# Efficiency of Double-barrier Magnetic Tunnel Junction-based Digital eNVM array for Neuro-inspired Computing

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

# Efficiency of Double-barrier Magnetic Tunnel Junction-based Digital eNVM array for Neuro-inspired Computing

Tatiana Moposita[1,2,*], Esteban Garzón[1], Marco Lanuzza[1], Lionel Trojman[2], Andrei Vladimirescu[2], and Felice Crupi[1]
[1]*DIMES, Università della Calabria, Rende, 87036, Italy*
[2]*LISITE, Institut Supérieur d'Électronique de Paris, Paris, 75006, France*
[*]email: tatiana.moposita@ext.isep.fr

***Abstract*—This work analyses the impact of spin-transfer torque magnetic random access memory (STT-MRAM) cells 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) alternative by means of a comprehensive evaluation carried out through a state-of-the-art device-to-algorithm simulation framework, considering the MNIST handwritten dataset, Verilog-A based MTJs compact models, and 0.8 V FinFET technology. Our results point out that the use of DMTJ-based STT-MRAM cell in a digital emerging 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 counterpart. This is also achieved by ensuring a learning accuracy of about 91%. This makes the DMTJ-based STT-MRAM cell a good eNVM alternative 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 realized in machine learning (ML) applications including image processing/classification/recognition, natural language processing, and visual intelligence [1], [2]. Thanks to features such as small cell area footprint, short programming time, and good endurance and data retention [3], there is an increasing interest in the field of neuro-inspired computing with 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 architecture could be competent in terms of energy and latency, it mainly suffers from low online learning accuracy. To deal with this, digital synapse based architecture has been widely considered [3], [4]. As potential eNVM candidate for digital synapse devices, STT-MRAM cell offers low operating voltage, high-speed operation, high density, relatively large endurance, low fabrication cost, low power consumption, and scability [5]–[7]. Typically, STT-MRAM based DNN implementations relies on conventional single-barrier MTJ (SMTJ) devices. However, it is requires high writing currents, limiting the overall energy-efficiency and latency of DNN. To solve this, a solution is to use double-barrier MTJ (DMTJ) with two reference pinned layers, enabling high-speed operation, low power consumption, and energy-efficient switching process [7]–[9].

To evaluate the impact of DMTJ-based STT-MRAM cell on DNN, we use Cadence-Virtuoso environment for circuit-level simulations, along with the multilayer perceptron (MLP) + NeuroSimV3.0 simulator computing-in-memory (CiM) based neural network accelerator [4]. NeuroSim is used to support a 2-layer MLP neural network to benchmark DNN architecture, relied on SMTJ-based and DMTJ-based digital synapse devices, in online learning and offline classification with MNIST handwritten dataset.

Our results points 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%, making the DMTJ-based STT-MRAM cell a promising candidate for digital synapse in neuro-inspired computing.

This work 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. Section IV concludes this work.

## II. Simulation Framework – MLP+NeuroSimV3.0

NeuroSim simulator estimates the algorithm-level performance by emulating the online learning and offline classification scenario with MNIST handwritten dataset in a 2-layer MLP neural network [4], [10]–[12]. As shown in Fig. 1, its framework consists from device and bitcell levels to memory architecture and algorithm levels. The input parameters of the simulation tool include memory types, non-ideal device parameters, transistor technology nodes, network topology and array size, training dataset and traces, etc [1]. The outputs

[1]For the full list of input parameters/variables, the reader is referred to [4].

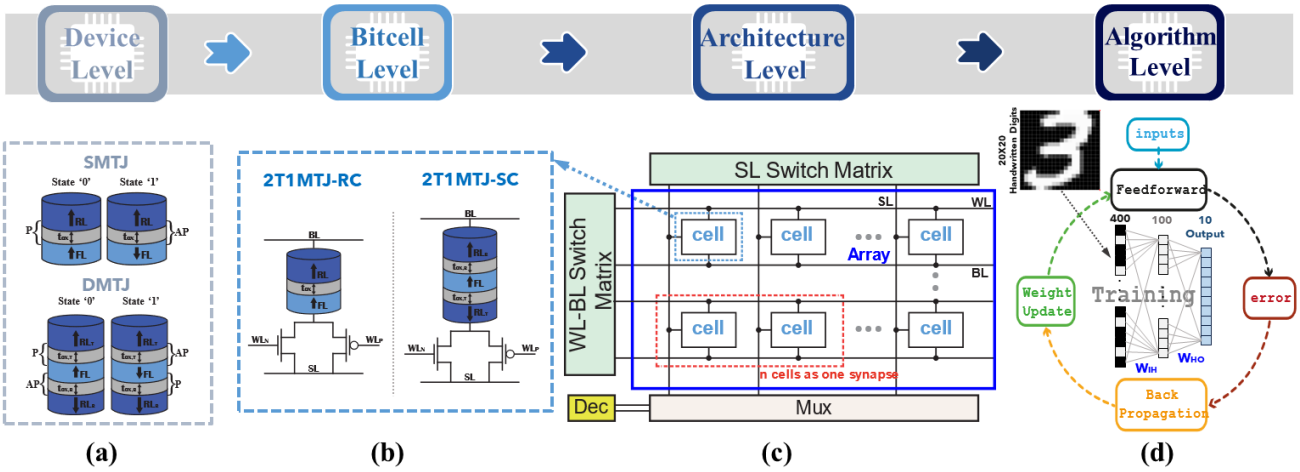


**Fig. 1:** Overview of NeuroSim framework from Device to Algorithm-level, (a) STMJ and DMTJdevice, (b)SMTJ-based and DMTJ-based bitcell configurations, (c) Circuit block diagram of digital eNVM synaptic core, (d) Training flow of Neural Network, the MNIST images are crooped and encoded into black and white data for simplification on hardware implementation.

of the simulator include: (1) the memory architecture-level performance metrics, such as area, latency, dynamic energy, leakage power consumption, and (2) algorithm-level learning accuracy in run-time. As for the design options of digital synaptic arrays, SRAM or eNVM can be used.

### A. Device Level

At the device-level, as shown in Fig. 1 (a), we consider STT-SMTJ/DMTJ devices, whose main parameters are listed in Table I. The STT-MTJs are described through Verilog-A based compact models [13], [14], calibrated with experimental data reported in [15].

The MTJ consists of two types of ferromagnetic (FM) layers, one with fixed magnetization direction called reference layer (RL), and other with a free magnetization direction named as free layer (FL), which can be changed by applying a switching current greater than the critical switching current of the device [7]. Based on the relative magnetization direction of the FL and RL, it can reside in one of two stable states: parallel (P) or antiparallel (AP).

*1) SMTJ-based:* It consists of RL and FL separated by a thin MgO oxide barrier ($t_{ox}$).

If two ferromagnetic layers have the same magnetization directions, i.e., RL and FL in P, the resistance of the MTJ is low, 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, indicating a "1" state [7].

*2) DMTJ-based:* 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 [7].

**TABLE I:** SMTJ and DMTJ device parameters [7].

| Parameter | Units | Value |
|---|---|---|
| Diameter (d)[a] | nm | 28 |
| Saturation magnetization (Ms)[a] | A/m | $1000 \times 10^3$ |
| Magnetic damping ($\alpha$)[a] | - | 0.025 |
| Spin-polarization factor ($\eta$)[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

### B. Bitcell- to Memory Architecture-Level

Fig. 1(b) shows the considered SMTJ-based and DMTJ-based bitcell configurations designed in a 28nm 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 [7], [16], these are the best write energy-efficient bitcell configurations.

At the architecture level, two synaptic cores of 2-layer MLP are considered. Each synaptic core is a computation unit specifically designed for weighted sum and weight update [4], [10]. Among the available design options for the synaptic cores, we considered the digital eNVM based on pseudo-crossbar array.

### 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 [4].

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. 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(d), by default, the network topology contains 400 neurons (20×20

**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 |

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 0.8V 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 $t_{ox}$, i.e., $t_{ox,SMTJ} = t_{ox,T,DMTJ} = 0.85nm$. 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 conventional SMTJ-based bitcell.

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

The number of images in MNIST dataset during training and testing are 8000 and 1000, respectively, with a total number of epoch (i.e., iterations) of 15, 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 Figs. 2 and 3. We can observe that the weighted sum and weight update operations associated to the DMTJ-based eNVM cell achieves the highest accuracy much faster as compared to the SMTJ-based counterpart, while at the same

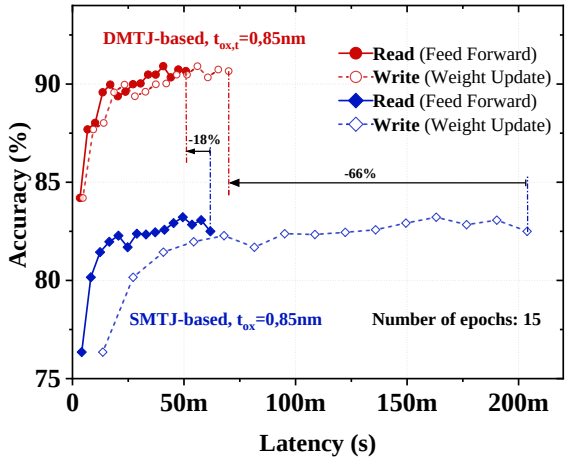


**Fig. 2:** Trace of Latency in feed forward and weight update

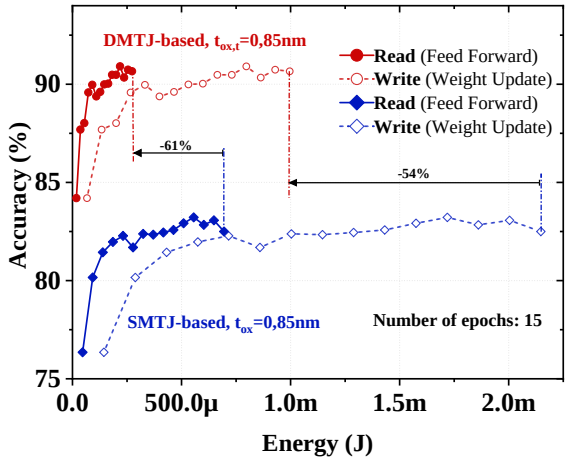


**Fig. 3:** Trace of Energy in feed forward and weight update.

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

As for the weight update operation, the delta latency between each epoch is 14ms and 4.7ms, respectively. This can be explained due to the larger pulse width. As compared with the SMTJ-based alternative, the DMTJ-based solution 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. 3. We observed that the DMTJ-based cell exhibits lower energy consumption as compared to the SMTJ-based alternative, owing to its reduced bitcell read/write energy. The results showed an upgrade of about 61% and 54% during feed forward and weight update, respectively.

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

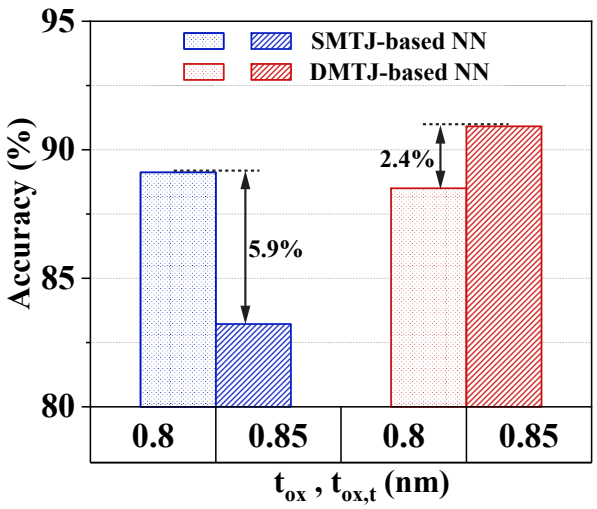


**Fig. 4:** Learning accuracy versus oxide thickness ($t_{ox}$ or $t_{ox,t}$) for SMTJ- and DMTJ-based neural networks.

83%. This is because of the lower ON/OFF ratio of the DMTJ-based bitcells. Indeed, the cause of degradation in terms of learning accuracy is attributed to the devices' poor ON/OFF ratio [17].

In addition, we showed the estimation of area and leakage power consumption obtained from NeuroSim. For the area, the total footprint for both SMTJ-based and DMTJ-based alternatives is $7.881 \times 10^{-9} m^2$ and $5.311 \times 10^{-9} m^2$, 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 device's conductance should be sufficiently large, i.e., the lowest conductance state (OFF-state) should be low enough to represent the zero weight in the algorithm [17]. 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 $t_{ox,SMTJ} = t_{ox,T} = 0.80nm$, the conductance ON/OFF ratio for SMTJ- and DMTJ-based cell are 1.91 and 1.88, respectively. The decrease of the ON/OFF conductance ratio in the DMTJ-based cell can be explained due to the presence of the second oxide barrier. Therefore, the accuracy for SMTJ-based cell increases by 5.9%, while DMTJ-based cell decreases by 2.5%, see Fig. 4.

## IV. Conclusion

In this work, we have exploited 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. The corresponding results show that, at parity of TMR and oxide thickness, as compared to the conventional SMTJ-based alternative, 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 size. This occurs while also achieving an accuracy closed to 91% when running the neural network with the MNIST dataset.

Compared with other architectures based on digital emerging non-volatile memory (eNVM) synaptic cores, we suggest 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," *Advances in neural information processing systems*, vol. 25, 2012.

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

[3] Y. Luo, X. Peng, and S. Yu, "MLP+ NeuroSimV3. 0: Improving on-chip learning performance with device to algorithm optimizations," in *Proceedings of the International Conference on Neuromorphic Systems*, 2019, pp. 1–7.

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

[5] N. Xu *et al.*, "STT-MRAM design technology co-optimization for hardware neural networks," in *IEEE IEDM*, 2018, pp. 15–3.

[6] K. Zhang *et al.*, "High On/Off Ratio Spintronic Multi-Level Memory Unit for Deep Neural Network," *Advanced Science*, p. 2103357, 2022.

[7] E. Garzon, R. De Rose, F. Crupi, L. Trojman, G. Finocchio, M. Carpentieri, and M. Lanuzza, "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, 2020.

[8] G. Hu *et al.*, "Low-current spin transfer torque MRAM," in *International Symposium on VLSI Design, Automation and Test*, 2017, pp. 1–2.

[9] E. Garzón, M. Lanuzza, R. Taco, and S. Strangio, "Ultralow voltage FinFET-versus TFET-based STT-MRAM cells for IoT applications," *Electronics*, vol. 10, no. 15, p. 1756, 2021.

[10] P.-Y. Chen *et al.*, "NeuroSim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 12, pp. 3067–3080, 2018.

[11] A. Lu, X. Peng, W. Li, H. Jiang, and S. Yu, "NeuroSim validation with 40nm RRAM compute-in-memory macro," in *IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 2021, pp. 1–4.

[12] 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 Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 40, no. 11, pp. 2306–2319, 2020.

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

[14] R. De Rose, M. d'Aquino *et al.*, "Compact modeling of perpendicular STT-MTJs with double reference layers," *IEEE Transactions on Nanotechnology*, vol. 18, pp. 1063–1070, 2019.

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

[16] E. Garzón, R. De Rose, F. Crupi, L. Trojman, G. Finocchio, M. Carpentieri, and M. Lanuzza, "Exploiting Double-Barrier MTJs for Energy-Efficient Nanoscaled STT-MRAMs," in *2019 16th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*, 2019, pp. 85–88.

[17] 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, 2018.