A Nonvolatile Compute-In-Memory Macro Using Voltage-Controlled MRAM and In-Situ Magnetic-to-Digital Converter

Vinod Kurian Jacob*, Jiyue Yang*, Haoran He, Puneet Gupta, Kang L Wang, Sudhakar Pamarti.
University of California, Los Angeles, CA. 90095.
(*authors with equal contribution)

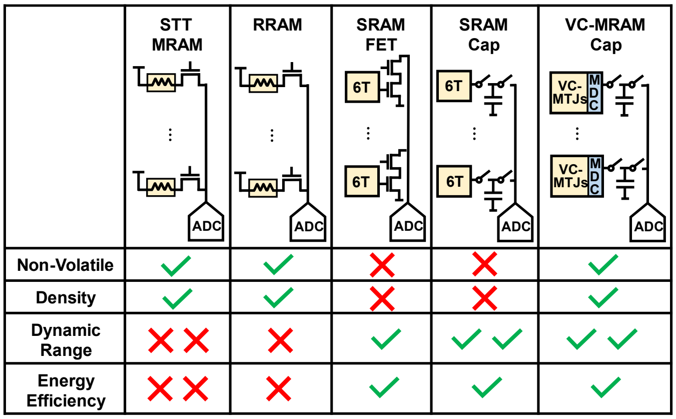
***Abstract*— Compute-in-Memory (CIM) accelerator has become a popular solution to achieve high energy efficiency for deep learning applications in edge devices. Recent works have demonstrated CIM macros using non-volatile memories (STT-MRAM, RRAM) to take advantages of their non-volatility and high density. However, effective computation dynamic range is far lower than their SRAM-CIM counterparts due to low device ON/OFF ratio. In this work, we combine a non-volatile memory based on a voltage-controlled magnetic tunneling junction (VC-MTJ) device, called voltage-controlled MRAM or VC-MRAM, and accurate switched-capacitor based CIM using a novel *in-situ* Magnetic-to-Digital Converter (MDC). The VC-MTJ device has demonstrated 10× lower write energy and switching time compared to STT-MRAM device and has comparable density, read energy and read latency. The *in-situ* MDCs embedded inside each VC-MRAM row convert magnetically stored weight information to CMOS logic levels and enable switched-capacitor based multiply–accumulate (MAC) operation with accuracy comparable to state-of-the-art SRAM-CIM. This paper describes the schematic and layout level design of a VC-MRAM CIM macro in 28nm. This is the first non-volatile CIM design to enable analog MAC computation with 256 parallel rows turned ON simultaneously without degradation in dynamic range (< 1 LSB). Detailed circuit simulations including experimentally validated VC-MTJ compact models show 1.5× higher energy efficiency and 2× higher density compared to state-of-the-art SRAM-based CIM.**

Index Term — Compute-in-Memory, Voltage-Control MRAM, Non-Volatile Memory, Deep Learning Accelerator.

I. Introduction

Deep learning algorithms have been widely used in computer vision, natural language processing and data analytics [1], [2]. Deep neural networks require many convolutional layers and a huge number of parameters learned from training data to achieve good inference accuracy. The latest image classification neural network has more than a hundred layers and several million weight parameters [3]. This poses a big challenge to the current computing architecture: model parameters and/or intermediate results must be moved, repeatedly, between off-chip memory and on-chip memory, or between the on-chip memory and processing elements. The energy cost and latency of data movement is much higher than the compute logic and can overwhelm the entire system's energy budget. As such, there is a huge demand for hardware accelerators that can process deep learning algorithms efficiently on edge devices.

Compute-in-Memory (CIM) is an emerging solution that reduces data movement by embedding the computing logic inside the memory. In typical CIM, the weight parameters are stored in the rows of a memory array, inputs are converted to analog voltage or pulse width-modulated signals and applied on the array's compute word lines, and dot-products between the weights and inputs are computed in current or charge domain by simultaneously enabling multiple rows; column ADCs digitize the result. The area and energy cost of the circuitry, especially that of the ADCs, is amortized by computing long dot products i.e., enabling as many rows as possible for the analog computation.


VC-MTJs
MDC
6T
ADC

	STT MRAM	RRAM	SRAM FET	SRAM Cap	VC-MRAM Cap
Non-Volatile	✓	✓	✗	✗	✓
Density	✓	✓	✗	✗	✓
Dynamic Range	✗✗	✗	✓	✓✓	✓✓
Energy Efficiency	✗✗	✗	✓	✓	✓

Fig. 1. Comparison between previous and proposed Compute-In-Memory (CIM) architectures.

Several works have demonstrated CIM-based deep learning processors or macros using SRAM that is available in the standard CMOS process [4], [5]. However, the number of parallelly enabled rows is limited by the large mismatch of the minimum sized transistors. The authors of [6] embed charge-based CIM using switched-capacitor circuits. A metal-oxide-metal (MOM) capacitor is added on top of each SRAM cell and, owing to its large size, provides good matching property that has been demonstrated to support over 1000 rows turning on at the same time without degrading the dynamic range. However, SRAM cells occupy a large area and limit the density of the macro.

In addition to high energy and area efficiency, ML accelerators in many edge devices also desire non-volatile storage of model parameters. Many embedded nonvolatile memory (eNVM) technologies such as Embedded Flash (eFlash), MRAM, and RRAM also offer higher storage density than SRAM. While eFlash technology has been popular in planar CMOS processes, it is increasingly difficult to scale in FinFET technology beyond 22nm [7]. MRAM technology has demonstrated better compatibility with advanced CMOS technology. It uses Magnetic Tunneling Junction (MTJ) as storage element that is fabricated between two interconnects in CMOS process backend. The 1T-1MTJ STT-MRAM cell has demonstrated more than 2 times higher density than SRAM [8].

There have been many recent efforts to embed compute logic in MRAM and RRAM to take advantage of their non-volatility and high density. However, only limited compute dynamic range is achieved. The authors of [9] and [10] demonstrated CIM macros using RRAM, however, they were only able to turn on 8 parallel rows simultaneously arguably due to the large variation of the RRAM resistances. The authors of [11] demonstrated current accumulation from 256 rows within the RRAM array during computation to achieve high energy efficiency. However, they employed accurate tuning of each RRAM device to meet the target resistance value during write operation. Although it helps reduce device mismatches, the calibration process requires off-chip equipment and takes a long time and large energy consumption. The device resistance also suffers from drifting over time.

The authors of [12] and [13] proposed CIM macros using STT-MRAM. However, the STT-MRAM's MTJs exhibit low resistances during both ON and OFF states, and a low ON/OFF ratio. The small MTJ resistance value causes large current consumption when many rows accumulate at the same time. The large current causes substantial IR drop caused by the wiring parasitic resistance, which makes the design challenging. The low ON/OFF ratio severely degrades achievable compute dynamic range in the presence of inevitable device mismatches. Note that the CIM macro in [12] achieves only 4 effective rows[1]. While [13] turns on 64 rows at the same time, achieved dynamic range is far less[2].

It is important to note that the primary challenge in realizing CIM in MRAM/RRAM is that of a low ON/OFF ratio. These technologies have an ON/OFF ratio less than 10 whereas each SRAM cell's read port can be completely turned off providing an ON/OFF ratio in the thousands. We show in the next section that the low ON/OFF ratio makes CIM very sensitive to device mismatch. Furthermore, prior MRAM CIM implementations have poor energy efficiency since the MTJs draw current for as long as they are being read.

In this work, we propose a robust and accurate nonvolatile compute-in-memory architecture using voltage-controlled MRAM (VC-MRAM) that addresses the aforementioned challenges. Our work has three main contributions:

1) We demonstrate the feasibility of using new Voltage Controlled MRAM technology for highly parallel, and accurate, in-memory computing. The core device, VC-MTJ has been demonstrated to achieve 10× lower write energy and smaller write time than STT-MRAM, thus making it a promising candidate for the next-generation MRAM technology [18]. Due to its large resistance (>10× than STT-MRAM), MTJ current is very low and VC-MRAM can achieve very low-power read/CIM operation. A detailed discussion follows in subsequent sections.

2) We propose compact magnetic-to-digital converters (MDCs) that can be embedded inside the VC-MRAM array to overcome the aforementioned challenges posed by a low MTJ ON/OFF ratio. The MDC is a single-ended 8T-1C offset-cancelling sense amplifier that translates information from magnetic domain to CMOS logic HIGH/LOW with high accuracy and is embedded "*in-situ*" or within each VC-MRAM row. Since stored bits are available as CMOS logic levels, high accuracy CIM such as capacitor based CIM is enabled with high parallelism (> 1000 rows) without degrading the signal dynamic range [6] i.e., the low ON/OFF ratio problem is effectively resolved.

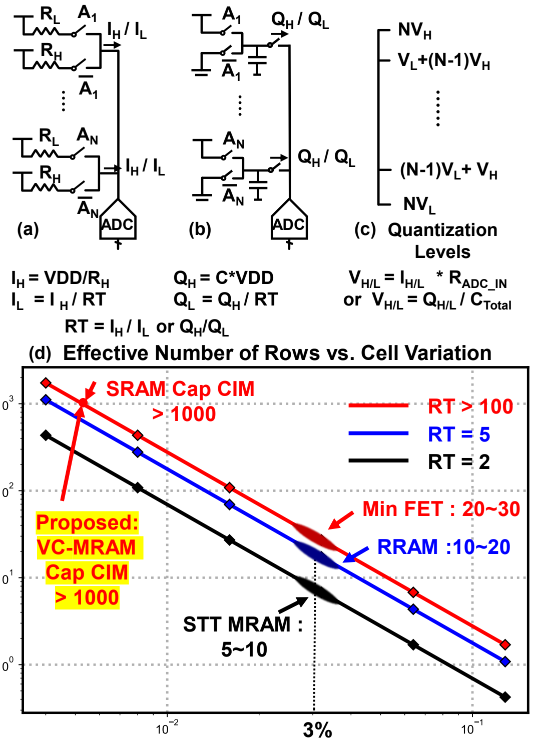


Fig. 2. Unified CIM circuit model for (a) SRAM/RRAM/MRAM (b) Switched- capacitor based compute.; (c) ADC Quantization levels.; (d) Effective number of rows vs. cell mismatch standard variation.

3) We propose a new "bit-serial weight" CIM macro architecture that improves reuse and further amortizes the MDC's energy and area overhead. Essentially, since the MDC generates CMOS logical levels, the stored weight information is reused over several compute lines that operate on the same weights.

This paper presents circuit and system level design of a non-volatile CIM macro and appropriate simulations that establish the feasibility and utility of the proposed approach. The simulations include an experimentally validated compact model of the VC-MTJ device. Section II introduces the challenges of compute in RRAM/MRAM due to large resistance variation and small ON/OFF ratio. Section III describes our proposed solution and Section IV discusses the results and performance evaluation.

II. Challenges of Compute in Non-Volatile Memory

As mentioned before, typically, Compute-in-Memory macros turn on as many rows as possible simultaneously to increase processing parallelism and cut down the area/energy overhead of the column circuitry such as ADCs. To understand the challenge posed by limited ON/OFF ratio consider the simplified in-memory compute circuit model shown in Fig. 2. Fig. 2.(a) corresponds to CIM scenarios that employ current summation. This is common in RRAM/MRAM based CIM and early SRAM based CIM. The model shows a differential

[1] As evident from the plot in Fig. 5 of [12].

[2] The statistical plot of the measured vs. ideal MAC result (Fig. 2.(c) in [13]) shows that ADC input has an effective variation of 7 (out of a maximum of 64) presumably due to MTJ resistance variation and/or mismatch i.e., the column ADC cannot distinguish MAC results less than 7 (out of a maximum of 64).

implementation where each weight is stored in two complementary cells, which is common in many such CIM solutions[3]. The LOW/HIGH resistance represents MTJ or FET resistance during ON/OFF state. Each cell will draw high current (I_H) or low current (I_L) from the shared bit line, BL, depending on multiplication results with input (A). The capacitor-based model of Fig. 2.(b) applies to switched-capacitor-based compute. The capacitor stores a weight as charge. During accumulation, charges on all the capacitors are shared.

Whichever model is considered, when N rows are turned on together, in the absence of device mismatches, the sum signal can be one of at most ($N+1$) possible levels. Typically, a column ADC is designed to reliably resolve these levels either in the current domain, or after converting into a proportional voltage, or time domain. Fig.2.(c) shows the quantization levels assuming current or charges are converted to voltage that has a $LSB = V_H\text{-}V_L$.

Invariably, mismatches between the cells in different rows degrade the effective resolution and the total variation increases with the number of rows. Assuming that the mismatches are independent zero-mean Gaussian random variables with a normalized[4] standard deviation of σ, the worst-case standard deviation of the MAC sum is $\sigma_{sum} = \sqrt{N}\sigma$. To reliably achieve no degradation of dynamic range, half the LSB should be greater than $3\sigma_{sum}$. It can be shown that

$$N \leq \left[\frac{1}{6\sigma}\left(1-\frac{1}{RT}\right)\right]^2$$

The effective number of rows is inversely proportional to the mismatch and the ON/OFF ratio, RT, as shown in Fig. 2.(d) which plots N vs σ for different values of RT.

Now, SRAM-based, RRAM-based, and MRAM-based CIM can be compared. STT-MRAM with Tunneling Magnetoresistance (TMR) ratio of 200% has been reported and corresponds to RT = 3 [14][5]. However, due to the larger resistance value of the access transistor compared to the MTJ resistance and MTJ resistance variations, the effective bit cell ON/OFF ratio is much lower. In fact, [15] claims that the tail bit in a large STT-MRAM array only has 20% TMR ratio. Since both the access transistor and the MTJ contribute to mismatch, a 3% σ-mismatch is optimistic and would limit the effective number of rows to 8. RRAM has a much higher ON/OFF ratio (5-10) compared to STT-MRAM but a 3% device mismatch would limit the number of rows to 20 during compute; state-of-the-art in RRAM based CIM has demonstrated 16 rows [9], [10]. SRAM-based CIM that sums FET currents, has >100 cell ON/OFF ratio but the mismatch in the minimum sized FETs can easily be up to 3-5% limiting operation to only about 8-32 simultaneously enabled rows. In contrast, SRAM-based CIM using large MOM capacitors, which owing to their relatively large size, achieve much better matching and can achieve more than 1000 parallel rows computation without reduction in dynamic range [6]. Note that the large MOM capacitor does not require additional die area since it is overlaid on top of the SRAM cell.

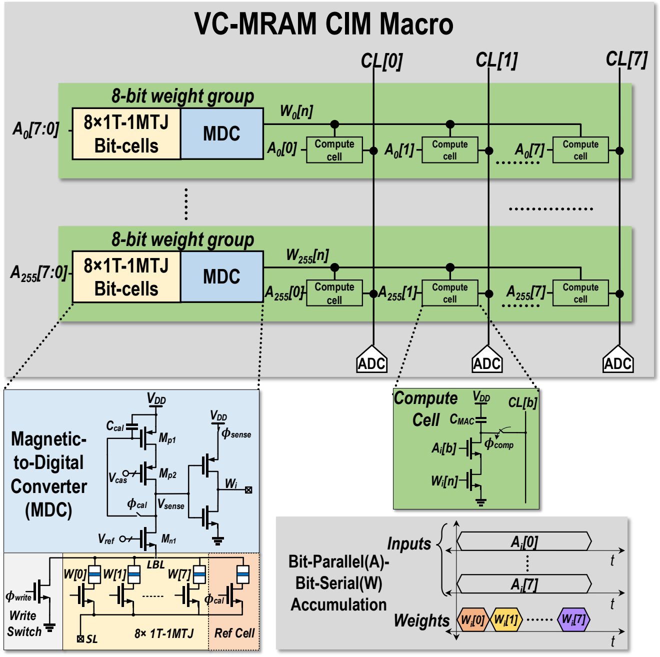


Fig. 3. Proposed VC-MRAM Compute-in-Memory Macro

Our proposed solution combines the benefits of the VC-MRAM technology, and the high dynamic range and parallelism of charge-based accumulation.

III. Proposed VC-MRAM CIM Solution

Fig. 3 shows the architecture of a VC-MRAM Compute-in-Memory macro. The VC-MRAM array is divided into 256 'weight-groups' and each weight-group contains 8 VC-MTJs, 1 MDC and 8 compute cells. A compact *in-situ* MDC converts the resistance state of a VC-MTJ cell into a CMOS logic HIGH/LOW (VDD/VSS) value. The resultant electrical bit is used for switched-capacitor based CIM on the compute line (CL). To amortize the area overhead of the MDC, it is shared between the 8 MTJs in a time-interleaved manner. Essentially, accurate charged based accumulation with bit-serial weights and bit-parallel activations is realized. The following subsections explain the details of our proposed solution.

A. Voltage-Controlled MRAM

Voltage-Controlled Magnetic Tunnel Junction (VC-MTJ) is a magnetic storage device that uses two magnetic layers sandwiching an oxide tunneling barrier. A simplified stack structure of the VC-MTJ is shown in Fig. 4. The parallel/anti-parallel state exhibits different resistance value (R_P/R_{AP}) and the

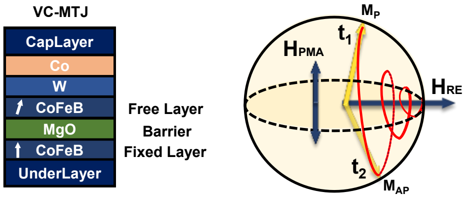


Fig. 4. VC-MTJ stack; Precessional switching operation.

[3] The model can be extended to single-ended versions with modifications, but it is not discussed here for the sake of brevity.

[4] I_H becomes $(1+\varepsilon)I_H$ and $I_L = I_H/RT$. This is reasonable and could be caused by a device size mismatch, for example.

[5] TMR ratio = $(R_H\text{-}R_L)/R_L = RT - 1$.

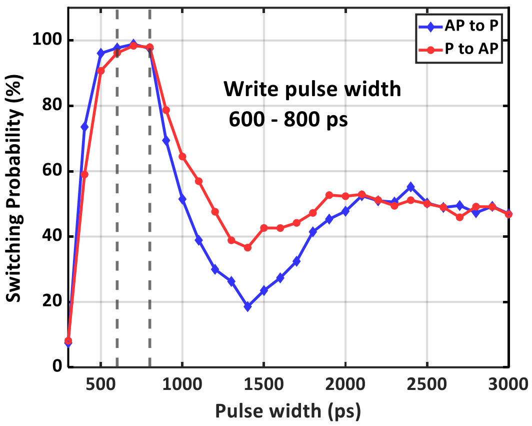


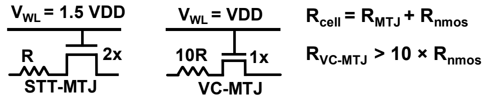
Fig.5. VC-MTJ switching probability vs pulse width

TMR ratio is expressed as $(R_{AP} – R_P)/R_P$. The VC-MTJ device is similar to STT-MRAM but has a thicker MgO layer and the write operation is based on Voltage Control of Magnetism (VCM). The mechanism of the VCM is due to the modulation of the charge carrier density by an applied electric field, which has an impact on the magnetic properties [16]. The thicker MgO barrier leads to much higher resistance and lower current than STT-MRAM. During switching, the applied voltage eliminates the Perpendicular Magnetic Anisotropy field (H_{PMA}) and the free layer starts to precess around the in-plane reference field, which is provided by the stray field from the *in-situ* reference layer [17]. The free layer's magnetic moment starts at t_1 and then precesses to the opposite direction noted as t_2 in Fig. 4. If the voltage pulse is removed when the magnetic moment reaches t_2, the moment becomes stable as H_{PMA} is recovered.

The switching of the VC-MTJ is ultrafast and typically less than 1 nsec [18], which is 10× faster than STT-MRAM and RRAM. Fig. 5 plots the experimental switching probability vs write pulse width. The VCM-based switching only changes the VC-MTJ into the opposite state, and a read-verify-write procedure is needed to write the same state. The write voltage is inverse proportional to the VCMA coefficient. 0.8 V write voltage and 115 fJ/V*m VCMA coefficient are measured in the literature [19]. Due to the low write voltage and fast speed, the VCM write operation consumes very low power comparable to STT-MRAM. A summary of the VC-MTJ is provided in Fig. 6.

The larger resistance of VC-MTJ helps the readout circuitry achieve lower power consumption and smaller area. A resistance-area product (RA) of 600 Ω·µm^2 is shown and leads to parallel resistances around 100 KΩ [18]. The total resistance of the 1T-1MTJ cell is the sum of MTJ and access transistor resistances. The typical TMR ratio in VC-MTJ (100%-200%) is comparable to STT-MRAM. Unlike STT-MRAM, VC-MTJ's resistance is 10× higher than the access transistor, therefore achieving higher effective cell TMR ratio. During read operation, a small voltage in opposite polarity to the write pulse is applied that can enhance device stability and lower disturbance rate [20]. Fig. 6 shows a summary of the comparison between VC-MRAM and STT-MRAM.

The proposed VC-MRAM CIM macro includes peripheral circuitry which supports normal memory read/write operations. To write the VC-MTJ device, a voltage pulse is applied on the source line (SL) shared along the column and the local bit line (LBL) is shorted to ground through the write switch as shown



$R_{cell} = R_{MTJ} + R_{nmos}$

$R_{VC\text{-}MTJ} > 10 \times R_{nmos}$

VCMA Coefficient	115 fJ/V*m
RA	600 Ω·µm^2
R_p	100KΩ
TMR	150%
Write Voltage	0.8 V
Write Duration	0.7n sec

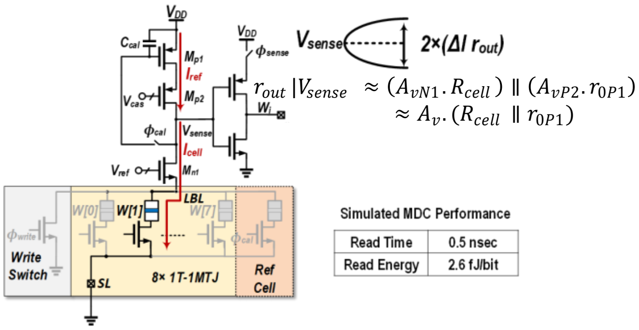
	STT MTJ	VC MTJ
Density	1×	2×
R_{MTJ}	1×	10×
Cell TMR Ratio	1×	1.6×
Write Time	1×	0.1×
Write Energy	1x	0.1x

Fig. 6. MTJ cell access transistor sizing; Summary of the VC-MTJ's specifications; Comparison between STT and VC-MTJ.

in Fig. 3. To read the VC-MTJ device the appropriate wordline is asserted and the MDC within the selected weight-group converts the VC-MTJ state into an electrical bit as described in the following subsection. ϕ_{sense}, ϕ_{comp} and the corresponding input line are set HIGH. If the MDC decision is a '1', the compute-cell will discharge the compute line (CL) and the decision is read out through the column peripheral circuit. The focus of this work is the in-memory compute operation in VC-MRAM CIM macro, so the details of the normal read/write operation are not further elaborated.

B. *Compact In-Situ Magnetic-to-Digital Converter*

As mentioned before, embedding the MDC inside each CIM row allows accurate capacitor based CIM. Such an MDC needs to be very compact and consume low power. However, since the MDC is essentially a sense-amplifier, large devices may be needed in principle, to keep V_{th} mismatches and other offsets to a minimum. Fig. 7 shows the proposed 8T-1C implementation of the MDC based on a local offset cancellation scheme, thereby eliminating large devices. In the sensing phase (ϕ_{sense}), V_{ref} sets the read voltage at LBL to about 200 mV nominally and this generates the cell current $I_{cell} = I_P$ or I_{AP} depending on whether the MTJ is in P or AP state. The wordline is asserted and M_{p1} and M_{p2} form a cascoded current source that pushes a reference current $I_{ref} = (I_P + I_{AP})/2$ into the sense node V_{sense}. Note that VC-MRAM, by virtue of its higher RA, exhibits comparable cell resistance to the r_o of minimum sized FETs in saturation, leading to a large cascode impedance of several 100 KΩ at V_{sense}. This translates to a large trans-resistance gain at V_{sense}. I_{ref} and I_{cell} are compared at V_{sense} and a large voltage swing of ~400 mV is obtained. The swing at V_{sense} far exceeds the variation range of the decision threshold of a subsequent gain stage, eliminating the need for a precise second stage. A



Read Time	0.5 nsec
Read Energy	2.6 fJ/bit

Fig. 7. MDC operation in sensing phase (ϕ_{sense})

simple minimum sized inverter suffices as the second stage to drive the decision to full rail logic levels. Furthermore, the lower read currents in VC-MRAM ensure that all devices remain in saturation during ϕ_{sense}, ensuring reliable operation in a low VDD of 0.8 V.

The accuracy of the first stage is key to achieving a low read-error-rate (RER). The effect of such errors on the compute accuracy is discussed in Section IV. Previous works [8],[21],[22] generate a precise reference current and mirror it to be compared with the cell current at V_{sense}. The V_{th} mismatch in the mirror transistors as well as the clamping transistors in the two distinct current paths lead to errors in the sense current ΔI. Mismatch is typically controlled by upsizing the current conducting devices. In contrast, the MDC generates I_{ref} locally (as described below) by sampling a corresponding V_{GS} on C_{cal} during ϕ_{cal}. The stored V_{GS} is reused in the sense phase (ϕ_{sense}) to compare with the cell current I_{cell}. Since I_{ref} and I_{cell} see the same current paths, the V_{th} variations of the FETs in the sense path do not generate an error current, leading to accurate read. The once sampled reference can be reused for several reads before recalibrating. Furthermore, cascode FET M_{p2} prevents coupling of swing on V_{sense} to C_{cal}, allowing better calibration reuse. A small C_{cal} of 0.5 fF is enough to allow re-use for 8 reads. This calibration scheme cancels the circuit offsets and allows to use minimum sized FETs for all devices in the compact 8T-1C MDC.

To generate an accurate I_{ref} over PVT corners, one extra VC-MTJ is added in each weight-group. Ideally, the reference VC-MTJ should present a conductance of $(G_P + G_{AP})/2$ during the calibration phase (ϕ_{cal}) to maximize the sense-margin. This is practically implemented as shown in Fig. 8 by combining two MTJs: one in P state and the other in AP state. The reference MTJs of adjacent weight-groups store these complementary states. The Local Bit Lines (LBL) of the adjacent weight-groups are connected by a switch controlled by ϕ_{cal}. During ϕ_{cal}, adjacent LBLs are shorted and the two complementary reference VC-MTJs are connected in parallel and present an equivalent conductance of $(G_P + G_{AP})$. V_{ref} sets the voltage across the VC-MTJs, and the current is provided by diode connected M_{p1} within the two identical MDCs. The two MDCs share the generated current and each effectively see $I_{ref} = (I_P + I_{AP})/2$. Since the reference VC-MTJ is inside the local array, it closely tracks the on-chip variation and provides a reference current that maximizes the sensing margin.

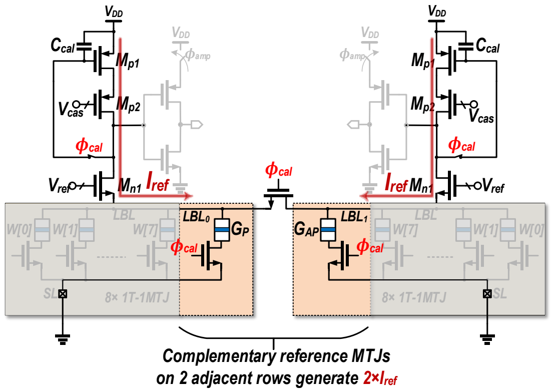


Fig.8. MDC during calibration phase (ϕ_{cal})

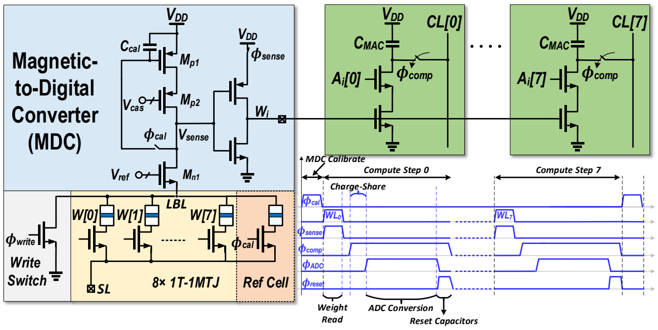


Fig.9. Switched capacitor-based compute and weight bit re-use strategy

C. Switch-Capacitor-Based Bit-Serial Bit-Parallel Compute

While the energy and area costs of the proposed MDC are small, we propose to amortize them further by reusing the MDC's output bit among multiple switched capacitor-based compute cells. Note that without the MDC, it is difficult to share/reuse the bit stored in the MTJ.

Fig. 9 shows the switched capacitor compute cells and the reuse strategy. Each read out weight bit is reused 8 times using 8 switched capacitor compute cells. Inside each compute cell, the weight bit is AND-ed to an input bit, and according to the result, a small capacitor which is pre-charged to VDD is either discharged or left alone. During a subsequent "compute" phase set by HIGH ϕ_{comp}, the capacitors of multiple parallel rows are connected to a shared compute line and CIM summation is achieved by charge sharing. Since advanced back-end processing in modern CMOS technology allows up to 0.8% mismatch for a 1.2 fF capacitance [23], high dynamic range of the summation is achieved.

Note that the input bits of the different compute cells are chosen to be the binary bits of 8b activations and are brought into the row in parallel. The MAC result on each compute line is digitized by column ADCs, binary weighted and added in the digital domain. This corresponds to a 1b Weight × 8b Input operation. The same sequence repeats for 8 compute cycles corresponding to each weight bit. The partial sum in each cycle is shifted and added in the digital domain to complete 8b Weight × 8b Input operation. The sequence described above essentially implements bit-serial weight bit-parallel input. It may seem that although only 1 out of 8 weight bits is used for compute in a cycle, leading to a lower throughput, the 8-way reuse of the weight bit effectively compensates for the apparent throughput loss. It is important to note that without the *in-situ* MDC, such weight reuse is not feasible in today's STT-MRAM or RRAM CIM solutions.

IV. Result and Discussion

The proposed VC-MRAM and capacitor-based CIM architecture is designed and evaluated in 28nm CMOS. The VC-MTJ compact model from [24] is used for evaluating performance. VC-MTJ resistance variation and transistor mismatch are considered in the accuracy evaluation. The 8b VC-MRAM CIM unit cell is laid out to show the feasibility of the physical design. The area and the throughput of the macro is also estimated for comparison.

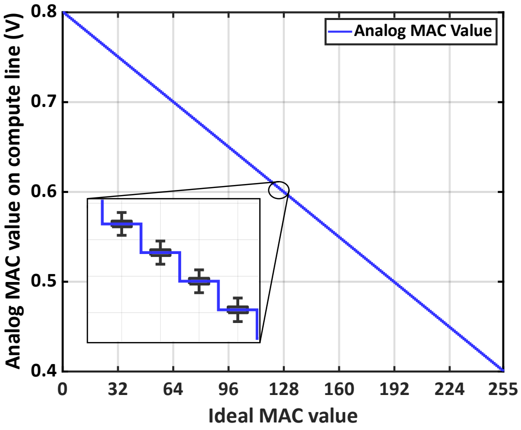


Fig. 10. Compute line voltage vs ideal MAC result.

A. Compute Accuracy

We demonstrate, in simulation, a 256 tall macro performing 256-way dot product accumulation using a 0.5 fF MOM capacitor in each compute cell. As described later, the MOM capacitor is overlaid on top of the compute cell's transistors. These 0.5 fF MOM capacitors have a mismatch standard deviation of 1.2%, which is still within the 3σ margin of 256-way i.e., 8-bit compute, as shown in Fig. 2. Note that using a larger 1.2 fF MOM reduces mismatch to 0.8% allowing a dynamic range up to 1000 rows but would result in larger compute cells.

In the compute phase (ϕ_{comp}), the compute line parasitic capacitance siphons away signal charge during charge redistribution among the MOM capacitors and leads to a gain error in the MAC v/s voltage characteristics as shown in Fig. 10. Every row adds 0.5 fF of parasitic capacitance on the compute line limiting the full-scale range to ½VDD. However, this doesn't impact compute accuracy as the resulting shrunk LSB is still an order of magnitude above the kT/C noise limit at the 8-bit level. Furthermore, this confirms that thermal noise does not limit the MAC SNR.

B. MDC Read Accuracy

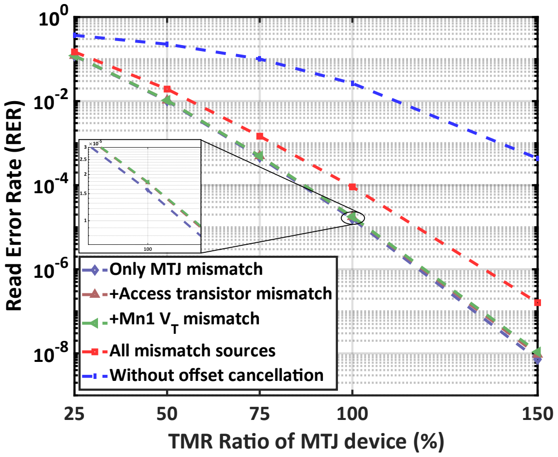


Fig. 11. Read error rate vs TMR Ratio.

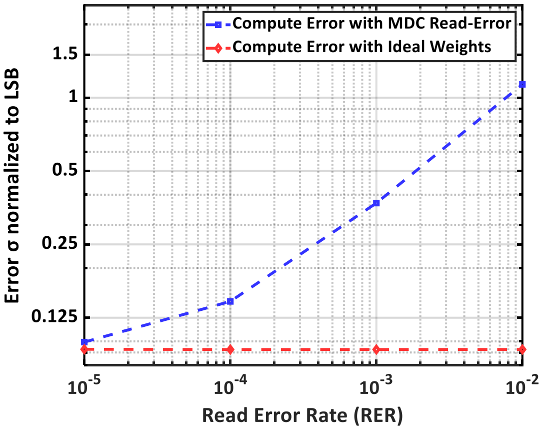


Fig. 12. Compute error vs MDC read error rate

To ensure that the compact MDC does not limit the MAC accuracy, we evaluated (a) the MDC's read error rate (RER) and, (b) the effect of MDC RER on MAC accuracy. To evaluate the former, we simulated the MDC's read error rate as a function of the VC-MTJ's TMR for a typical 5% mismatch with and without the proposed offset cancellation scheme. We observe that for a typical 100% cell TMR of VC-MTJ, the offset-cancelling sensing scheme discussed achieves better than 10^{-4} RER offering at least 2 orders of magnitude of improvement as shown in Fig. 11.

To evaluate the effect of read errors on the compute accuracy, we model each of the MOM capacitors *(C_k)* as a Gaussian distributed random variable with a mean of 0.5 fF and *(X_k)* and weight bits (W_k) are Bernoulli distributed random variables with equal chances for 0 and 1. In a charge-based accumulation scheme, this corresponds to the worst case for the compute error as pointed out in [25]. We further introduce random errors on the weight bits to model the MDC error rate. The weight bits with injected errors are represented as $\widetilde{W_k}$. The analog MAC value on the compute line is then evaluated as:

$$MAC = \frac{VDD\ (\sum_{k=0}^{255} C_k\, X_k.\widetilde{W_k})}{\sum_{k=0}^{255} C_k + C_{parasitic}}$$

This is then compared with the ideal MAC value to estimate the compute error. Fig. 12 shows the compute error standard deviation normalized to an LSB at the 8-bit level as a function of MDC read error rate, using 1-million-point monte-carlo sampling. We also observe that at the 10^{-4} RER level, we add an excess compute error standard deviation of just 5% LSB at the 8-bit level.

C. 8-bit VC-MTJ CIM unit layout

The 8-bit VC-MTJ weight-group along with the MDC and compute cells are laid out in 28nm CMOS technology to show the feasibility of physical design and allows us to evaluate the density of the proposed solution. The layout of the local array is 2.8 µm × 1.8 µm, shown in Fig. 13. There are 8 VC-MTJs within the weight-group in a 4×2 arrangement connected to the same local bit line (LBL) . The VC-MTJ is fabricated between M4 and M5. Every group of two access transistors share the same source diffusion to save area. The MDC is laid out on the top of the array, along with the reference MTJ and write switch. The output of the MDC is shared with 8 compute cells which are placed on the right side of the MTJ array, also in a 4×2 arrangement. A MOM capacitor is placed on top of each compute cell in M6 to M8 and does not occupy extra area.

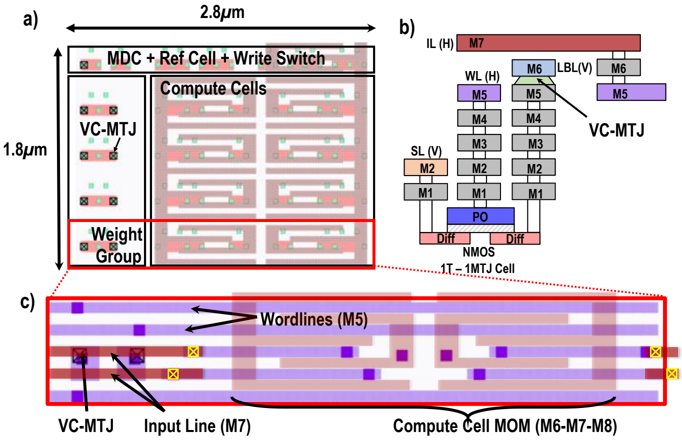


Fig. 13. a) Layout of the 8-bit VC-MTJ CIM unit cell. b) Cross-section in the MRAM region. c) Wordline and Input Line horizontal routing

As mentioned before, to maintain high compute density and throughput the bit-serial weight bit-parallel input scheme was implemented. This requires routing 8 parallel input lines (IL) and 8 wordlines (WL) horizontally within the compact local array vertical pitch. While the 16 lines can be readily accommodated in a single horizontal routing layer (M5 was chosen) within the 1.8 μm row height, the routing is non-trivial since the MTJs partially block M1 to M5 and the MOM capacitors above the compute cells completely block M6 to M8. So, the input lines use a bridge on M7 in the MRAM region as shown in Fig. 13(b), (c). The compute lines are routed vertically in M4. This arrangement frees up lower metals M2 and M3 for control signal routing and MDC output. LBL runs vertically on M6. VSS and source line (SL) runs vertically on M2 and VDD runs vertically on M6-M7-M8 on the shared MOM capacitor top-plate.

D. Energy Efficiency

The energy efficiency of the VC-MRAM CIM macro is evaluated at 0.8 V supply and estimated based on SPICE simulation that incorporates parasitic capacitances. The macro has 32 slices, each implementing a different weight filter. Each slice has 256 rows arranged vertically.

The MDC consumes 2.6 fJ per read operation on average. This read cost is amortized by a factor of 8 through the bit-serial weight and bit-parallel input scheme described in Section III. The input bits are applied from outside of the macro and will consume communication energy from the input buffer next to the macro. The bottom plate of the MOM capacitors in the compute cells are pre-charged to VDD during the reset phase and discharge to VSS if the multiplication result is 1. The compute cells consume energy from switching the NMOS-based multiplication circuit and charging the capacitors. A 50% switching activity factor is assumed for the input buffers and compute cells. Furthermore, the parasitic capacitance on the compute line consumes energy due to charging up to VDD and down to analog MAC value during the pre-charge and evaluation phase respectively. Each slice has 8 compute lines, and the analog voltage is converted to digital bits by 8 ADCs. The 6-bit SAR ADC accounts for 33% of the macro energy, as shown in Fig.14. The 6-bit ADC does not use the full precision of the MAC result, but previous works of capacitor-based CIM accelerator such as [26] have demonstrated negligible loss in network classification accuracy with respect to fixed point implementation for the same ratio between MAC and ADC precision. The macro energy consists of four main components: input buffer, *in-situ* sense amplifier (MDC), compute cell and column ADCs. Each slice consumes 2.02 pJ and corresponds to 32 TOPS/W for 8-bit operation.

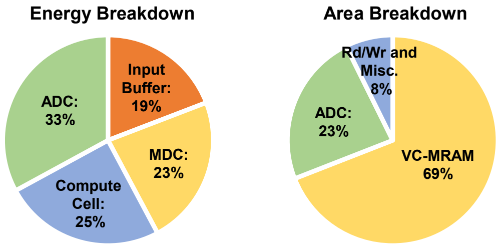


Fig. 14. Energy and area break-down of the macro.

E. Throughput and Area

The macro operates at 250 MHz. Before the compute starts, the *in-situ* MDC first reads one MTJ from the weight-group. Unlike conventional NV-MRAM memory access, the specific MDC implementation along with smaller weight group (8-cells) choice decouples bit-line capacitance from the read time constant, enabling fast read time of 0.5 nsec. The weight group's size can be increased to 16 or 32 to amortize the area/energy overhead of MDC further, but might lower the macro utilization ratio. The compute cells multiply and store the results in the capacitors. The 1-bit multiplication in the compute cell happens simultaneously with the MDC read. The charge sharing between capacitors on the same compute line takes just 0.2 nsec, since all the rows in the compute line participate in the charge sharing irrespective of the individual 1-bit multiply result and the time constant is of the order of just 20 psec. This contrasts with current summation CIM architectures, where it is possible that only 1 cell discharges the compute line in the worst case and could potentially limit the throughput. The 6-bit SAR ADC uses a 2 GHz clock to convert the analog MAC result on the compute line in 3.5 nsec. The macro performs 8-bit MAC operations with a throughput of 303 GOPS/mm^2.

F. Comparison with Other Works

Table 1. shows the comparison table with the other state-of-the-art works. Our proposed VC-MRAM CIM solution achieve 256 rows in parallel during compute, while not sacrificing the theoretical maximum signal dynamic range. The effective dynamic range in each CIM array is calculated as the ratio between maximum signal amplitude and worse case error (at a 3σ confidence interval) from the reported statistical measured v/s ideal MAC result plot. Although [12] achieves 128 rows in parallel, the effective dynamic range is only 12dB. Due to the capacitor-based compute, no active current is drawn during compute and VC-MRAM CIM solution achieve 1.5× – 30× higher energy efficiency compared to other non-volatile solutions. The throughput density is 4× – 30× higher than RRAM/STT-MRAM solutions because of the high parallelism. Compared to the SRAM capacitor based CIM solution [6], our solution achieves 2× higher energy efficiency and 1.4× higher throughput density, while having the benefits of non-volatility.

TABLE 1. COMPARISON WITH OTHER WORKS

	Nonvolatile CIM						SRAM CIM	
	This Work	[9] Hung ISSCC' 22	[10] Xue ISSCC' 21	[12] Deaville VLSI' 22	[13] Jung Nature' 21	[27] Chiu ISSCC'22	[6] Jia ISSCC' 21	[4] Yue ISSCC' 21
Technology	**VC-MRAM, 28nm**	ReRAM, 22nm	ReRAM, 22nm	STT-MRAM, 22nm	STT-MRAM, 28nm	STT MRAM, 22nm	SRAM, MOM Cap,16nm	SRAM, FET 65nm
CIM or Near Mem(NM)	**CIM**	CIM	CIM	CIM	CIM	NM	CIM	CIM
Parallel Rows	**256**	8	4	128	64	1	1152	16
Effective Dynamic Range (dB)	**48.2 dB**	18.1 dB	12dB	12dB	19.1dB	6dB	61.2dB	24.1dB
ADC Precision	**6b**	3b	2b	6b	4b	No ADC	8b	4b
Supply Voltage	**0.8**	0.8	0.8	0.8	0.9	0.8	0.8	1
Macro Energy Efficiency (8-bit, TOPS/W)	**32**	21.6	11.9	0.65	6.3	25.1	15*	14.6*
Throughput Density (8-bit, GOPS/mm^2)	**303**	11.4	11.6	85.6	69.2	14.4	222.3*	116*

* Scaled to 28nm technology

Effective dynamic range (dB) is calculated by maximumsignal / worst case noise ratio in the statistical measured vs ideal MAC result plot.

The macro in [6] achieves 1152 parallel rows with less than 1 LSB degradation of dynamic range due to the large MOM capacitor on top of the SRAM cell. The VC-MRAM CIM macro can also achieve more than 1000 parallel rows if the compute cell uses a larger MOM capacitor, but the density and energy efficiency will be sacrificed.

V. CONCLUSION

We have proposed a robust CIM architecture using a new Voltage-Controlled MRAM technology combining with the high-parallel capacitor-based compute. We have presented a compact *in-situ* Magnetic to Digital Converter (MDC) that is offset tolerant, compact and ultra-low power. The proposed solution is evaluated by simulation and shows much higher energy efficiency and throughput than other non-volatile memory based CIM solutions.

REFERENCES

[1] R. Girshick, "Fast R-CNN," *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440-1448

[2] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *Computing Research Repository (CoRR)*, Sept. 2016.

[3] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.

[4] J. Yue et al., "A 2.75-to-75.9TOPS/W Computing-in-Memory NN Processor Supporting Set-Associate Block-Wise Zero Skipping and Ping-Pong CIM with Simultaneous Computation and Weight Updating," *ISSCC*, 2021.

[5] Q. Dong et al., "A 351TOPS/W and 372.4GOPS Compute-in-Memory SRAM Macro in 7nm FinFET CMOS for Machine-Learning Applications," *ISSCC*, 2020.

[6] H. Jia et al., "A Programmable Neural-Network Inference Accelerator Based on Scalable In-Memory Computing," *ISSCC*, 2021.

[7] Y. -C. Shih *et al.*, "A Reflow-capable, Embedded 8Mb STT-MRAM Macro with 9nS Read Access Time in 16nm FinFET Logic CMOS Process," *2020 IEEE International Electron Devices Meeting (IEDM)*, 2020

[8] L. Wei *et al.*, "13.3 A 7Mb STT-MRAM in 22FFL FinFET Technology with 4ns Read Sensing Time at 0.9V Using Write-Verify-Write Scheme and Offset-Cancellation Sensing Technique," *ISSCC*, 2019.

[9] J. -M. Hung *et al.*, "An 8-Mb DC-Current-Free Binary-to-8b Precision ReRAM Nonvolatile Computing-in-Memory Macro using Time-Space-Readout with 1286.4-21.6TOPS/W for Edge-AI Devices," *ISSCC*, 2022.

[10] C. -X. Xue *et al.*, "16.1 A 22nm 4Mb 8b-Precision ReRAM Computing-in-Memory Macro with 11.91 to 195.7TOPS/W for Tiny AI Edge Devices," *ISSCC*, 2021.

[11] Wan, W., Kubendran, R., Schaefer, C. *et al.* A compute-in-memory chip based on resistive random-access memory. *Nature* 608, 504–512 (2022)

[12] P. Deaville, B. Zhang and N. Verma, "A 22nm 128-kb MRAM Row/Column-Parallel In-Memory Computing Macro with Memory-Resistance Boosting and Multi-Column ADC Readout," *2022 IEEE Symposium on VLSI Technology and Circuits*, 2022.

[13] Jung, S., Lee, H., Myung, S. *et al.* A crossbar array of magnetoresistive memory devices for in-memory computing. *Nature* 601, 211–216 (2022).

[14] V. B. Naik *et al.*, "Manufacturable 22nm FD-SOI Embedded MRAM Technology for Industrial-grade MCU and IOT Applications," *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 2.3.1-2.3.4.

[15] H. L. Chiang *et al.*, "Cold MRAM as a Density Booster for Embedded NVM in Advanced Technology," *2021 Symposium on VLSI Technology*, 2021, pp. 1-2.

[16] B. Dai *et al.*, "Review of voltage-controlled magnetic anisotropy and magnetic insulator," *Journal of Magnetism and Magnetic Materials*, vol 563, 2022.

[17] Wu, Y. C., et al. "Deterministic and field-free voltage-controlled MRAM for high performance and low power applications." *2020 IEEE Symposium on VLSI Technology*. IEEE, 2020.

[18] C.Grezes *et al.*, "Ultra-low switching energy and scaling in electric-field-controlled nanoscale magnetic tunnel junctions with high resistance-area product." *Appl. Phys. Lett.* 2016.

[19] L.Xiang *et al.*, "Enhancement of voltage-controlled magnetic anisotropy through precise control of Mg insertion thickness at CoFeB|MgO interface." *Appl. Phys. Lett.* 2017.

[20] C. Grezes et al., "Write Error Rate and Read Disturbance in Electric-Field-Controlled Magnetic Random-Access Memory," in IEEE Magnetics Letters, vol. 8, pp. 1-5, 2017.

[21] P. Jain *et al.*, "13.2 A 3.6Mb 10.1Mb/mm2 Embedded Non-Volatile ReRAM Macro in 22nm FinFET Technology with Adaptive Forming/Set/Reset Schemes Yielding Down to 0.5V with Sensing Time of 5ns at 0.7V," *ISSCC*, 2019, pp. 212-214.

[22] Q. Dong et al., "A 1-Mb 28-nm 1T1MTJ STT-MRAM With Single-Cap Offset-Cancelled Sense Amplifier and In Situ Self-Write-Termination," in *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 231-239, Jan. 2019.

[23] V. Tripathi and B. Murmann, "Mismatch Characterization of Small Metal Fringe Capacitors," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 8, pp. 2236-2242, Aug. 2014.

[24] H. Lee *et al.*, "Analysis and Compact Modeling of Magnetic Tunnel Junctions Utilizing Voltage-Controlled Magnetic Anisotropy," in *IEEE Transactions on Magnetics*, vol. 54, no. 4, pp. 1-9, April 2018.

[25] H. Valavi *et al.*, "A 64-Tile 2.4-Mb In-Memory-Computing CNN Accelerator Employing Charge-Domain Compute," in *IEEE Journal of Solid-State Circuits*, vol. 54, no. 6, pp. 1789-1799, June 2019

[26] Z. Jiang, S. Yin, J. -S. Seo and M. Seok, "C3SRAM: An In-Memory-Computing SRAM Macro Based on Robust Capacitive Coupling Computing Mechanism," in *IEEE Journal of Solid-State Circuits*, vol. 55, no. 7, pp. 1888-1897, July 2020.

[27] Y. -C. Chiu *et al.*, "A 22nm 4Mb STT-MRAM Data-Encrypted Near-Memory Computation Macro with a 192GB/s Read-and-Decryption Bandwidth and 25.1-55.1TOPS/W 8b MAC for AI Operations," *ISSCC*, 2022