error rate) on the BNN accuracy is also analyzed.

Edge-AI (artificial intelligence) based on DNNs (deep neural networks) has emerged as an area of prime focus for researchers as well as industry due to widespread applications such as smart cities, autonomous systems, and pervasive computing. Conventional DNN implementations require high-precision computing using floating-point computations, which escalates energy costs. Additionally, due to physical separation between the storage/memory unit and the processor, a memory ↔ compute bottleneck causes further limitations due to the increasing size of networks. These factors further challenge hardware implementations in terms of computation, memory, and communication (refer to Appendix S1 in the supplementary material). To address this, low-precision DNN techniques have been proposed, where networks are realized using binary precision that makes them feasible for edge-deployment by cutting down memory requirements (upto 32×). To alleviate the memory wall issue, IMC (in-memory computing) approaches based on NVM (non-volatile memory) devices have emerged as a promising candidate. Some of the NVM technologies explored for IMC applications for analog multiplication, including: Flash,[1,2] RRAM (resistive random access memory),[3–6] and MRAM (magnetoresistive random access memory).[7] RRAM based XNOR (exclusive NOR) bitcells provide the following advantages: (i) less area and non-volatility compared to SRAM (static random access memory) (≈150 F^2 per bitcell); (ii) lower operating voltages and faster memory access time compared to Flash; and (iii) lower fabrication cost, reduced area, and write energy compared to MRAM. In BNNs (binarized neural networks), multiplication is realized in the form of XNOR operations, while accumulation is implemented as bit-counting (*popcount*) of bitwise XNOR outputs (refer Appendix S5 of the supplementary material for more details). Using RRAM based IMC, it is possible to implement XNOR BNN operations either in row[3,4] or in column configurations[5] with both being demonstrated in the literature separately. In this work, we propose a 2T-2R XNOR IMC bitcell using SiO_x RRAM devices and further exploit it for realizing BNNs on hardware. Compared to the literature, we present the following aspects in this

work: (i) experimental demonstration and validation of a dual-configuration (row-wise and column-wise) 2T-2R XNOR bitcell using fabricated SiO_x based 1T-1R RRAM device arrays, (ii) performance benchmarking of the state-of-the-art BNN[8] (no integer weights/activations even at input) for person detection using VWW (visual wake words) dataset[9] and CIFAR-10,[10] (iii) analysis on impact of the 2T-2R array size for both XNOR IMC bitcell configurations, and (iv) analysis of simulated BER (bit error rate) on performance of the XNOR IMC bitcell and BNN accuracy.

Figure 1(a) shows the SEM cross section of the fabricated SiO_x RRAM device integrated on top of 130 nm CMOS technology in the BEOL (back end of line) between the fourth and fifth metal layers. The device stack is composed of TiN as an inert BE (bottom electrode), non-stoichiometric SiO_x as an active switching layer, followed by a Ti layer (acting as an oxygen scavenging layer) and a TiN layer. Memory dots are obtained by etching, followed by passivation layer deposition. Then the TE (top electrode) contact is opened, and the fifth metal line is processed. A single resistive memory cell comprises 1T-1R (1 transistor-1 resistor), where the NMOS transistor acts as the selection element in the array. The optimized 1T-1R bitcell area is $30F^2$. An electro-forming procedure is required on pristine devices before executing SET/RESET programming operations. Figure 1(b) shows electro-forming, SET and RESET programming characteristics highlighting D2D (device-to-device) variability for the SiO_x RRAM device based 1T-1R array. Figure 1(c) shows SET/RESET switching characteristics highlighting C2C (cycle-to-cycle) variability. *Electroforming/SET switching*: The atomic SiO_x layer material has wide distribution of bond lengths and angles, thereby having large site-to-site variations and defects (such as Frenkel pair). On application of high electric fields, aforementioned defects decay into doubly charged oxygen interstitial (I_o) and doubly charged oxygen vacancy (V_o). V_o can carry an inelastic trap-to-trap tunneling current; as a result, a CF (conductive filament) is formed due to increased V_o density, and RRAM switches to a LRS (low resistance state). Since the trap-to-trap conduction mechanism is inelastic, heat is generated within oxide during SET and RESET switching, resulting in a positive feedback loop between the tunneling current and CF heating.[11] By limiting the electric field (using current compliance), this run-away effect is prevented across the oxide layer. *RESET switching*: When the negative bias is ramped up at the TE, the reverse current and heat generation increase, this activates the reverse mode of the interface reaction. The opposite polarity of an electric field pushes back the I_o into the bulk oxide (towards the TE), where they recombine with V_o to weak spots resulting in rupturing of the CF. As a result, RRAM switches to a HRS (high resistance state). Endurance characteristics and statistical distribution of LRS/HRS regions are shown in Fig. S1 of the supplementary material. Fabricated test chip to validate XNOR application consists of 8×8 1T-1R RRAM device arrays. Figure S2(a) of the supplementary material illustrates the schematic representation of the 8×8 RRAM array topology highlighting (i) *Wordlines (WLs)*: the gate terminal of transistors, (ii) *Bitlines (BLs)*: the source terminal of all transistors, and (iii) *Select lines (SLs)*: the TE of all devices. Note that the BL and SL run parallel to each other along a column on purpose. While other XNOR implementations in the literature[2,3,5] are array orientation sensitive (i.e., they can function only across rows or columns but not both within the same array), parallel SL and BL topology of the proposed XNOR IMC bitcell helps it in realizing both row-wise and column-wise functionality within the same array. To access a desired memory bitcell, we select row address by enabling corresponding WL and column address by selecting BL/SL. Since LRS and HRS resistance values are positive physical quantities, conventional RRAM devices cannot directly encode/represent negative weight values ("+1" and "−1") as required in BNNs. Hence, we have proposed the "2T-2R XNOR-RRAM" bitcell for realizing XNOR-Net based BNNs. The topology of the fabricated array provides support for two XNOR-RRAM bitcell configurations: (i) $XNOR_{row}$ and (ii) $XNOR_{col}$ (refer to Fig. 2 and Fig. S3 of the supplementary material). For both configurations, binary weights are mapped onto the HRS/LRS values of RRAM devices, whereas the binary activations are mapped onto the differential SLs (for $XNOR_{row}$) and WLs (for $XNOR_{col}$). Since two devices are utilized for representing each Logical weight (shown in Fig. 3 and Fig. S3 of the supplementary material), array utilization is reduced by 50%. In $XNOR_{row}$ implementation, one row is selected at a time by enabling the corresponding WL. The 2T-2R bitcells of that row are effectively realized in parallel through application of complementary signals (V_{read}, 0) by the SL decoder on pairs of consecutive columns as shown in Fig. 2(a). Summed-up current values (i.e., integrated Bitline current, $I_{BL,n}$) from the respective 2T-2R bitcells are passed to the I_{BL} input of the 1-bit CSA (current sense amplifier) in parallel. All n CSA blocks function as a comparator (comparing I_{BL} and I_{REF}) and generate binary voltage outputs (V_{DD},0), which are then fed to an analog adder (referred to as *Popcount block*). CSA functioning and choice of I_{REF} are detailed in Appendix S4 of the supplementary material. The output of the *Popcount block* is then compared against an input-specific

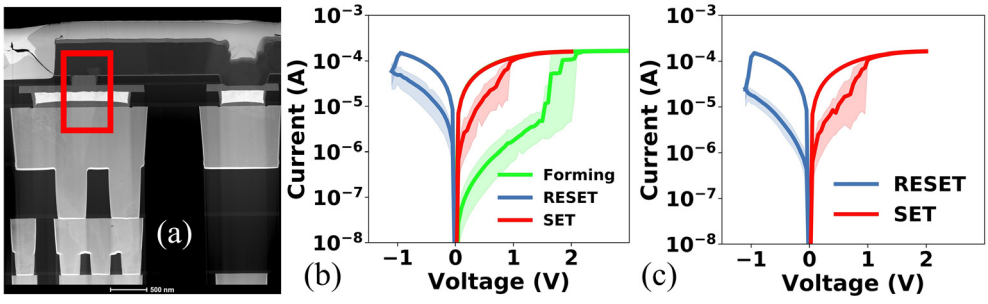

FIG. 1. (a) SEM cross section of the SiO_x RRAM cell integrated on top of the 130 nm CMOS, (b) IV characteristics showing electro-forming, SET and RESET operation highlighting D2D variability (20 devices), and (c) C2C variability during SET/RESET distribution over ten cycles.

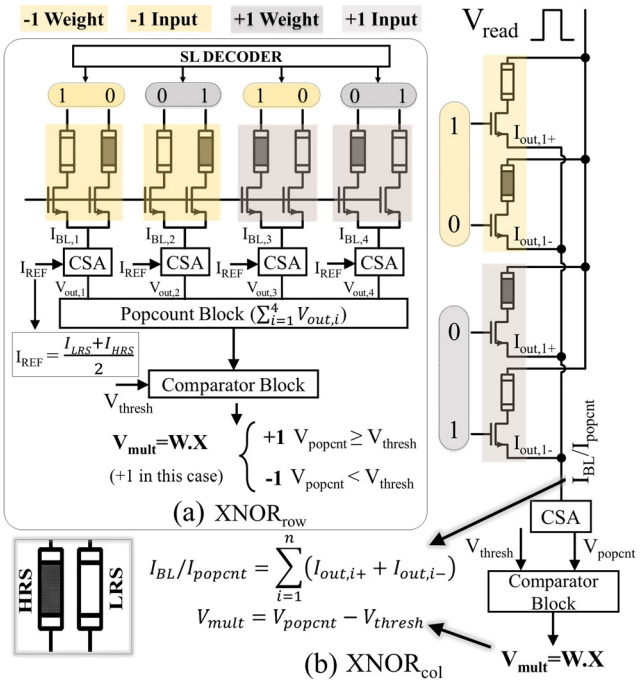

FIG. 2. (a) BNN computation mapping on $XNOR_{row}$ bitcells and its corresponding *popcount* implementation. The output current from the 2T-2R bitcell is converted to voltage using a CSA followed by summation in the *popcount block*. The *popcount* is then compared to a pre-fixed threshold to obtain the output of VVM. (b) BNN computation mapping using the $XNOR_{col}$ configuration and *popcount* is implemented inherently over the fabricated RRAM array.

threshold (V_{thresh}) to generate final output for the row-wise multiplication using a digital comparator. In $XNOR_{col}$ implementation, multiple WLs are activated in parallel to map inputs to all columns simultaneously. I_{BL} is sensed across the CSA in the column as shown in Fig. 2(b) (for details refer to Appendix S4 of the supplementary material). If the length of the input vector is greater than the number of columns (for $XNOR_{row}$)/rows (for $XNOR_{col}$), weight vectors corresponding to a neuron are partitioned and allocated on a separate set of rows/columns. To obtain the dot-product of the applied input vector and weights, "*popcount*" is computed over the XNOR outputs using analog circuits as shown in Fig. 2 and Fig. S3 of the supplementary material. The output of VVM (vector–vector multiplication) operations (i.e., $V_{out,j}$ across column **j** of the matrix) is defined in Eq. (1). The output current ($I_{i,j}$) of each XNOR cell is summed after post-processing using a CSA (represented by *F*). The summed output voltage is then compared with a threshold (V_{thresh}) based on the array size to obtain the final multiplication output. A single 8×8 1T-1R array can realize VMM (vector matrix multiplication) operations for a weight matrix of size 4×8. Input and weight vectors of width 4 can be utilized in either row or column configuration

$$V_{out,j} = \sum_{i=1}^{n} F(I_{i,j}) - V_{thresh}. \quad (1)$$

For weight matrices larger than the array size, *popcount* is realized using a multi-stage compute scheme where the XNOR bitcell

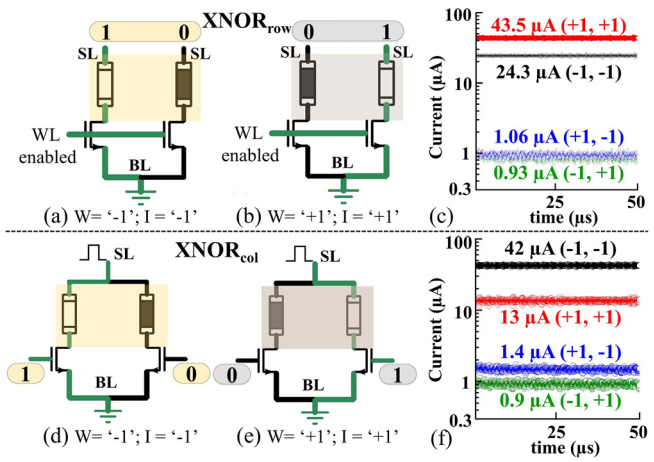

FIG. 3. Schematic representation/operand mapping corresponding to possible combinations of input activations ("−1," "+1") and weights ("−1," "+1") are shown for (a) and (b) $XNOR_{row}$ and (d) and (e) $XNOR_{col}$. (c) and (f) Experimentally characterized bitcell output current of four possible operand combinations for $XNOR_{row}$ and $XNOR_{col}$.

output is first amplified using the CSA. The partial sum output corresponding to the partition of input and weight is then computed and converted to a binary output. Binary outputs from all such partitions are summed up with a weighting scheme based on array utilization, i.e., the number of rows or columns occupied for the operation. A final binarized value is then generated corresponding to complete input and weight vectors. Summation performed at each row/column of the array is shown in Figs. S3(a) and S3(b) of the supplementary material. In the case of $XNOR_{col}$ implementation, all bitcells in the same column are computed in parallel [Fig. S3(b) of the supplementary material] by asserting all WLs simultaneously, thereby implementing binary MAC (multiply-accumulate) computation in a single step.

Experimental results corresponding to XNOR IMC bitcell configurations are shown in Fig. 3. Due to differential mapping of weights/activations, proposed bitcells have higher error tolerance for process variation.[7] Figures 3(a) and 3(b) illustrate all possible binary operand mappings for $XNOR_{row}$ bitcell operation. Input activations "1" and "0" mapped on the SL indicate $V_{read} = 0.2$ and 0 V, respectively. When both weight and input activation are of the same polarity, as per the mapping scheme shown in Figs. 3(a) and 3(b), I_{read} corresponding to the LRS device is sensed and thereby results in effective logic "+1" output. This satisfies XNOR bitcell criteria because of input activation/weight value combinations of "−1"/"−1" and "+1"/"+1"; the effective resistance range is the same. However, when weight and input activations have opposing polarities, I_{read} corresponding to the HRS device is sensed and thereby results in effective logic "−1" output. Experimental validation and current measurements for all four possible cases are presented in Fig. 3(c). The $XNOR_{row}$ approach has been widely studied in the literature due to its relative robustness to device variation.[3] Figures 3(d) and 3(e) depict the mapping strategy to realize XNOR operation along the column of the 1T-1R bitcell array. Input activations are mapped on transistor gates of two selected 1T-1R bitcells (for effectively realizing 2T-2R), storing binary weight in differential format. Input activation and binary weight mapping for $XNOR_{col}$

are shown in Figs. 3(d) and 3(e), and measured I_{read} outputs corresponding to four possible combinations are shown in Fig. 3(f). If $I_{read} \geq 10\ \mu A$ for a single bitcell, the output logic is "+1" indicating both binary weight/input activation have the same polarity. In $XNOR_{col}$ realization, *popcount* operation is inherently possible across the bit strings/column (since the SL is common) with only need for a comparator, whereas for $XNOR_{row}$ this requires additional circuitry (refer to Fig. 2).

Classification accuracy results for simulations performed using (i) VWW and (ii) CIFAR-10 datasets are summarized in Table I. It can be observed that the proposed IMC bitcell-based realization of FracBNNs (refer to Appendix S7 of the supplementary material for more details) shows good learning performance (i.e., $\geq$ 80% inference accuracy) for both datasets. Inference accuracy refers to the accuracy estimated over the test split of the dataset (i.e., examples from the dataset not observed during the training phase). To estimate energy and latency of the proposed hardware, a matrix (MAT) size = 256 × 256 is selected to maximize parallel operations with $T_{read} = 10\ \mu s$. MAT ($x \times y$) implies a memory matrix of specified bitcells with "x" rows and "y" columns. Here, the MAT size refers to size of the 2T-2R array matrix, i.e., it would result in 1T-1R MAT size of 256 × 512 for $XNOR_{row}$ and 512 × 256 for $XNOR_{col}$, respectively. Corresponding operation energy of the MAT is estimated for both $XNOR_{row}$ and $XNOR_{col}$ configurations. XNOR operation mapping for the network is performed for each MAT based on which the total operation count is computed. This count is used to estimate the inference energy and latency. For the $XNOR_{row}$ configuration, the operations are performed row-wise in a sequential manner, whereas for the $XNOR_{col}$ configuration, all operations are performed simultaneously in the MAT. This leads to a higher delay for $XNOR_{row}$ bitcells as shown in Table I. However, it also presents an opportunity for energy savings in terms of operation counts for asymmetric block matrices that have sizes smaller than the total MAT size. Energy estimates reported in this study account for CMOS periphery (CSA and row decoders) and additional read operations performed on the MAT (refer to Appendix S8 of the supplementary material). During the inference operation, the same MAT is not reused for storing a separate set of weight values. Hence, no additional delay/energy cost owing to multiple write operations for the bitcell needs to be considered.

TABLE I. Performance of the trained BNN implemented using XNOR IMC bitcells. Performance parameters used for the simulation are MAT size: 256 × 256; T_{read}: 10 μs; V_{read}: 0.2 V.

Parameter	Configuration	VWW	CIFAR-10
Hardware platform		2T-2R XNOR IMC	
Model precision		Binary	
Weight memory		3.37 MB	34.39 kB
Training accuracy (%)		79.6	96.0
Inference accuracy (%)	$XNOR_{row}$	80.3	84.9
Inference energy (mJ)		0.87	0.095
Inference latency (s)		0.18	1.62
Inference accuracy (%)	$XNOR_{col}$	76.31	83.4
Inference energy (mJ)		1.31	0.095
Inference latency (ms)		9.54	0.69

For performance comparison of the two proposed XNOR IMC schemes, extensive simulations have been carried out based on RRAM LRS/HRS distributions [refer to Fig. S1(b) of the supplementary material]. For both $XNOR_{row}$ and $XNOR_{col}$ configurations, VMM operation is characterized using the 8 × 8 1T-1R array. Figures 4(a) and 4(b) illustrate the variability in CSA's output for different *popcount* outputs considering proposed XNOR IMC configurations. The $XNOR_{row}$ configuration provides reliable sensing margins [see Fig. 4(a)] as mostly resistance state distributions affects the sense_out$_{row}$. Hence, RRAM devices with higher MW (memory window-$= \frac{I_{LRS}}{I_{HRS}}$) will boost the system efficiency using the $XNOR_{row}$ IMC approach. On the other hand, the $XNOR_{col}$ approach offers better speed, but there is a larger trade-off in terms of accuracy as shown in Fig. 4(b). Since analog *popcount* computation is directly impacted by D2D RRAM variability observed along the column devices, there is a considerable overlap between sense_out$_{col}$ values for different *popcount* outputs. Furthermore, the impact of MAT sizes in terms of energy is analyzed for both XNOR IMC approaches to implement the CIFAR-10 workload. The results are shown in Figs. 4(c) and 4(d). It can be noticed that $XNOR_{row}$ offers better energy optimization when considering pure XNOR operation cost. However, the overall energy increases due to an increase in MAT accesses. As a result, $XNOR_{col}$ emerges as a better IMC scheme in terms of total energy costs.

Since RRAM devices may experience different potential error-inducing factors such as switching failures, sense/read failures, variability, stochasticity, device-ageing, and read/program-disturbs, a simulated BER analysis is also performed. Here, BER refers to the percentage of output bits that have errors relative to the total number of trials performed for a single XNOR gate realized using 2T-2R IMC bitcells. In Fig. 5(a), the overall BER has been simulated as a lumped parameter independent of any specific source of error to investigate

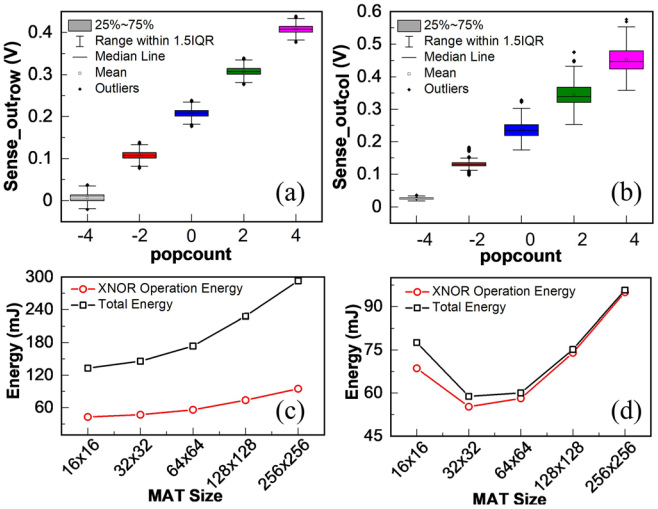

FIG. 4. Statistical distribution for VMM output variability for (a) $XNOR_{row}$ and (b) $XNOR_{col}$ configurations. The simulations are performed on the 8 × 8 1T-1R array with all input combinations applied for ≥1000 trials. Energy trade-off analysis based on MAT sizes for CIFAR-10 workload in terms of the XNOR operation energy and total energy (XNOR operations + CMOS periphery) for (c) $XNOR_{row}$ and (d) $XNOR_{col}$.

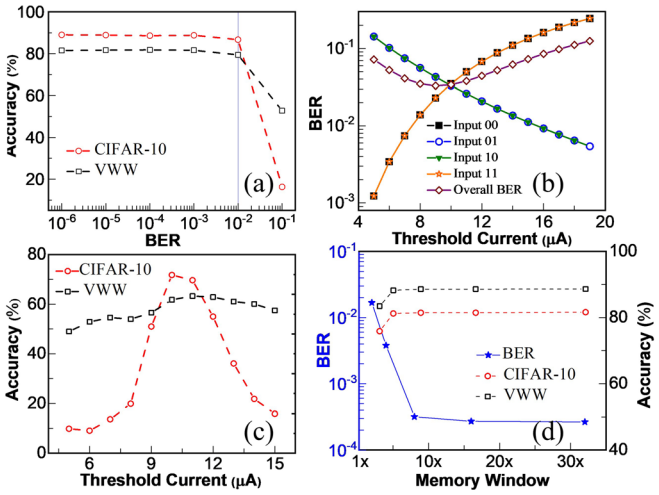

FIG. 5. (a) Impact of BER on BNN accuracy for VWW and CIFAR-10 workloads. For the $XNOR_{row}$ bitcell, the impact of $I_{sense,th}$ on (b) BER (for 1 million instances) and (c) BNN accuracy for VWW and CIFAR-10 workloads. (d) Impact of the MW on BER and inference accuracy (for VWW and CIFAR-10 workloads). All BNN accuracy simulations have been averaged over 10 trials and exhibit negligible variability (≈1%).

the robustness of the IMC based BNN networks. A clear roll-off in network accuracy is observed when the BER exceeded 10^{-2}. Accuracy loss in the case of VWW at higher BER is less (compared to CIFAR-10) as it is a binary classification problem. It is observed that even in cases where the RRAM devices are reliable and robust, the network performance may be impacted due to the nature of RRAM LRS/HRS distribution or the precision of the sense (read) circuitry. Using experimental LRS/HRS characterization [Fig. S1(b) of the supplementary material], statistical distribution parameters (mean, sigma) for the fabricated 1T-1R device array have been extracted. Multiple IMC inference simulations are performed using the extracted device array distribution parameters. Along with the resistance distributions, the impact of varying the sensing (read) threshold current [$I_{sense,th}$, labeled as I_{REF} in Fig. S3(c) of the supplementary material] has also been considered as shown in Figs. 5(b) and 5(c). A low $I_{sense,th}$ value indicates the memory threshold is defined closer to HRS, minimizing any LRS sensing error. Similarly, a higher $I_{sense,th}$ value indicates the memory threshold is defined closer to the LRS, minimizing the HRS sensing error. It is interesting to note that two different trends emerge depending upon the applied input combinations to the IMC bitcell. For input combination = "00"/"11," IMC bitcell's logic output = "+1," i.e., the LRS device should be read. Similarly, logic output = "−1" (i.e., the HRS device should be read) when input combination = "01"/"10." Thus, it becomes essential to select a sensing threshold that can minimize overall sensing error for accurate bitcell operation. Here, all input combinations are assumed to be equi-probable for an ideal workload, and the BER is characterized for the bitcell. An error minima can be observed for $I_{sense,th} = 10\ \mu A$ for the proposed IMC bitcell based on measured device array parameters. To analyze the impact of $I_{sense,th}$ on the BNN inference mode, two independent error models are implemented depending upon the output states, i.e., "+1" and "−1." Figure 5(c) presents the impact of $I_{sense,th}$ on network accuracy for both aforementioned datasets. The network accuracy trend for CIFAR-10 matches well with the error trend of the IMC bitcell shown in Fig. 5(b). However, for VWW, the network performance improves even when the error for "+1," i.e., error for reading an LRS device increases. It can be hypothesized that this effect occurs due to inherent sparsity of the network. Sparsity, in general, indicates data have higher count of "0"s. Here, in BNNs, it indicates majority of devices accessed during read operation are in the HRS. Clearly, using binary RRAM states (HRS/LRS) for IMC limits the adverse impact of variability that would otherwise reflect in analog VMM implementations. For the current implementation as shown in Fig. 5(c), the accuracy performance of the network is limited. An effective way of countering this is to increase the MW. The MW is a function of the statistical resistance distribution arising from the fabrication process related aspects, choice of active material stack, programming conditions, read voltage, sensing precision, device ageing, and other extrinsic factors such as temperature. Consequently, even with fixed programming conditions, the MW may vary or degrade with cycling. Analysis summarizing the impact of variability of the MW on BER and inference accuracy is shown in Fig. 5(d). In this analysis, the LRS is fixed at 20 kΩ and by sweeping the HRS from 42 to 800 kΩ, and the corresponding BER/inference accuracy is estimated using repeated simulations (≈1 million). As observed, the small MW leads to read/write errors, thereby resulting in higher BER in the IMC architecture. Using BER estimates, the outputs at each layer are computed, and further network accuracy (inference accuracy) is analyzed. It can be observed that MW ≈10 is sufficient for achieving reasonable learning performance (within 2% of maximum). A constant value of $I_{sense,th} = 10\ \mu A$ has been used for all simulations of the MW.

In summary, a state-of-the art BNN implementation using XNOR IMC based on SiO_x RRAM device arrays for both row and column configurations was experimentally demonstrated. The inference accuracy of ≈80.3% and ≈84.9% was obtained for VWW and CIFAR-10 workloads, respectively. For RRAM IMC MAT size = 256 × 256, per image inference energy (and latency) was estimated to be 1.3 mJ (9.54 ms) and 95 μJ (0.69 ms) for VWW and CIFAR-10 workloads, respectively, using the $XNOR_{col}$ scheme. Extensive analysis was performed comparing performance of $XNOR_{row}$ and $XNOR_{col}$ configurations in terms of energy, accuracy, and peripheral overhead. The impact of BER, RRAM device variability, and MW on the IMC based BNN network accuracy was also analyzed in detail.

See the supplementary material for details about the experimental setup, mapping strategy for BNN applications on the fabricated RRAM array, training of the network used as workload, energy estimation methodology for RRAM based IMC, and benchmarking of the current work with other literature NVM based XNOR IMC.

This work was supported in part by SERB-CRG/2018/001901, IITD-MFIRP grant, and CYRAN AI Solutions.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

S.K.K. and V.P. contributed equally to this work.

DATA AVAILABILITY

The data that support the findings of this study are openly available in the following links: (i) https://www.cs.toronto.edu/~kriz/cifar.html and (ii) https://cocodataset.org.

REFERENCES

[1]H.-T. Lue, P.-K. Hsu, M.-L. Wei, T.-H. Yeh, P.-Y. Du, W.-C. Chen, K.-C. Wang, and C.-Y. Lu, in *2019 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2019), pp. 38.1.1–38.1.4.

[2]W. H. Choi, P.-F. Chiu, W. Ma, G. Hemink, T. T. Hoang, M. Lueker-Boden, and Z. Bandic, in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2020), pp. 1–5.

[3]M. Bocquet, T. Hirztlin, J.-O. Klein, E. Nowak, E. Vianello, J.-M. Portal, and D. Querlioz, in *2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2018), pp. 20.6.1–20.6.4.

[4]P. Huang, Z. Zhou, Y. Zhang, Y. Xiang, R. Han, L. Liu, X. Liu, and J. Kang, APL Mater. **7**, 081105 (2019).

[5]S. Yin, X. Sun, S. Yu, and J.-S. Seo, IEEE Trans. Electron Devices **67**, 4185 (2020).

[6]A. Sebastian, M. L. Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, Nat. Nanotechnol. **15**, 529 (2020).

[7]T. Hirtzlin, B. Penkovsky, J.-O. Klein, N. Locatelli, A. F. Vincent, M. Bocquet, J.-M. Portal, and D. Querlioz, in *2019 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)* (IEEE, 2019), pp. 1–5.

[8]Y. Zhang, J. Pan, X. Liu, H. Chen, D. Chen, and Z. Zhang, in *The 2021 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, FPGA '21* (Association for Computing Machinery, New York, 2021), pp. 171–182.

[9]A. Chowdhery, P. Warden, J. Shlens, A. Howard, and R. Rhodes, arXiv preprint arXiv:1906.05721 (2019).

[10]A. Krizhevsky, V. Nair, and G. Hinton, see http://www.cs.toronto.edu/kriz/cifar.html (2014) for details regarding the CIFAR-10 dataset.

[11]W. Goes, D. Green, P. Blaise, G. Piccolboni, A. Bricalli, A. Regev, G. Molas, and J.-F. Nodin, in *2021 IEEE International Memory Workshop (IMW)* (IEEE, 2021), pp. 1–4.