Efficiency of Double-barrier Magnetic Tunnel Junction-based Digital eNVM Array for Neuro-Inspired Computing

Moposita, Tatiana; Garzon, Esteban; Crupi, Felice; Trojman, Lionel; Vladimirescu, Andrei; Lanuzza, Marco

DOI
10.1109/TCSII.2023.3240474

Publication date
2023

Document Version
Final published version

Published in
IEEE Transactions on Circuits and Systems II: Express Briefs

Citation (APA)
Moposita, T., Garzon, E., Crupi, F., Trojman, L., Vladimirescu, A., & Lanuzza, M. (2023). Efficiency of Double-barrier Magnetic Tunnel Junction-based Digital eNVM Array for Neuro-Inspired Computing. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *70*(3), 1254-1258. https://doi.org/10.1109/TCSII.2023.3240474

Efficiency of Double-Barrier Magnetic Tunnel Junction-Based Digital eNVM Array for Neuro-Inspired Computing

Tatiana Moposita, *Student Member, IEEE*, Esteban Garzón, *Member, IEEE*,
Felice Crupi, *Senior Member, IEEE*, Lionel Trojman, *Senior Member, IEEE*,
Andrei Vladimirescu, *Life Fellow Member, IEEE*, and Marco Lanuzza, *Senior Member, IEEE*

***Abstract*—This brief deals with the impact of spin-transfer torque magnetic random access memory (STT-MRAM) cell based on double-barrier magnetic tunnel junction (DMTJ) on the performance of a two-layer multilayer perceptron (MLP) neural network. The DMTJ-based cell is benchmarked against the conventional single-barrier MTJ (SMTJ) counterpart by means of a comprehensive evaluation carried out through a state-of-the-art device-to-algorithm simulation framework. The benchmark is based on the MNIST handwritten dataset, Verilog-A MTJ compact models developed by our group, and 0.8 V FinFET technology. Our results point out that the use of DMTJ-based STT-MRAM cells to implement digital embedded non-volatile memory (eNVM) synaptic core allows write/read energy and latency improvements of about 53%/61% and 66%/17%, respectively, as compared to the SMTJ-based equivalent design. This is achieved by ensuring a reduced area footprint and a learning accuracy of about 91%. Such results make the DMTJ-based STT-MRAM cell a good eNVM option for neuro-inspired computing.**

***Index Terms*—STT-MRAM, double-barrier magnetic tunnel junction (DMTJ), multilayer perceptron (MPL), online classification, MNIST dataset, energy-efficiency.**

I. Introduction

NEURO-INSPIRED computing systems such as deep neural networks (DNNs) have been successfully demonstrated in machine learning (ML) applications including image processing/classification/recognition, natural language processing, and visual intelligence [1], [2], [3]. Due to features such as small cell area footprint, short programming time, and good endurance and data retention [4], [5], there is an increasing interest in the field of neuro-inspired computing exploiting emerging non-volatile memories (eNVMs) such as resistive RAM (RRAM), phase change memory (PCM), spin-transfer-torque magnetic random access memory (STT-MRAM), and ferroelectric field-effect transistor (FeFET), allowing flexibility to the development of DNNs. Although analog synapse eNVM-based architectures could be competitive in terms of energy and latency, they mainly suffer from low online learning accuracy [6]. To deal with this issue, digital synapse based architectures have been widely considered [4], [7]. As potential eNVM candidate for digital synapse devices, STT-MRAM cell offers low operating voltage, enough good speed operation, high-density, relatively large endurance, low fabrication cost, low-power consumption, and scalability [8], [9], [10]. Typically, STT-MRAM based DNN implementations are based on conventional single-barrier MTJ (SMTJ) devices [11]. However, it is required high writing current, thus limiting the overall energy-efficiency and latency of DNN. To counteract with this, a solution consists of using double-barrier MTJ (DMTJ), with two reference layers, to enable higher-speed operation, lower power consumption, and more energy-efficient switching process [10], [12], [13].

We evaluate the impact of DMTJ-based STT-MRAM cell on DNN, by using Cadence-Virtuoso environment for circuit-level simulations, along with the multilayer perceptron (MLP) + NeuroSimV3.0 simulator computing-in-memory (CiM) based neural network accelerator [7]. More precisely, NeuroSim is used to support a 2-layer MLP neural network to benchmark the DNN architecture, relied on SMTJ-based and DMTJ-based digital synapse devices, in online learning and offline classification with MNIST handwritten dataset.

Our results point out that the use of DMTJ-based STT-MRAM cell in a digital eNVM synaptic core allows write/read energy and latency improvements of about 53%/61% and 66%/17%, respectively, as compared to the SMTJ-based counterpart. This is also achieved by ensuring a learning accuracy of about 91%, suggesting that the DMTJ-based STT-MRAM cell could be a promising candidate for digital synapse in neuro-inspired computing.

Manuscript received 22 December 2022; accepted 23 January 2023. Date of publication 30 January 2023; date of current version 6 March 2023. This work was supported by the Italian Ministry of University and Research (MUR) through Project PRIN under Grant 2020LWPKH7. This brief was recommended by Associate Editor A. L. Zimpeck. *(Corresponding author: Tatiana Moposita.)*

Tatiana Moposita is with the Department of Computer Engineering, Modeling, Electronics and Systems, University of Calabria, 87036 Rende, Italy, also with the Laboratoire d'Informatique, Signal et Image, Électronique et Télécommunications, Institut Supérieur d'Électronique de Paris, 75006 Paris, France, and also with Sorbonne University, 75006 Paris, France (e-mail: tatiana.moposita@dimes.unical.it).

Esteban Garzón, Felice Crupi, and Marco Lanuzza are with the Department of Computer Engineering, Modeling, Electronics and Systems, University of Calabria, 87036 Rende, Italy (e-mail: esteban.garzon@unical.it; felice.crupi@unical.it; m.lanuzza@dimes.unical.it).

Lionel Trojman is with the Laboratoire d'Informatique Signal et Image Telecom et Electronique, Institut Supérieur d'Électronique de Paris, 75006 Paris, France (e-mail: lionel.trojman@isep.fr).

Andrei Vladimirescu is with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA, and also with Technical University Delft, 2628 CD Delft, The Netherlands (e-mail: andreiv@berkeley.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSII.2023.3240474.

Digital Object Identifier 10.1109/TCSII.2023.3240474

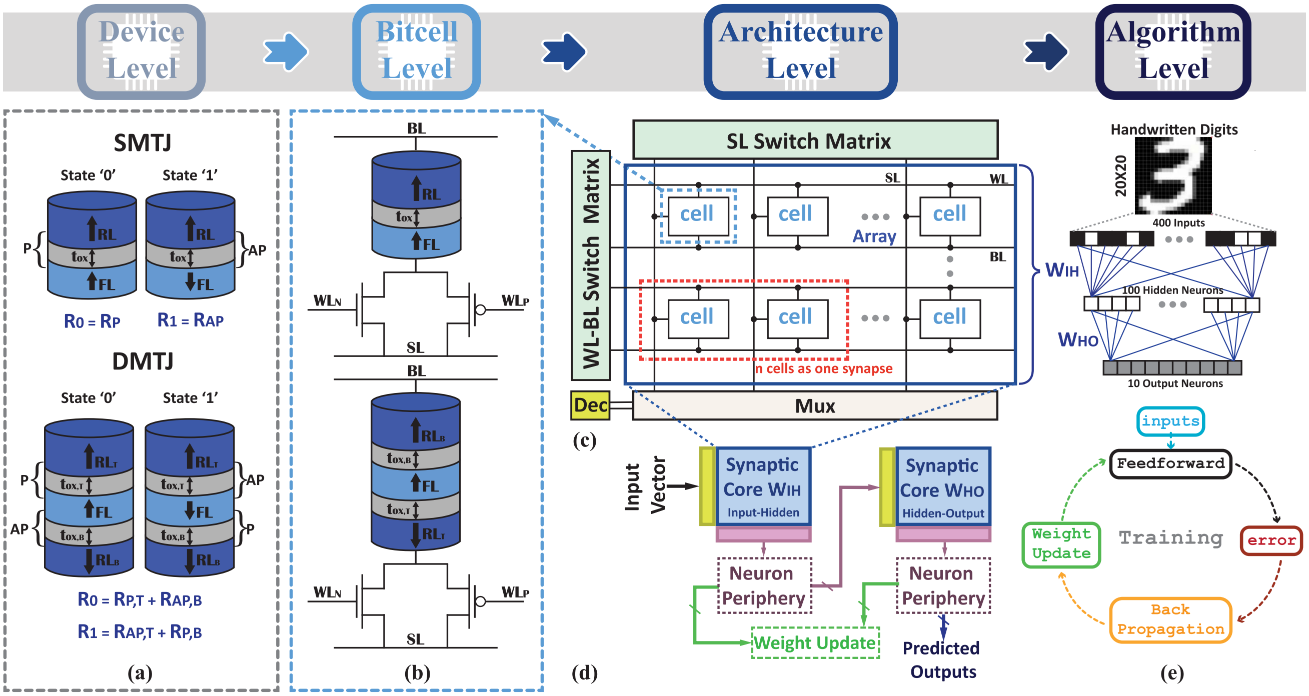


Fig. 1. Overview of NeuroSim framework from Device to Algorithm-level, (a) STMJ and DMTJdevice, (b) SMTJ-based and DMTJ-based bitcell configurations, (c) Circuit block diagram of digital eNVM synaptic core, (d) Circuit block diagram for hardware implementation of the 2-layer MLP NN. The weights are mapped to synaptic cores, which are computation units designed for performing weighted sum and weight update. (e) Training flow of Neural Network, the MNIST images are crooped and encoded into black and white data for simplification on hardware implementation.

This brief is organized as follows. Section II details the simulation framework, its customization and setting from device-to-algorithm level. Section III discusses the system level performance evaluation in terms of accuracy, area, latency and energy. Finally, Section IV concludes this brief.

II. Simulation Framework – MLP + NeuroSimV3.0

NeuroSim simulator allows to estimate the algorithm-level performance by emulating the online learning and offline classification scenario with MNIST handwritten dataset in a 2-layer MLP neural network [7], [14], [15], [16]. As shown in Fig. 1, the evaluation framework takes into account the whole system from device and bitcell levels to memory architecture and algorithm levels. The input parameters of the simulation tool include memory type, non-ideal device parameters, transistor technology node, network topology and array size, training dataset and traces, etc. For the full list of input parameters/variables, the reader is referred to [7]. The outputs of the simulator include: (1) the memory architecture-level performance metrics, such as area, latency, dynamic energy, and (2) algorithm-level learning accuracy in run-time. As for the design options of digital synaptic arrays, SRAM or eNVM bitcells can be used.

A. Device Level

As shown in Fig. 1 (a), we consider STT-SMTJ/DMTJ devices, whose main physical and performance parameters are listed in Table I. The STT-MTJs are described through Verilog-A based compact models [17], [18], which have been validated against full micromagnetic and experimental results. In particular, the STT-MTJ models utilize experimental data reported in [19]. These models further account for the impact of process variability on the STT-MTJs. Specifically, the variability, modeled by incorporating Gaussian-distributed variations, was set to 1% for both the free-layer and oxide thickness, 3% for the tunnel magnetoresistance (TMR) ratio, and 5% for the cross-section area [10].

TABLE I
SMTJ and DMTJ Device Parameters [10]

Parameter	Units	Value
Diameter (d)[a]	nm	28
Saturation magnetization (Ms)[a]	A/m	1000×10^3
Magnetic damping (α)[a]	-	0.025
Spin-polarization factor (η)[a]	-	0.67
FL thickness (t_{FL})[a]	nm	1.2
SMTJ oxide thickness	nm	0.85
DMTJ top oxide thickness	nm	0.85
DMTJ bottom oxide thickness	nm	0.4
TMR at 0 V (TMR(0))[c]	%	150

[a] Same value for SMTJ and DMTJ devices.
[c] Same value for SMTJ barrier and DMTJ top/bottom barriers

1) SMTJ: The SMTJ consists of two types of ferromagnetic (FM) layers, one with fixed magnetization called reference layer (RL), and the other with a free magnetization named as free layer (FL), whose magnetization direction can be changed by applying a switching current greater than the critical switching current of the device [10]. Based on the relative magnetization direction of the FL and RL, the SMTJ can reside in one of two stable states: parallel (P) or antiparallel (AP). If two FM layers have the same magnetization directions, i.e., RL and FL in P, the resistance of the MTJ is low (R_0), indicating a "0" state. Conversely, if the two layers have different magnetization directions, i.e., RL and FL in AP, the resistance of the MTJ is high (R_1), indicating a "1" state [10].

2) DMTJ: The FL is sandwiched between two MgO oxide barriers, each of them interfaced with one RL. The low resistance state (“0”) corresponds to FL in P and AP with respect to the RL top and RL bottom, respectively. As for the high resistance state (“1”), the FL is in AP and P with respect to RL bottom and RL top, respectively. Accordingly, the DMTJ resistances in states “0” and “1” can be calculated as $R_0=R_{P,T}+R_{AP,B}$ and $R_1=R_{AP,T}+R_{P,B}$, respectively, [10]. Due to the presence of the second reference layer, the spin-transfer torque is enhanced [18]. Therefore, the write switching currents is reduced as compared to the conventional SMTJ device.

TABLE II
BITCELL-LEVEL PARAMETERS

	Parameter	Unit	STMJ	DTMJ
bitcell	Cell Area	F^2	231	131
	Resistance ON	Ω	9513	11370
	Resistance OFF	Ω	16390	22170
	Conductance ON/OFF	–	1.79	1.97
	Read Voltage	V	0.338	0.121
	Read Energy	fJ	20.9	5.76
	Read Pulse Width	ns	1.00	1.00
	Write Energy	fJ	185	4.80
	Write Voltage LTD	V	0.788	1.09
	Write Voltage LTP	V	0.898	0.564
	Write Pulse Width	ns	3.39	1.16

B. Bitcell- to Memory Architecture-Level

Fig. 1(b) shows the considered SMTJ-based and DMTJ-based bitcell configurations designed exploiting a 28 nm FinFET technology featuring a nominal supply voltage of 0.8 V. These are referred to the two complementary transistors and one MTJ (2T1MTJ) cells in reverse and standard connection (2T1MTJ-RC and 2T1MTJ-SC) for the SMTJ- and DMTJ-based bitcells, respectively. According to the study carried out in [10], those considered are the most write energy-efficient bitcell configurations.

At the architecture level shown in Fig. 1(c) and Fig. 1(d), two synaptic cores of 2-layer MLP are considered. Each synaptic core is a computation unit specifically designed for weighted sum and weight update [7], [14]. Among the available design options for the synaptic cores, we considered the digital eNVM based on pseudo-crossbar array.

For the sake of an accurate modeling of the MLP NN, we have adapted NeuroSim to match the considered FinFET technology node. To this aim, a fine-grained electrical characterization of the transistors was carried out exploiting Cadence Virtuoso tool. More specifically, the MLP NN utilizes transistor information like gate capacitance, mobility, threshold voltage, ON/OFF current, etc. Therefore, the synaptic core and periphery neuron of the MLP NN are accurately built for the considered 28 nm FinFET process.

C. Algorithm Level

At the algorithm level, the standard MNIST benchmark data is used for online learning (6k images for training dataset and 10k images for testing dataset) and offline classification [7].

The considered MLP is a fully connected neural network, where each neuron node in one layer connects to every neuron node in the following layer. Fig. 1(e) shows the flow of Neural Network, where the MNIST images are cropped and encoded into black and white data for simplification on hardware implementation. The network consists of an input layer, hidden layer and output layer. The connections between input-hidden and hidden-output layers represent the weight matrix W_{IH} and W_{HO}, respectively. As shown in Fig. 1(e), the network topology contains 400 neurons (20×20 MNIST image) of input layer, 100 neurons of hidden layer, and 10 neurons (10 classes of digits) of output layer.

III. SIMULATION RESULTS

NeuroSim framework shown in Fig. 1 was properly calibrated with the 0.8 V FinFET technology parameters, along with the bitcell electrical characteristics of the considered 2T1MTJ-based bitcells, which are the cells of the pseudo-crossbar eNVM digital synaptic core. Bitcell-level results consider both SMTJ/DMTJ and FinFET device-to-device variability through extensive Monte Carlo simulations. Table II shows the bitcell-level parameters of the energy-optimal cell size and configurations (refer to Fig. 1(b)). It is worth to mention that these results are carried out at parity of tunnel magnetoresistance ratio (TMR), and oxide thickness, i.e., $t_{ox,SMTJ} = t_{ox,t,DMTJ} = 0.85$ nm. Performance results for write and read operations are obtained, assuring a write-error-rate (WER) of 10^{-7} and read disturbance rate (RDR) of 10^{-9}, respectively. From Table II, it is clear that thanks to the reduced switching and read currents, the DMTJ-based bitcell is the most energy-efficient alternative under write/read operations. Overall, at bitcell-level, the DMTJ-based alternative shows energy savings of about 72% and 97% for read and write operations, while assuring faster (65.7%) switching in contrast to the SMTJ-based bitcell.

The parameters reported in Table II were used as input in NeuroSim to evaluate the algorithm-level performance.

Considering the training time, we employ 15 epochs (i.e., number of training iterations), 8000 and 1000 MNIST images for training and testing, respectively, giving a total of 12000 MNIST images being trained. We used the online learning in hardware configuration, which handle testing and training for both weight sum and weight update all in hardware.

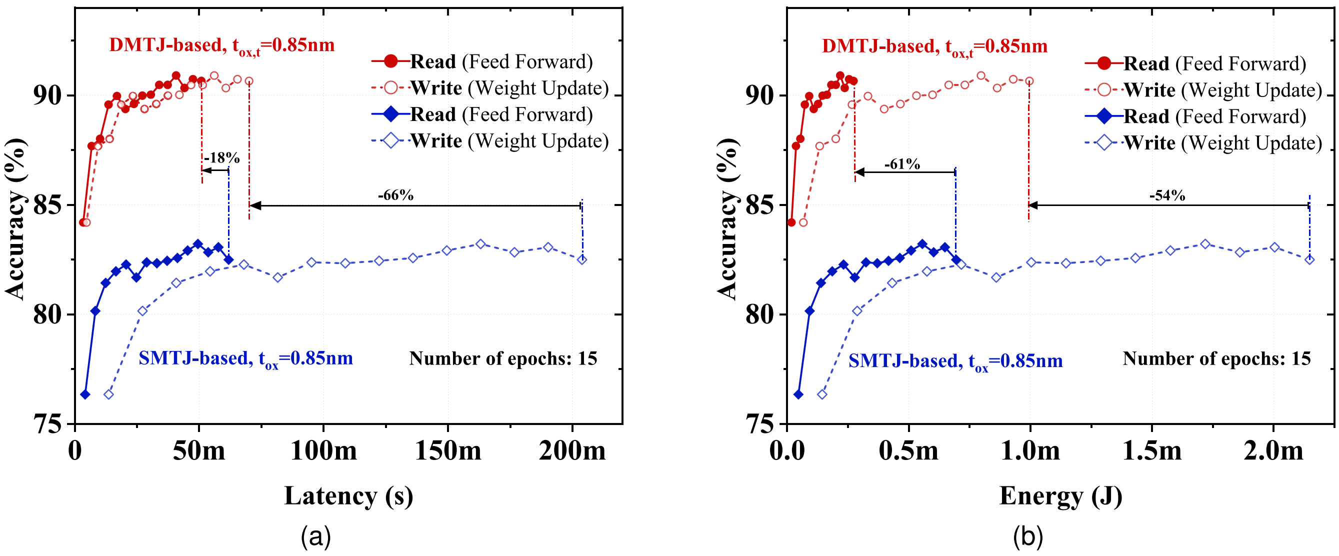
A. Performance Analysis

The SMTJ- and DMTJ-based 2-layer MLP neural network performance is evaluated in terms of learning accuracy versus latency and energy consumption, calculated at the run-time.

The read (weighted sum-feed forward operation) and write (weight update operation) latency and energy are shown in Fig. 2. We can observe that the weighted sum and weight update operations associated to the DMTJ-based eNVM cell achieve the highest accuracy much faster as compared to the SMTJ-based counterpart, while at the same time ensuring less energy consumption. This is due to the reduced energy/write-pulse width of the DMTJ-based bitcell (refer to Table II).

From Fig. 2(a), it is worth noting that the delta latency (i.e., time between iterations) in feed forward operation, for both STMJ- and DMTJ-based alternatives, is roughly the same, mainly do to the similar requirement for the read pulse width.

As for the weight update operation, the delta latency between each epoch is 14ms and 4.7ms, respectively. This can



(a) (b)

Fig. 2. Trace of Latency and Energy in feed forward and weight update during online learning for both STMJ- and DMTJ-based when considering a top barrier of $t_{ox,SMTJ} = t_{ox,t,DMTJ} = 0.85$ nm.

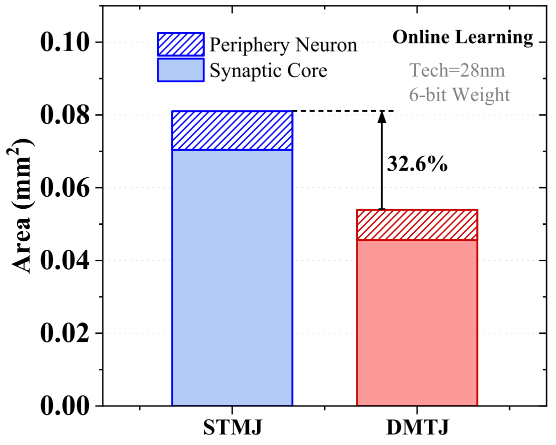


Fig. 3. Area of MLP NN architecture for both SMTJ-based and DMTJ-based synaptic cores.

be explained due to the larger pulse width required for writing operation. As compared with the SMTJ-based alternative, the DMTJ-based cell shows an improvement in terms of latency, of about 18% and 66% in feed forward and weight update operations, respectively, during online learning. Similar results have been obtained for the energy consumption, as shown in Fig. 2(b). The DMTJ-based cell shows lower energy consumption as compared to the SMTJ-based alternative, owing to its reduced bitcell read/write energy. The results showed an improvement of about 61% and 54% during feed forward and weight update, respectively.

The benchmark results show that, while the DMTJ-based solution achieves a good accuracy of (> 90%), the SMTJ-based neural network reaches a learning accuracy of about 83%.

The cause of degradation in terms of learning accuracy is attributed to the devices' poor ON/OFF ratio [6].

In addition, we estimate the area occupation as extracted from NeuroSim. Fig. 3 shows the total area footprint. The area occupation for the SMTJ-based and DMTJ-based alternatives is 0.0788 mm^2 and 0.0531 mm^2, respectively. DMTJ-based bitcell can achieve the smallest area footprint due to the smaller bitcell area (see Table II), which corresponds to the energy-optimal cell size.

B. *Impact of Synaptic Device Properties on Accuracy*

During the weight update, the conductance of the device should be sufficiently large, i.e., the lowest conductance state (OFF-state) should be low enough to represent the zero weight in the algorithm [6]. To quantify the impact of the device properties on the learning accuracy, we carried out an analysis for both STT-MTJ alternatives by varying $t_{ox}/t_{ox,t}$.

If we decrease the oxide thickness for both devices, the ON and OFF resistance of the bitcell will be affected. When considering a top barrier of $t_{ox,SMTJ} = t_{ox,t,DMTJ} = 0.80$ nm, the conductance ON/OFF ratio for SMTJ- and DMTJ-based cell are 1.91 and 1.88, respectively. The reduced ON/OFF conductance ratio in the DMTJ-based cell can be explained by the presence of the second oxide barrier. Therefore, the accuracy for SMTJ-based cell increases by 5.9%, while DMTJ-based cell decreases by 2.4%, see Fig. 4. Note that the use of very thin oxide barriers could lead to breakdown of the MTJ structure. To deal with this reliability issue, the write voltages have to be reduced [20].

Table III shows the assessment of energy, latency, accuracy, and area results obtained at different values of oxide thickness, for SMTJ- and DMTJ-based cells. From table III, the SMTJ-based cell at t_{ox}=0.85 nm has less latency and energy consumption compared with SMTJ-based cell at t_{ox}=0.80 nm in feed forward operation. In contrast, during the weight update, the latency and energy consumption increases when t_{ox}=0.85 nm. Moreover, during the feed forward and weight update operation the DMTJ-based cell at t_{ox}=0.80 nm results less energy hungry than its t_{ox}=0.85 nm counterpart. Furthermore, the DMTJ-based cell at t_{ox}=0.85 nm is faster compared with t_{ox}=0.80 nm along the weight sum. During the weight update, the DMTJ-based cell at t_{ox}=0.80 nm has improved latency over the t_{ox}=0.85 nm counterpart.

TABLE III
BENCHMARK RESULTS OF SMTJ- AND DMTJ-BASED CELL AT $T_{OX,SMTJ} = T_{OX,T,DMTJ} = 0.85$ nm AND $T_{OX,SMTJ} = T_{OX,T,DMTJ} = 0.80$ nm

Parameter	Oxide thickness	Energy (mJ)		Latency (ms)	
		Read	Write	Read	Write
SMTJ	**0.80** nm	0.712	2	76.2	128
	0.85 nm	0.695	2.15	61.8	204
DMTJ	**0.80** nm	0.273	0.899	58.8	55.4
	0.85 nm	0.2731	0.996	50.9	70.0
SMTJ (0.8 nm vs 0.85 nm) (%)		2.45	-6.98	23.3	-37.25
DMTJ (0.8 nm vs 0.85 nm) (%)		-0.04	-9.74	15.52	-20.86
DMTJ vs SMTJ (@ 0.80 nm) (%)		-61.66	-55.05	-22.83	-56.72

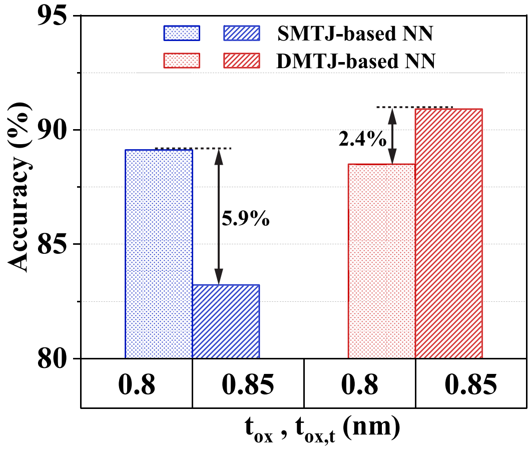


Fig. 4. Learning accuracy versus oxide thickness (t_{ox} or $t_{ox,t}$) for SMTJ- and DMTJ-based neural networks.

Finally, we have also performed the comparative study of the DMTJ- and SMTJ-based solutions considering t_{ox}=0.80 nm. The DMTJ-based cell shows an improvement in terms of latency, of about 23% and 57% in feed forward and weight update operations, respectively, compared with the SMTJ-based cell. As for the energy consumption, the analysis shows similar results compared with $t_{ox,SMTJ} = t_{ox,t,DMTJ} = 0.85$ nm, showing accuracy improvements of about 62% and 55% during feed forward and weight update, respectively.

IV. CONCLUSION

In this brief, we have explored the STT-MTJ synaptic pseudo-crossbar array architecture and device/transistor models in NeuroSim. We have used the NeuroSim emulator to evaluate the learning accuracy with 2-layer MPL neural networks at the run-time of online learning in eNVM devices such as MTJ-based STT-MRAM. Our results show that, at parity of TMR and oxide thickness, as compared to the conventional SMTJ-based alternative, the DMTJ-based solution proves to be faster during feed forward and weight update operations of about 18% and 66%, respectively, more energy efficient under read (−60.7%) and write operation (−53.7%), and less area hungry (−35%) at an energy-optimal bitcell configuration/size. This occurs while also achieving an accuracy closed to 91% when running the neural network with the MNIST dataset. Our study suggests that DMTJ-based eNVM synaptic cores are good candidates to replace conventional SRAM-based solutions.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–9.

[2] E. Garzón, A. Teman, M. Lanuzza, and L. Yavits, "AIDA: Associative in-memory deep learning accelerator," *IEEE Micro*, vol. 42, no. 6, pp. 67–75, Nov./Dec. 2022.

[3] B. B. Traore, B. Kamsu-Foguem, and F. Tangara, "Deep convolution neural network for image recognition," *Ecol. Inform.*, vol. 48, pp. 257–268, Nov. 2018.

[4] Y. Luo, X. Peng, and S. Yu, "MLP+NeuroSimV3.0: Improving on-chip learning performance with device to algorithm optimizations," in *Proc. Int. Conf. Neuromorphic Syst.*, 2019, pp. 1–7.

[5] E. Garzón, B. Zambrano, T. Moposita, R. Taco, L.-M. Prócel, and L. Trojman, "Reconfigurable CMOS/STT-MTJ non-volatile circuit for logic-in-memory applications," in *Proc. IEEE 11th Latin Amer. Symp. Circuits Syst. (LASCAS)*, 2020, pp. 1–4.

[6] P.-Y. Chen and S. Yu, "Technological benchmark of analog synaptic devices for neuroinspired architectures," *IEEE Design Test*, vol. 36, no. 3, pp. 31–38, Jun. 2019.

[7] P.-Y. Chen, X. Peng, and S. Yu, "NeuroSim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures," in *IEEE Int. Electron Devices Meeting Tech. Dig.*, 2017, pp. 1–6.

[8] N. Xu et al., "STT-MRAM design technology co-optimization for hardware neural networks," in *IEEE Int. Electron Devices Meeting Tech. Dig.*, 2018, pp. 15.3.1–15.3.4.

[9] K. Zhang et al., "High on/off ratio spintronic multi-level memory unit for deep neural network," *Adv. Sci.*, vol. 9, no. 13, 2022, Art. no. 2103357.

[10] E. Garzón et al., "Assessment of STT-MRAMs based on double-barrier MTJs for cache applications by means of a device-to-system level simulation framework," *Integration*, vol. 71, pp. 56–69, Mar. 2020.

[11] A. Khvalkovskiy et al., "Basic principles of STT-MRAM cell operation in memory arrays," *J. Phys. D, Appl. Phys.*, vol. 46, no. 7, 2013, Art. no. 74001.

[12] G. Hu et al., "Low-current spin transfer torque MRAM," in *Proc. Int. Symp. VLSI Design Autom. Test*, 2017, pp. 1–2.

[13] E. Garzón, A. Teman, and M. Lanuzza, "Embedded memories for cryogenic applications," *Electronics*, vol. 11, no. 1, p. 61, 2021.

[14] P.-Y. Chen, X. Peng, and S. Yu, "NeuroSim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 12, pp. 3067–3080, Dec. 2018.

[15] A. Lu, X. Peng, W. Li, H. Jiang, and S. Yu, "NeuroSim validation with 40nm RRAM compute-in-memory macro," in *Proc. IEEE 3rd Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, 2021, pp. 1–4.

[16] X. Peng, S. Huang, H. Jiang, A. Lu, and S. Yu, "DNN+NeuroSim V2. 0: An end-to-end benchmarking framework for compute-in-memory accelerators for on-chip training," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 40, no. 11, pp. 2306–2319, Nov. 2021.

[17] R. De Rose et al., "A compact model with spin-polarization asymmetry for nanoscaled perpendicular MTJs," *IEEE Trans. Electron Devices*, vol. 64, no. 10, pp. 4346–4353, Oct. 2017.

[18] R. De Rose, M. D'Aquino, G. Finocchio, F. Crupi, M. Carpentieri, and M. Lanuzza, "Compact modeling of perpendicular STT-MTJs with double reference layers," *IEEE Trans. Nanotechnol.*, vol. 18, pp. 1063–1070, Oct. 2019.

[19] Y. Zhang et al., "Compact model of subvolume MTJ and its design application at nanoscale technology nodes," *IEEE Trans. Electron Devices*, vol. 62, no. 6, pp. 2048–2055, Jun. 2015.

[20] E. Garzón, R. De Rose, F. Crupi, L. Trojman, A. Teman, and M. Lanuzza, "Relaxing non-volatility for energy-efficient DMTJ based cryogenic STT-MRAM," *Solid-State Electron.*, vol. 184, Oct. 2021, Art. no. 108090.

Open Access funding provided by 'Università della Calabria' within the CRUI CARE Agreement