Fully-Binarized, Parallel, RRAM-based Computing Primitive for In-Memory Similarity Search

Authors: Sandeep Kaur Kingra, Vivek Parmar, Deepak Verma, Alessandro Bricalli, Giuseppe Piccolboni, Gabriel Molas, Amir Regev, Manan Suri
Publisher: IEEE Transactions on Circuits and Systems II: Express Briefs (Early Access)
Date: 16 September 2022
Copyright © 2022, IEEE
https://ieeexplore.ieee.org/document/9893900

*IEEE Author Posting Policy:

IEEE seeks to maximize the rights of its authors and their employers to post the author-submitted, peer-reviewed, and accepted manuscript of an article on the author’s personal web site or on a server operated by the author’s employer. Additionally, IEEE allows its authors to follow mandates of agencies that fund the author’s research by posting author-submitted, peer-reviewed, and accepted manuscript of their articles in the agencies’ publicly accessible repositories. No third party (other than authors and employers) may post IEEE-copyrighted material without obtaining the necessary licenses or permissions from the IEEE Intellectual Property Rights Office or other authorized representatives of the IEEE.
Fully-Binarized, Parallel, RRAM-based Computing Primitive for In-Memory Similarity Search

Sandeep Kaur Kingra1, Vivek Parmar1, Deepak Verma1, Alessandro Bricalli2, Giuseppe Piccolboni2, Gabriel Molas2, Amir Regev2, and Manan Suri††

Abstract—In this work, we propose a fully-binarized XOR-based IMSS (In-Memory Similarity Search) using RRAM (Resistive Random Access Memory) arrays. XOR (Exclusive OR) operation is realized using 2T-2R bitcels arranged along the column in an array. This enables simultaneous match operation across multiple stored data vectors by performing analog column-wise XOR operation and summation to compute HD (Hamming Distance). The proposed scheme is experimentally validated on fabricated RRAM arrays. Full-system validation is performed through SPICE simulations using open source Skywater 130 nm CMOS PDK demonstrating energy of 17 fJ per XOR operation using the proposed bitcell with a full-system power dissipation of 145 μW. Using projected estimations at advanced nodes (28 nm) energy savings of $\approx 1.5 \times$ compared to the state-of-the-art can be observed for a fixed workload. Application-level validation is performed on HSI (Hyper-Spectral Image) pixel classification task using the Salinas dataset demonstrating an accuracy of 91%.

Index Terms—RRAM, In-Memory Computing, Similarity Search, Edge-AI, Low-power computing

I. INTRODUCTION

Associative memories (or CAM - Content Addressable Memory) are an important component of intelligent systems that can perform fast search operations [1]. CAMs accept a query and perform search over multiple data points stored in memory to find one or more matches based on a distance metric and return locations of matches. This information can be potentially used for applications such as nearest neighbour searches for classification or unsupervised labeling [2]–[5]. One of the basic distance metrics that can be used for such applications is HD (Hamming Distance) [2]. For any pair of strings or words of equal length, HD is defined as the total number of positions where the symbols/characters of the pair differ from each other. Conventional CAMs are designed using standard memory technologies such as SRAM (16T [6], 9T NOR-based and 10T NAND-based bitcels [7]) or DRAM [8], [9]. However, such volatile memory-based circuits have performance limitations that can be potentially addressed by using emerging NVM (Non-Volatile Memory) devices [10]–[12]. Use of NVM devices provides additional design flexibility by reducing circuit complexity and providing opportunity to exploit low-area analog IMC (In-Memory Computing) [10], [13]. Associative memory architectures that exploit NVM based IMC have been recently demonstrated using RRAM (Resistive RAM) devices based on XNOR [1], [5], [14]–[17] and XOR [3], [18] logic using HD as the distance metric. In this work, an end-to-end scheme is proposed to realize IMSS (In-Memory Similarity Search) in hardware by using RRAM devices and binarizing data and queries through a custom pre-processing pipeline. XOR gate functionality is realized using 2T-2R RRAM circuits where one input is encoded in form of RRAM device conductance states and the other input is applied as voltage signals. Reasons for adopting the 2T-2R bitcell for the study are: (i) Improved tolerance for D2D (Device-to-Device)/C2C (Cycle-to-Cycle) variability compared to 1T-1R resulting in reliable operations even at relaxed programming conditions potentially increasing endurance [19]; (ii) Improved resilience to impact of resistance drift or read-disturbs for the programmed states and (iii) High signal margins easing the sensing requirements [17]. In an array structure when such circuits are arranged in column-orientation, QI (Query Input) can be simultaneously applied to multiple columns of SD (Stored Data). Key contributions of the work are: (i) Experimental validation of 2T-2R XOR bitcell operation on fabricated 8×8 1T-1R RRAM arrays, (ii) Scheme for performing fully-binarized XOR-based IMSS through analog computation of HD, (iii) Validation of IMSS peripheral circuits through detailed SPICE simulations using 130 nm Skywater PDK, (iv) Validation of proposed methodologies on HSI (Hyper-Spectral Image) classification application with
Section VI provides concluding remarks.

on HSI pixel classification task using the Salinas dataset and the application level validation of the proposed IMSS architecture considering the periphery blocks. Section V presents

Section IV presents the SPICE simulation results for IMSS architecture. In addition to XOR circuit functionality, the proposed scheme also improves robustness to programming errors as opposed to a single 1T-1R bitcell [19] due to differential variability on classification accuracy.

Salinas dataset and (v) Analysis depicting impact of array size on sensing margin and variability on classification accuracy. The manuscript is organized as follows: Section II summarizes the RRAM device fabrication flow. Section III presents the experimental validation of proposed RRAM IMSS scheme. Section IV presents the SPICE simulation results for IMSS architecture considering the periphery blocks. Section V presents the application level validation of the proposed IMSS scheme on HSI pixel classification task using the Salinas dataset and finally Section VI provides concluding remarks.

II. FABRICATED RRAM ARRAY

The fabricated test chip used in our study is of size 16kb consisting of 256 8×8 1T-1R RRAM arrays [20]. The SEM cross-section of fabricated RRAM device integrated on top of 130 nm CMOS technology is shown in the inset of Fig. 1(a). The device stack has TiN as BE (Bottom Electrode), non-stoichiometric SiOx as switching layer and TiN as TE (Top Electrode). The 1T-1R bitcell occupies 30F2 on-chip area. Fig. 1(a,b) shows the electro-forming characteristics (where an initial conductive filament is formed) and cumulative distribution for pristine- and post-forming device resistance (in kΩ).

Fig. 1(c) shows SET and RESET switching characteristics highlighting D2D variability for the 8×8 RRAM device array. LRS (Low Resistance State) and HRS (High Resistance State) device resistance distributions are shown in Fig. 1(d). It is observed that the LRS ranges from 3 kΩ to 20 kΩ and HRS ranges from 110 kΩ to 1 MΩ.

H(D(SD, QI)) = \sum_{i=0}^{n} QI[i] \oplus SD[i] (1)

Eq. (1).

Fig. 3(a) presents the truth table summarizing 2T-2R XOR bitcell operation, (b) Custom designed PCB for XOR IMC validation, (c) Experimental validation for all possible input QI combinations of 2T-2R XOR bitcell.
storage. Fig. 3(b) shows the custom experimental setup and RRAM test chip used in the study. Programming signals are applied using high speed pulse measurement unit (Keithley 4225-PMU) from semiconductor parameter analyzer (Keithley 4200-SPA). The signals from PMU channels are multiplexed and applied to different signal lines (WL,SL,BL) using the custom switch board. Experimental validation of a single 2T-2R bitcell based XOR gate for all possible input combinations across 32 2T-2R XOR bitcells is presented in Fig. 3(c). Average $I_{\text{read}}$ (Read Current) is found to be less than 1 $\mu$A (i.e. RRAM device in HRS gets selected) when QI matches SD. When QI doesn’t match SD i.e mismatch, average is $I_{\text{read}} \geq 6$ $\mu$A (i.e. RRAM device in LRS gets selected). A reliable sense-margin $\geq 5$ $\mu$A is realized between match and mismatch states. For a given QI vector, the current will be the lowest through BL storing match/nearest match data vector. To perform IMSS, $V_{\text{read}} = 0.2$ V is applied on SLs with SD vectors to be compared against applied QI vector. HD is computed by applying an $n$-bit long binary QI vector and comparing it against $n$-bit binary SD vectors; at each SL, there are $n$ 2T-2R XOR gates participating. Bit level comparison is carried out by each XOR gate and summed up $I_{\text{BL}}$ (Bit-line Current) from n XOR gates is sensed on corresponding BL. BL/SL with maximum match bits will result in the lowest current. $I_{\text{BL}}$ is converted to $V_{\text{out}}$ (voltage sensed by SA (Sense Amplifier)).

### IV. SPICE SIMULATIONS

To validate the proposed IMSS architecture, SPICE simulations using the Opensource Skywater 130 nm PDK [21] are performed with ngspice. Simulations included RRAM arrays (modeled through resistance matrices) and periphery circuits shown in Fig. 2 and Fig. 4(a). Decoder-circuit/logic blocks and SA-circuits are designed with $V_{\text{DD}}=1.4$ V and 1.8 V respectively. Experimentally measured resistance values of LRS/HRS are used for the 1T-1R array during simulations. Transfer curve for the SA used in the simulation analysis is shown in Fig. 4(c). Gain of SA is selected based on length of the column vector so that full-mismatch leads to $V_{\text{out}} \approx V_{\text{DD}}$ and full-match leads to $V_{\text{out}} \approx 0$ V. $V_{\text{out}}$ as a function of HD, between a 4-bit QI vector and a 4-bit SD vector is shown in Fig. 4(d). Timing waveform showcasing all possible QI vector combinations and SA outputs for a $4 \times 8$ 2T-2R array are shown in Fig. 4(e). $V_{\text{out}}$ is minimum when QI vector matches the SD vector. For instance, when applied QI is “0100”, $V_{\text{out}}$ for BL$_7$ is minimum because the SD along BL$_7$ is “0100”. After incorporating measured D2D variability (Fig. 1(b),d)) in the simulations, the proposed IMSS scheme shows no overlap between output voltage levels for neighbouring HD values thus demonstrating a reliable operation. To estimate energy dissipation, current waveforms for all supply voltages ($I_{\text{dd,dec}}, I_{\text{read}}, I_{\text{dd,SA}}$) over a set of 32 input combinations were obtained through simulations. The simulations utilized a clock of 50 MHz. Based on average current dissipation for all operations, average power was estimated. Total power dissipation was estimated to be 145 $\mu$W with energy cost per XOR operation of 17 fJ. A complete breakdown of power estimation is shown in Fig. 4(g) with methodology for estimation explained in Eq. (2)-(3). In this study, we have exclusively focused on array size $= 8 \times 8$ in order to align with experimental measurements. However, array size has a significant contribution in determining overall performance for the IMC arrays [20]. For a constant workload, with increase in array size latency would decrease due to reduction in number of operation. However the cost of a single operation increases due to periphery overhead. Additionally increasing array size may lead to limitations in terms of sensing margin (see Fig. 4(b)).

Algorithm 1: Proposed IMSS method based on bitwise XOR.

**Require:** Query input vector QI, Stored Data SD

**Ensure:** Match Index $m$

**Pre-processing:**

$q_1 = \text{PCA}(QI) [0:20]$

$q_2 = \text{sign}(q_1) \times \log_{10}(|q_1|)$

$q_3 = \frac{q_1}{q_2}$

$q_4 = \frac{\max(q_3) - \min(q_3)}{\max(q_3) + \min(q_3)}$

$q_5 = \text{round}(q_4 \times 255)$

for $i=0; i<8; i++$ do

$q_d[i] = q_5 > (31 + 32 \times i)$

end for

**Similarity Search:**

for $k=0; k<\text{len}(\text{SD}); k++$ do

$d_s[k] = \text{popcount}(q_5 \oplus SD[k])$

end for

$m = \text{index}(\text{mode}(\text{topk}(-d_s)))$

**V. HSI CLASSIFICATION TASK**

HSI classification on the ‘Salinas’ dataset is used as an example usecase for the proposed IMSS architecture where every pixel in the image is classified to identify the type and age of the vegetation present on ground. Dataset includes HSI of size $512 \times 217$ with 224 bands acquired using AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) [24] sensor flying over Salinas Valley, California. RGB representation of the image is shown in Fig. 5(a). Ground truth comprises of 16 types of vegetation classes (see Fig. 5(b)). Design of HSI application-specific pre-processing pipeline becomes essential in order to obtain a simplified representation of the data without losing information. In case of HSI since both data precision and spectral resolution is high, compression becomes important. Custom pre-processing steps used for performing IMSS using Salinas dataset are summarized in Algorithm 1. First step involves performing PCA (Principal Component Analysis) [25] to extract relevant feature data, followed by log-scaling to compress data representation especially in case of large integers. To preserve the dynamic range, sign-multiplication is performed followed by normalization using two methods: mean-sigma followed by min-max. At the final step, 8-bit unsigned integer ($\text{uint8}$) representation is created. In the proposed HSI classification pipeline, use of only XOR operations is ensured to exploit the 2T-2R IMC array. Due to position specific weight assigned to each

\[ P_{\text{total}} = P_{\text{read}} + P_{\text{SA}} + P_{\text{dec}} \]  

\[ E_{\text{XOR}} = \frac{P_{\text{total}} \times T_{\text{read}}}{\text{Array size}} \]
Fig. 4. (a) Schematic of the simulated SA circuit using two-stage amplification. (b) Impact of array sizing on sensing margin. RBSTM is resistance-based sensing margin representing ratio in terms of R (in dB) for all-match and 1-mismatch case for the given block size. (c) SA transfer characteristic based on 130 nm Skywater PDK technology used for converting column-wise integrated current to voltage. (d) Simulations based SA outputs for different HD values using 4×8 2T-2R array. (e) Timing waveforms for QI vector (applied at WLs) and V_{out} for different BLs (storing binarized feature vectors) demonstrates successful match operation when QI vector matches with SD vector (extracted from SPICE simulations using 130 nm Skywater PDK). (f) Parameters used for SPICE simulations. (g) Estimated power values for each component.

Fig. 5. HSI classification results using proposed IMSS engine. (a) RGB data, (b) Ground truth of classification, (c) Prediction results using proposed IMSS engine. Decision space representation based on t-SNE for: (d) Ground truth and (e) Predicted results. (f) Accuracy comparison for Linear SVM (Support Vector Machine), proposed algorithm and HW-aware simulations. (g) Impact of variability on performance of proposed scheme with variability expressed in terms of ratio to mean for device state (LRS/HRS).

| Ref. | Device | Stack | Set Pulse | Reset Pulse | Read Pulse | E_{search} (pJ) | Tech. Node | Array | Application | Dataset |
|------|--------|-------|-----------|-------------|------------|----------------|------------|-------|--------------|---------|
| This work | SiOx | 2V, 1µs | -1.4V, 1µs | 0.2V, 20ns | 71.26 | 130 nm | 16 kB | Similarity Search | Salinas |
| [5] | HfOx | 3.7V, 1µs | -2.7V, 1µs | 0.1V, 10ns | 28.67 | 28 nm | | One-shot learning | Omniglot |
| [17] | MoS2 Tran. + HfOx | 2V, 0.2µs | -3.5V, 0.2µs | 30mV, 100ns | 42.76 | 40 nm | 64 kB | Similarity Search | NA |
| [23] | HfOx | 2V, 0.1µs | -2.5V, 0.1µs | 0.6V, 10ns | 44.46 | 130 nm | 4 kB | One-shot learning | Omniglot |
| [16] | FeFET | 4V, 10µs | 4V, 10µs | 1V, 1ns | 4.62 | 45 nm | 45 kB | One-shot learning | NA |
| [15] | PCM | 2.5V, 10ns | 1.2V, 1µns | 32.13 | 90 nm | 1 Mb | Similarity Search | NA |
| [14] | MRAM | 50µA | -150µA | 1.5V | NA | 140 nm | 9kB | Similarity Search | NA |

| Parameter | Value |
|-----------|-------|
| V_{diss} | 0.2V |
| V_{th,SA} | 1.0V |
| V_{th,⁑} | 1.4V |
| t_{ref} | 20 ns |
| Array size | 4×8 |

TABLE I

COMPARISON OF IMSS REALIZED USING STRUCTURES WITH 2T-2R CIRCUITS.
on the proposed differential encoding scheme, would require an RRAM IMSS chip (Fig. 2) of size ≈24MB. Since this requires a modest chip size for high-density RRAM, software simulations were performed using the PyTorch framework to validate the full Salinas classification application based on proposed RRAM IMSS. Pre-processed feature vectors from the test set are applied as QI to compute HD through XOR operations between SD vector and QI vector. Index of the SD vector with least distance is computed and class label from corresponding index is assigned to the test vector. Inference results using proposed IMSS scheme are shown in Fig. 5(c). To visualize learning performance in form of 2D-decision spaces, t-SNE (t-Distributed Stochastic Neighbour Embedding) [27] plots are generated for both ground-truth and predicted results as shown in Fig. 5(d) and Fig. 5(e) respectively. A comparison of the proposed HW-based approach against standard classification algorithms and software-based realization is shown in Fig. 5(f). For HW-aware simulations we incorporate the experimentally measured D2D variability of individual resistance states as shown in Fig. 1(d). To further demonstrate the resilience benefits of the 2T-2R structure, inference accuracies were estimated by sweeping the device variability for both HRS and LRS states simultaneously (see Fig. 5(g)). Table I presents comparison with other RRAM-based IMSS studies in literature. The proposed methodology enables simultaneous processing of 8 rows to determine HD as opposed to prior work where only 4 rows can be sensed in a single cycle of operation [5]. This however comes at the cost of increased energy dissipation from the SA circuit. When comparing device read energy costs for processing IMSS across 32 vectors of 128-bit, it can be observed that current implementation consumes ≈2× more energy compared to state-of-the-art array-based implementations. For the case of FeFET and MoS2 transistor-based realizations, the array structures have been realized in simulation and hence haven’t been considered for comparison. However, projected estimations using parameters from a 28 nm fabricated 1T-1R arrays [22] demonstrates energy savings of ≈2.5× compared to present implementation and ≈1.5× compared to SOTA IMC bitcells.

VI. CONCLUSION

Successful realization of XOR operations based on 2T-2R circuits using fabricated $SiO_x$ RRAM was experimentally validated. Analog computation of HD based on column-wise current integration to perform binarized IMSS was realized through extensive SPICE simulations including peripheral circuits using the Skywater 130 nm PDK. Simulations validated energy of 17 fJ per XOR operation with a full-system power dissipation of 145 $\mu$W for 8×8 RRAM array. Using projected estimations at advanced nodes (28 nm) energy savings of ≈1.5× compared to the state-of-the-art can be observed for a fixed workload. Proposed IMSS scheme was used for HSI pixel classification demonstrating 91% accuracy.

REFERENCES

[1] Y. Pan et al., “A ram-based associative memory cell,” in ISCAS 2021, Daegu, South Korea, May 22-28, 2021. IEEE, 2021, pp. 1–5.

[2] M. Imani et al., “A binary learning framework for hyperdimensional computing,” in DATE. IEEE, 2019, pp. 126–131.

[3] T. F. Wu et al., “Hyperdimensional computing exploiting carbon nanotube fets, resistive ram, and their monolithic 3d integration,” IEEE JSSC, vol. 53, no. 11, pp. 3183–3196, 2018.

[4] M. Imani et al., “Dual: Acceleration of clustering algorithms using digital-based processing in-memory,” in MICRO, 2020, pp. 356–371.

[5] H. Li et al., “Sapiens: A 64-kb ram-based non-volatile associative memory for one-shot learning and inference at the edge,” IEEE TED, pp. 1–7, 2021.

[6] N. Mohan et al., “Low-leakage storage cells for ternary content addressable memories,” IEEE TVLSI, vol. 17, no. 5, pp. 604–612, 2009.

[7] K. Pagiantzis et al., “Content-addressable memory (CAM) circuits and architectures: a tutorial and survey,” IEEE JSSC, vol. 41, no. 3, pp. 712–727, 2006.

[8] H. Noda et al., “A cost-efficient high-performance dynamic TCAM with pipelined hierarchical searching and shift redundancy architecture,” IEEE JSSC, vol. 40, no. 1, pp. 245–253, jan 2005.

[9] V. Lines et al., “60 MHz 2.3 m ternary dynamic content addressable memory,” in Records of the IEEE International Workshop on Memory Technology, Design and Testing. IEEE Comput. Soc., 2000.

[10] R. Karam et al., “Emerging trends in design and applications of memory-based computing and content-addressable memories,” Proceedings of the IEEE, vol. 103, no. 8, pp. 1311–1330, 2015.

[11] X. Sun et al., “Xor-ram: A scalable and parallel resistive synaptic architecture for binary neural networks,” in DATE, 2018, pp. 1423–1428.

[12] S. Shrey et al., “Energy-efficient all-spin bnn using voltage-controlled spinorbit torque device for digit recognition,” IEEE TED, vol. 68, no. 1, pp. 385–392, 2019.

[13] D. Ielmimi et al., “In-memory computing with resistive switching devices,” Nature Electronics, vol. 1, no. 6, pp. 333–343, jun 2018.

[14] S. Matsunaga et al., “Implementation of a perpendicular mbit-based read-disturb-tolerant 2t-2r nonvolatile tcam based on a reversed current reading scheme,” in ASPDAC, 2012, pp. 475–476.

[15] J. Li et al., “1 mb 0.41 µm² 2t-2r cell nonvolatile tcam with two-bit encoding and clocked self-referenced sensing,” IEEE JSSC, vol. 49, no. 4, pp. 896–907, 2014.

[16] K. Ni et al., “Ferroelectric ternary content-addressable memory for one-shot learning,” Nature Electronics, vol. 2, no. 11, pp. 521–529, nov 2019.

[17] R. Yang et al., “Ternary content-addressable memory with MoS2 transistors for massively parallel data search,” Nature Electronics, vol. 2, no. 10, pp. 590–600, 2019.

[18] Y. Halawani et al., “Stateful memristor-based search architecture,” IEEE TVLSI, vol. 26, no. 12, pp. 2773–2780, 2018.

[19] M. Bocquet et al., “In-memory and error-immune differential ram implementation of binarized deep neural networks,” in IEDM, 2018, pp. 20.6.1–20.6.4.

[20] S. K. King et al., “Dual-configuration in-memory computing bitcells using siox rram for binary neural networks,” Applied Physics Letters, vol. 120, no. 3, p. 034102, 2022.

[21] “Skywater open source pdk,” https://github.com/google/skywater-pdk.

[22] L. Grenouillet et al., “16kbit 11t rram arrays embedded in 28nm fdsoi technology demonstrating low ber, high endurance, and compatibility with core logic transistors,” in IMW, 2021, pp. 1–4.

[23] D. B. Ly et al., “Novel 1t2t1t ram-based ternary content addressable memory for large scale pattern recognition,” in IEDM, 2019, pp. 35.5.1–35.5.4.

[24] G. Vane et al., “The airborne visible/infra red imaging spectrometer (aviris),” Remote sensing of environment, vol. 44, no. 2-3, pp. 127–143, 1993.

[25] S. Wold et al., “Principal component analysis,” Chemometrics and intelligent laboratory systems, vol. 2, no. 1-3, pp. 37–52, 1987.

[26] J. Buckman et al., “Thermometer encoding: One hot way to resist core logic transistors,” in ASPDAC, pp. 385–392, 2021.

[27] J. van der Maaten et al., “Visualizing data using t-sne,” JMLR, vol. 9, no. 66, pp. 2579–2605, 2008.