A Variation Robust Inference Engine Based on STT-MRAM with Parallel Read-Out

Yandong Luo[1], Xiaochen Peng[1], Ryan Hatcher[2], Titash Rakshit[2], Jorge Kittl[2], Mark S Rodder[2], Jae-sun Seo[3] and Shimeng Yu[1]
[1]Georgia Institute of Technology, Atlanta, GA 30332, USA,
[2]Samsung Semiconductor Inc., Austin, TX 78754, USA, [3]Arizona State University, Tempe, AZ 85281, USA
Email: shimeng.yu@ece.gatech.edu

***Abstract*—STT-MRAM is a promising candidate as embedded non-volatile memory (NVM) at 28nm and beyond. Due to its limited on/off ratio, STT-MRAM is often used as digital memory that only allows row-by-row read-out for near-memory computing. This work proposes design strategies to overcome this limitation with a new bit-cell design to enable parallel read-out for in-memory computing, which is of great interests for deep neural network (DNN) acceleration. We consider the non-ideal device properties that degrade inference accuracy including small on/off ratio, cell-to-cell MTJ conductance variation and current sense amplifier (CSA) offset. We propose three techniques to minimize inference accuracy degradation: 1) a 2T-2MTJ bit-cell design with high on/off ratio, 2) redundancy for MSB weights to mitigate the impact of MTJ conductance variations, and 3) a hybrid-layer mapping scheme to reduce column current thus mitigating CSA offset effect. DNN benchmarking results show that on CIFAR-10 dataset, the inference accuracy can be maintained at > 90% in the presence of 10% MTJ conductance variations, and >87.5% after considering CSA offset effect, with minimal 8% energy and 4% chip area overhead.**

Keywords—DNN, In-memory computing, STT-MRAM

I. INTRODUCTION

In-memory computing [1] is proposed to accelerate the intensive vector-matrix multiplication in DNN algorithms, where the multiplication is conducted in analog column current domain and the digital outputs are obtained after analog to digital converter (ADC). So far, in-memory computing mostly utilizes RRAM or PCM due to their analog nature and large on/off ratio [2]. Due to the lower programming voltage, STT-MRAM is a more promising technology for embedded NVM at 28nm and beyond. However, in-memory computing designs with multiple rows read-out in parallel are more sensitive to non-ideal device properties than binary memories. In particular, we find that the low on/off ratio (TMR=1.67-2.15 [3]), large cell-to-cell MTJ conductance variation induced by process variations (>7% [3]), aggregately degrade the DNN inference accuracy, as shown in Fig. 1(a) and 1(b). In addition, Flash ADC implemented by multilevel current sense amplifier (ML-CSA) suffers from offset due to process variations [4], which further degrades the inference accuracy by mis-quantizing the partial sum (Psum) current I_{psum}.

For the inference accuracy loss induced by device variation, solutions have been proposed to retrain neural network after chip fabrication. But they either increase the workloads during chip testing stage or need additional circuit modules to support on-chip re-training [5] [6]. On the other hand, training weights with noise during software training stage may degrade the baseline accuracy [6] [7]. Moreover, the robustness of these solutions against CSA offset has not been studied. In addition, for row-parallel reads, the relatively small on-state resistance (R_{ON}) of STT-MRAM becomes problematic, as larger area and energy are consumed in the peripheral circuitry to drive and read STT-MRAM than PCM/RRAM.

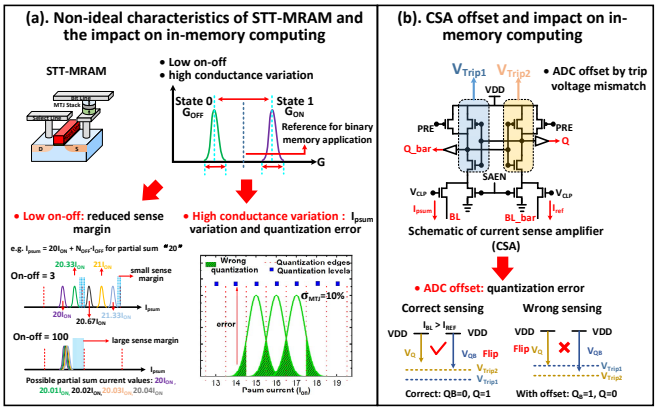


Fig. 1 (a) an illustration of the non-ideal characteristics of STT-MRAM: low on/off ratio, high MTJ conductance variations and the impacts on in-memory computing. (b) Current sense amplifier (CSA) offset leads to quantization error for the partial sum current.

In this work, circuit-device interaction strategies without re-training are proposed to maintain the inference accuracy considering the low on/off ratio, MTJ conductance variations and CSA offset. The proposed strategies are validated by software-hardware co-simulation for a 7-layer convolutional neural network (CNN) with 4-bit weight precision and 6-bit activations in TensorFlow platform for CIFAR-10 dataset.

II. CHALLENGES USING STT-MRAM FOR COMPUTING

The challenges for in-memory computing using STT-MRAM are discussed. Fig. 1(a) shows a process-element (PE) design with 1T-1MTJ cell for parallel read-out. The weights and input vectors are encoded to the 1T-1MTJ cell conductance and input voltage cycles, respectively. To implement 4-bit weights, 4 columns are grouped as one weight, and the partial sum digitized from each column goes through shift and add operations. 6-bit input is represented by 6 cycles with another round of shift and add. The convolution kernel is mapped into the array using the method proposed in [8]. The 7-layer CNN architecture used for simulation is listed in Table I. The PE size is assumed to be 64×128 memory cells in this work. Larger array size will be more area-efficient, however, the I_{psum} may be too large considering the relatively low R_{on} of STT-MRAM. The ADC precision is assumed to be 5-bit with linear quantization.

978-1-7281-3320-1/20/$31.00 ©2020 IEEE

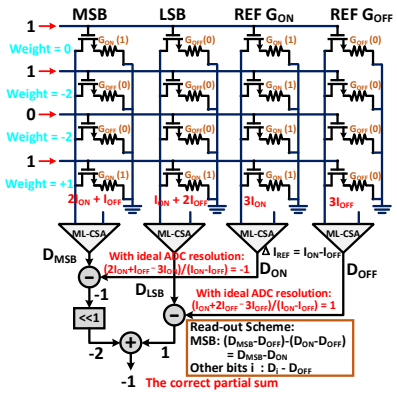
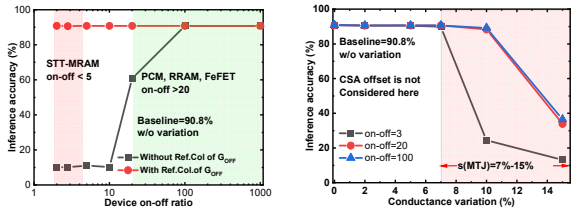


Fig. 2 Parallel read-out scheme to represent negative weight and eliminate the effect of low on/off ratio. I_{OFF} is eliminated by subtracting the LSB's partial sum from reference column of G_{OFF}. The partial sum from reference column of G_{ON} is subtracted from MSB's partial sum to represent negative weight.

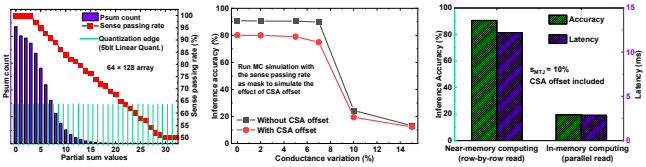
First, to overcome low on/off ratio of STT-MRAM and represent the negative weight, we use two dummy columns programmed to all on-states or all off-states, which are notated as Ref_G_{ON} and Ref_G_{OFF}, respectively. Fig. 2 shows this proposed parallel read-out scheme. To eliminate the contribution from I_{OFF} current, the Psums from Ref_G_{OFF} is subtracted from Psums of the regular columns except the MSB column. To represent the negative weight, the Psum from Ref_G_{ON} is subtracted from the MSB column (both MSB column and Ref_G_{ON} need to subtract the Psum from Ref_G_{OFF} so that it cancels out). Fig. 3 (a) shows that the inference accuracy (without dummy column) is only around 10% when on/off ratio < 10 due to accumulation of I_{OFF}. By adding dummy columns, it recovers the accuracy to 90.7% assuming zero MTJ conductance variations. However, for in-memory computing, when multiple rows are turned on, the I_{psum} will spread out due to the MTJ conductance variations. Smaller on/off ratio makes the distribution wider and extend further into the neighboring partial sum's range as shown in Fig. 1(a), thus making it easier to introduce an ADC quantization error. For digital memory applications, read accuracy is largely determined simply by the tail-to-tail gap between on/off distributions. However, for in-memory computing, the sigma of the distribution itself is more critical as the current is added up along the column in analog domain. Fig. 3(b) shows that the inference accuracy drops sharply below 30% as the σ_{MTJ} increases above 10%. As expected, devices with lower on/off ratio are more vulnerable.



(a) (b)

Fig. 3 (a). Inference accuracy vs. device on/off ratio for CIFAR-10. By using the dummy column, high inference accuracy can be achieved for devices with low on/off ratio. (b). Inference accuracy vs. device conductance variations for different on/off ratios. MTJ on/off ratio = 3, and σ_{MTJ} = 7%~15% [3].

CSA offset could mis-quantize I_{psum} to a wrong quantization level [4]. Fig. 4 (a) shows the sense passing rate (SPR) obtained from SPICE Monte-Carlo (MC) simulations with a foundry 28nm PDK. The SPR decreases as the I_{psum} increases as higher BL current leads to lower sense margin [4]. When such offset patterns are incorporated in the TensorFlow simulations, the inference accuracy drops from 90.8% to around 80% even without MTJ conductance variations (Fig. 4(b)). The comparison of near-memory computing with row-by-row read-out and in-memory computing with parallel read-out is shown in Fig. 4(c). Near-memory computing is more resistant to conductance variations and CSA offset, however, it leads to much longer latency as only one row is read-out at a time.

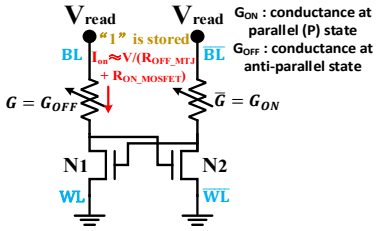


(a) (b) (c)

Fig. 4 (a) Partial sum distribution for CIFAR-10 and the corresponding sense passing rate (SPR) considering CSA offset. SPR is obtained by Monte Carlo simulations with a 28nm foundry PDK. (b) Inference accuracy vs. MTJ conductance variations with and without CSA offset. (c) A comparison between the near memory computing (row-by-row read) and in-memory computing (parallel read) regarding to inference accuracy and latency. Layer-by-Layer computing is assumed here.

III. Variation Robust Design Strategies

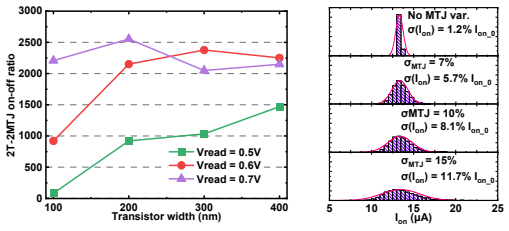
To enlarge the on/off ratio, a cross-coupled 2T-2MTJ bit-cell design is proposed in this work as shown in Fig. 5, where the conductance of the two MTJs G and G bar are always complementary to each other. When V_{read} is applied at BL and BL bar, the transistor connected to the MTJ with G_{OFF} will operate in triode region while the transistor connected to G_{ON} will be cut-off. $I_{ON} \approx V_{read} / (R_{OFF\ MTJ} + R_{ON\ MOSFET})$ will flow through BL when G=G_{OFF}, corresponding to stored weight of "1" while the I_{OFF} is determined by $I_{leakage}$ of the transistor. SPICE simulation was performed using the MTJ parameters in Table II. Fig. 6(a) shows that on/off ratio >1000 is obtained with transistor W > 100nm at 28nm node if the read voltage >0.6V. The I_{on} variation of the 2T-2MTJ cell is shown in Fig. 6(b), which is obtained by MC simulations considering variations of both MTJ conductance and transistors. The benefits of such 2T-2MTJ design include highly increased on/off ratio, reduced I_{ON}, which are both critical for in-memory computing.



Weight	G	$\bar{G}$
0	G_{ON}	G_{OFF}
1	G_{OFF}	G_{ON}

Fig. 5 Schematic of proposed 2T-2MTJ bit-cell, where the two MTJ are cross-coupled. $I_{ON} \approx V_{read}/(R_{OFF\_MTJ} + R_{ON\_MOSFET})$ will pass through BL if MTJ G is at G_{OFF}. The I_{OFF} of the bit cell is determined by the I_{OFF} of the transistor.

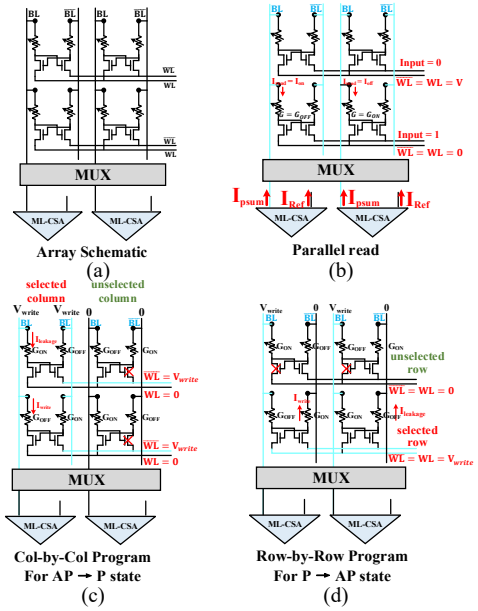
Fig. 7(a) shows the array design based on 2T-2MTJ cell. For parallel-read operation in Fig. 7(b), BL and BL bar are first precharged to the same voltage level. If input is "1", both WL and WL_bar are grounded. If the input is "0", a high voltage is



(a) (b)

Fig. 6 (a) Simulated on/off ratio for the 2T-2MTJ bit cell design. On-off > 1000 is obtained if V_{read} > 0.6V. (b) The Monte Carlo simulation results of the I_{ON} distribution for the proposed 2T-2MTJ bit cell, where different MTJ conductance variations are considered. The simulation is conducted with a foundry 28nm PDK. I_{OFF} for the bit cell is a few nA and therefore negligible.

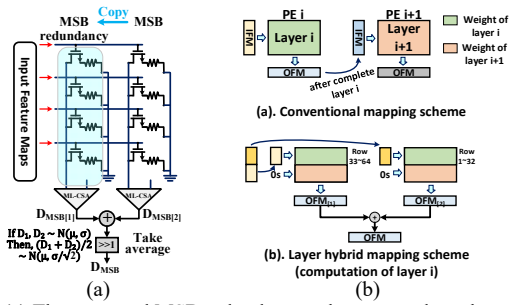
applied to both WL and WL bar so that the cell is turned off. I_{psum} is sensed by the Flash ADC based on multi-level current mode S/A through BL. For write operation, the AP to P programming is conducted column by column (Fig. 7(c)). For the selected column, V_{write} is applied to both BL and BL bar by SL switch matrix. WL is grounded if MTJ G is to be programmed and a high voltage is applied to WL bar for inhibiting MTJ G bar. Leakage path exists for MTJs at P state in the selected column. The P to AP programming is conducted row-by-row (Fig. 7(d)). V_{write} are applied to both WL and WL bar for the selected rows while they are grounded for the unselected rows. The BL (or BL bar) of the selected MTJ is grounded to allow programming current while V_{write} is applied to their BL (or BL bar) for inhibition. Leakage path exists for MTJs at AP state in the same row.



(a) (b) (c) (d)

Fig. 7 (a) Schematic of the array design with 2T-2MTJ cells. (b) Parallel read operation. The partial sum current at BL is sensed by multi-level CSA to get the digitized partial sum. (c) AP state to P state programming, which is conducted column by column. (d) P state to AP state programming, which is conducted row by row. Leakage current path exists but it will not disturb the state of the cells.

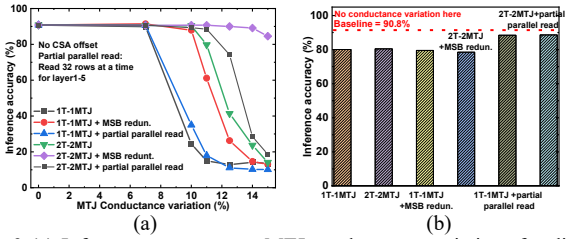
The variations of I_{psum} corresponding to MSB have more significant impact on the inference accuracy because it has higher weight significance and the quantization error will be magnified. We propose to have a redundant column for the MSB column. During read operation, the averages of the partial sums from these two columns are taken to reduce the variations, as shown in Fig. 8 (a).



(a) (b)

Fig. 8 (a) The proposed MSB redundancy scheme to reduce the variation of I_{psum} corresponding to MSB. (b) The proposed layer hybrid mapping scheme to reduce the number of rows read in parallel so that the I_{psum} distribution is shifted to the range with high SRP. The latency overhead is minimized by distributing the weights of the same layer into different PEs and adding up the partial sum outside PE.

As the sense passing rate reduces for large I_{psum}, a partial parallel read that activates only a part of the rows is proposed to reduce I_{psum} distribution to the range with higher SPR. In addition, the smaller 2T-2MTJ cell's I_{ON} (than 1T-1MTJ) also helps reduce I_{psum}. To mitigate the latency increase, a hybrid-layer mapping scheme is proposed as Fig. 8(b). In conventional mapping, one PE contains the weights from the same layer. If half of the rows are read out at a time, the PE latency is doubled. However, with hybrid-layer mapping scheme, the weights from the same layer are split into two parts and mapped to two PEs. The partial sum is added up from the two PEs externally, as shown in Fig. 8(b). Since different layers are computed sequentially in a layer-by-layer computing scheme, the two PEs can operate at the same time for one layer, thus the proposed scheme could maintain the same latency.



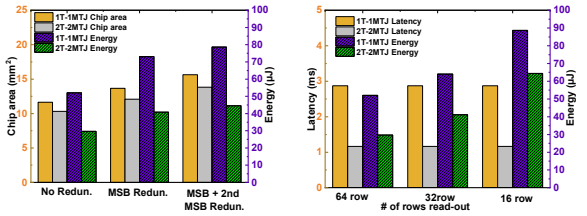
(a) (b)

Fig. 9 (a) Inference accuracy vs. MTJ conductance variations for different design schemes. CSA offset is not considered here. It shows that with 2T-2MTJ bit cell design and MSB redundancy, accuracy is resistant to MTJ conductance variations. (b) Inference accuracy considering CSA offset. The MTJ conductance variation is not considered here. The 32-row partial parallel read scheme is more robust to CSA offset. It can be explained by the fact that I_{psum} is more concentrated at range with high SRP when less rows are read-out in parallel.

IV. Benchmark Results and Discussions

Now we evaluate the efficacy of the proposed designs. First, we consider MTJ conductance variations. Fig. 9(a) shows that 2T-2MTJ cell maintains about 90.3% inference accuracy with σ_{MTJ} = 10% while the accuracy for 1T-1MTJ is reduced to ~20%. However, the inference accuracy for 2T-2MTJ cell is reduced to about 14% when σ_{MTJ} = 15%. With redundancy for MSB cell, the inference accuracy for 2T-2MTJ cell can be improved to ~84 % at σ_{MTJ} = 15%. It can also be noted that partial parallel

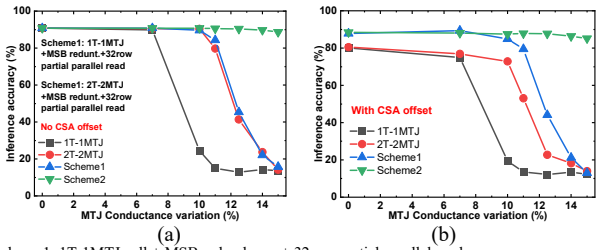
read-out does not increase the robustness to MTJ conductance variation but it is robust to CSA offset, as shown in Fig. 9(b) where only CSA offset is considered. The performance overhead is evaluated with the latest DNN+NeuroSim simulator at 28nm node [9]. Adding MSB redundancy increases the chip area and read energy as more weights are stored and read (Fig. 10(a)). The partial parallel read-out increases the energy consumption as more periphery operations are needed to read-out all the weights. It should be noted that the latency overhead is negligible with hybrid-layer mapping, as shown in Fig. 10(b).



(a) (b)

Fig. 10 (a) The overhead for implementing MSB redundancy. Both chip area and read energy increases when more redundancy is used. (b) The overhead for implementing partial parallel read. Energy consumption is increased when less rows are read-out as the periphery circuit needs to operate multiple times to read-out all the weights. However, with layer hybrid mapping, the overhead of latency is negligible.

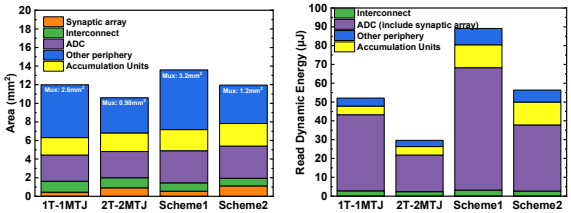
Then, two design schemes are considered: 1) 1T-1MTJ cell with MSB redundancy and 32-row partial parallel read; 2) 2T-2MTJ cell with MSB redundancy and 32-row partial parallel read. 1T-1MTJ and 2T-2MTJ design with full 64-row parallel read are used as baselines. Without CSA offset, Scheme 2) maintains about 89% accuracy at σ_{MTJ} = 15% while the accuracy for Scheme 1) drops from about 88% to 20% as σ_{MTJ} increases from 10% to 15% (Fig. 11(a)). This indicates that high device on/off ratio is important to make the chip resistant to the MTJ conductance variations. Considering CSA offset, inference accuracy can be maintained at 85% with Scheme 2) even at σ_{MTJ} = 15%, as shown in Fig. 11(b). This can be attributed to the fact that I_{psums} are more concentrated in the range with high SPR when less rows are read out simultaneously. The hardware performance of the proposed schemes obtained from NeuroSim are shown in Table III. Note that the cell size for 1T-1MTJ and 2T-2MTJ are assumed to be $36F^2$ and $72F^2$, respectively. The I_{ON} reduction in the 2T-2MTJ design enables a reduction in both the total chip area and read energy compared to the 1T-1MTJ option. Comparing Scheme 2 with 1T-1MTJ baseline, only 8% more energy consumption and 4% chip area overhead is observed, while it could maintain 87.5% inference accuracy at σ_{MTJ}=10% and with CSA offset. The chip area and energy breakdown are shown in Fig. 12.



(a) (b)

Scheme1: 1T-1MTJ cell + MSB redundancy + 32row partial parallel read
Scheme2: 2T-2MTJ cell + MSB redundancy + 32row partial parallel read

Fig.11 (a) Inference accuracy vs. MTJ conductance variations without CSA offset (b) Inference accuracy vs. MTJ conductance variations with CSA offset. Scheme2 shows robustness against conductance variations.



(a) (b)

Fig. 12 (a) Chip area and (b) read dynamic energy breakdown for the CIFAR-10 benchmark results from Table III. 2T-2MTJ based designs shows less area cost and energy consumption in the periphery circuits due to smaller I_{ON}. The area reduction of periphery circuits are mainly attributed to the area reduction of mux due to smaller transmission gate size.

V. Conclusions

In this paper, the impact of non-ideal effects of STT-MRAM is studied for in-memory computing. Design strategies including 2T-2MTJ bit cell, MSB redundancy and hybrid-layer mapping scheme are proposed. Benchmark results suggest that with an optimized design, the parallel read-out with high accuracy is feasible for STT-MRAM array even under significant MTJ conductance variations and CSA offset induced by process variations. Compared to a 1T-1MTJ bit-cell, the 2T-2MTJ bit-cell enables an overall reduction in chip area, read latency and energy despite the increase in the bit-cell area. This is due to a significant reduction in the on-state conductance, which in turn reduces the overhead of the peripheral circuitry. This work paves the way for a practical STT-MRAM based inference engine tape-out.

Acknowledgment

This work is in part supported by ASCENT, one of the SRC/DARPA JUMP centers, and Samsung Electronics.

TABLE I. The 7-layer CNN for CIFAR-10

Layer	Kernel Size
1	(3,3,3,64)
2	(3,3,64,64)
	MaxPool
3	(3,3,64,128)
4	(3,3,128,128)
	MaxPool
5	(3,3,128,256)
6	(3,3,128,256)
	MaxPool
7	(4096,10)

TABLE II. The device parameters for SPICE simulation for 2T-2MTJ bit cell

Parameters	Value
Technology node	28nm
R_{on} (R_p)	14.8kΩ [10]
R_{off} (R_{ap})	41.4kΩ [10]
TMR	1.8 [3]
σ_{MTJ}	7% ~ 15% [3]
t_{read}	10ns

TABLE III. Estimated chip performance for different design schemes (28nm node)

	1T-1MTJ	2T-2MTJ	Scheme 1	Scheme 2
CIFAR10 Inference accuracy (σ_{MTJ}=10%,w/CSA offset)	~19.45%	~73%	~85%	~87.5%
Chip area (mm^2)	11.65	10.33	13.68	12.09
Read Dynamic Energy (layer-by-layer, μJ)	52.09	29.67	89.09	56.26
Leakage Energy (μJ)	0.11	0.044	0.130	0.053
Latency (ms)	2.875	1.167	2.881	1.173
Energy efficient (TOPS/W)	2.93	5.14	1.712	2.71
Throughput (FPS)	347.86	856.625	347.16	852.74

Scheme1: 1T-1MTJ cell + MSB redundancy + 32row partial parallel read
Scheme2: 2T-2MTJ cell + MSB redundancy + 32row partial parallel read
Software accuracy baseline = 90.8%

REFERENCES

[1]. S. Yu, "Neuro-inspired computing with emerging nonvolatile memorys," in Proceedings of the IEEE, vol. 106, no. 2, pp. 260-285, Feb. 2018.

[2]. G. W. Burr et al., "Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: Comparative performance analysis (accuracy, speed, and power)," 2015 IEEE International Electron Devices Meeting (IEDM), Washington, DC, 2015, pp. 4.4.1-4.4.4.

[3]. Y. J. Song et al., "Highly functional and reliable 8Mb STT-MRAM embedded in 28nm logic," 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2016, pp. 27.2.1-27.2.4.

[4]. C. Lo et al., "A ReRAM Macro Using Dynamic Trip-Point-Mismatch Sampling Current-Mode Sense Amplifier and Low-DC Voltage-Mode Write-Termination Scheme Against Resistance and Write-Delay Variation," in IEEE Journal of Solid-State Circuits, vol. 54, no. 2, pp. 584-595, Feb. 2019.

[5]. A. Mohanty, X. Du, P. Chen, J. Seo, S. Yu and Y. Cao, "Random sparse adaptation for accurate inference with inaccurate multi-level RRAM arrays," 2017 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2017, pp. 6.3.1-6.3.4.

[6]. C.C. Chang, M. H. Wu, J.W. Lin, C. H. Li, V. Parmar, H. Y. Lee, J. H. Wei, S. S. Sheu, M. Suri, T. S. Chang, and T. H. Hou, "NV-BNN: An Accurate Deep Convolutional Neural Network Based on Binary STT-MRAM for Adaptive AI Edge". Annual Design Automation Conference, Las Vegas, NV, USA, 2019

[7]. Y. Long, X. She and S. Mukhopadhyay, "Design of Reliable DNN Accelerator with Un-reliable ReRAM," Design, Automation & Test in Europe Conference & Exhibition (DATE), Florence, Italy, 2019

[8]. X. Peng, R. Liu and S. Yu, "Optimizing Weight Mapping and Data Flow for Convolutional Neural Networks on RRAM Based Processing-In-Memory Architecture," 2019 IEEE International Symposium on Circuits and Systems (ISCAS), Sapporo, Japan, 2019, pp. 1-5.

[9]. https://github.com/neurosim/DNN_NeuroSim_V1.0

[10]. Y. Kim et al., "Integration of 28nm MJT for 8~16Gb level MRAM with full investigation of thermal stability," 2011 Symposium on VLSI Technology - Digest of Technical Papers, Honolulu, HI, 2011, pp. 210-211.