

An Effective SNN Macro with Real-Time STDP and Dynamic LIF Model Based on Thermally Interplayed Spin-Orbit Torque MTJ

Changyu Li*, Linjun Jiang*, Liangchen Li, Dehang Zhu, Junda Zhao, Hongxi Liu, Wang Kang†, Wenlong Cai†, He Zhang†, Weisheng Zhao

School of Integrated Circuit Science and Engineering, Beihang University, China

Truth Memory Corporation, Beijing, China

Changyu Li and Linjun Jiang contributed equally to this work

†: corresponding authors

{wang.kang, caiwenlong1993, zhanghe}@buaa.edu.cn

Abstract—Spiking neural networks (SNNs) have emerged as a promising paradigm for effective event-driven computation. However, CMOS-based SNN designs are limited by power consumption and complexity, while nonvolatile memory (NVM)-based SNN designs often lack biological characteristics and require active capacitive circuits to emulate neuronal dynamics. In this paper, we propose a thermally interplayed spin-orbit torque magnetic tunnel junction (TI-MTJ) macro that integrates core SNN functionalities. Our neuron array autonomously achieves leaky integrate-and-fire (LIF) model within the TI-MTJ device, thus improving power efficiency and simplifying circuit structure. Additionally, the proposed synaptic array provides adaptive in-situ responses based on a simplified spike-timing-dependent plasticity (STDP) rule. To enhance biological plausibility, our macro incorporates real-time spike monitoring and inhibition mechanisms. A comprehensive device-circuit-algorithm co-optimization framework validates the high performance of the TI-MTJ macro, achieving a synaptic energy consumption of 6.07fJ per spike, an inference accuracy of 97.76% on the MNIST dataset, and an energy efficiency of 22.8TOPS/W.

Keywords—MTJ, STDP, LIF model, thermally interplayed, SNN

I. INTRODUCTION

Biological brains exhibit an extraordinary capability to process real-time information effectively [1]. Inspired by brain architecture, spiking neural networks (SNNs) enable real-time problem-solving with remarkably low power consumption [2]. Over the past two decades, CMOS technology has been used to emulate the behavior of spiking neurons and synapses [3–5]. However, fundamental differences between the physical mechanisms of CMOS devices and biological components necessitate complex circuitry to simulate even the simplest neural functions [6]. Additionally, the implementation of unsupervised learning rules like spike-timing-dependent plasticity (STDP) in CMOS circuits requires a continuous power supply, which limits the flexibility to simulate SNN learning mechanisms.

To address these challenges, emerging nonvolatile memory (NVM) devices, including the memristor, phase-change memory (PCM), and magnetic tunnel junction (MTJ), offer potential for enhancing biological fidelity in SNNs, such as the leaky integrate-and-fire (LIF) model [7–9]. Among these devices, although the memristor mimics synapses with low power consumption, it suffers from the inherent problem of sneak path currents in crossbar array [10]. PCM provides multi-state storage for representing synaptic weights but suffers from high write energy and slow speed [11]. Spin-transfer torque magnetic tunnel junction (STT-MTJ) offers

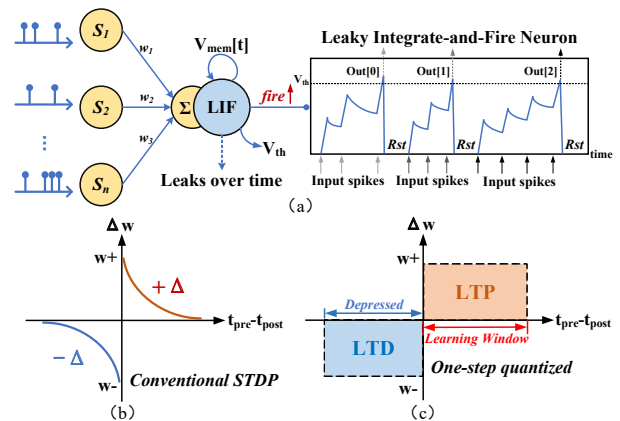


Fig. 1. (a) LIF neuromorphic model in SNN. (b) STDP learning rule. (c) Simplified STDP learning rule.

non-volatility for stable weight storage but faces the challenges of high write power and scalability [12]. Conversely, spin-orbit torque magnetic tunnel junction (SOT-MTJ) offers high speed, low energy consumption, non-volatility, and excellent endurance, making it a promising candidate for neuromorphic systems [13].

However, SOT-MTJ still faces challenges in directly emulating core neural behaviors at the device level for SNNs. While prior work [14] explored using SOT-MTJs for both synapses and neurons, these designs typically relied on actively charging CMOS capacitors to implement neuron functionalities including the LIF model and STDP rules, leading to increased area and power consumption. Furthermore, the lack of real-time spike monitoring in prior designs also severely impeded system-level SNN implementation. For example, the work in [2] not only relied on extra clock cycles for MTJ state control, but also lacked real-time neuron firing monitoring. As a result, it struggled to support essential SNN mechanisms such as refractory periods and lateral inhibition, especially under conditions of concurrent neuronal firing.

In this work, we develop and experimentally validate a novel thermally interplayed spin-orbit torque magnetic tunnel junction (TI-MTJ) device. The macro achieves 97.76% inference accuracy on the MNIST dataset, and an energy efficiency of 22.8TOPS/W, confirming its functionalities. In comparison with other designs, the macro shows significant advantages in terms of energy efficiency, inference latency, and area compactness. The primary contributions include:

- **STDP-Enabled TI-MTJ Synaptic Array:** A TI-MTJ synaptic array is developed that implements a

simplified STDP rule, enabling real-time and self-adaptive synaptic weight updates with high energy efficiency.

- **Cap-Free TI-MTJ LIF Neuron:** A LIF neuron is realized using the intrinsic thermal dynamics of the TI-MTJ, eliminating the need for explicit capacitors and reducing area overhead. The monitoring circuit allows for real-time tracking of firing event.
- **Multiple Inhibitory Mechanisms:** Our SNN macro incorporates refractory period and winner-takes-all (WTA) lateral inhibition functions, which enhance biological plausibility and enable more accurate emulation for neuromorphic applications.

II. PRELIMINARIES

A. LIF Neuron Model and STDP Learning Rule

In the SNNs, the LIF neuron model and STDP unsupervised learning rule are essential to mimic biologically plausible computation and learning behaviors.

The LIF model emulates the spiking mechanism of biological neurons by capturing the temporal dynamics of the membrane potential [15]. As shown in Fig. 1(a), the membrane potential V_{mem} of the neuron increases upon the arrival of an input spike. It subsequently decays over time due to a leakage current until the next spike is received. Once V_{mem} exceeds a predefined threshold V_{th} , the neuron fires an output spike and V_{mem} is reset. The LIF model can be expressed as:

$$C_{mem} \frac{dV_{mem}}{dt} = -\frac{V_{mem}}{R_{mem}} + \sum_j \delta(t - t_j)w_j \quad (1)$$

where C_{mem} and R_{mem} are the membrane capacitance and resistance respectively. $\delta(t - t_j)$ represents the arrival of a spike at time t_j and w_j represents the synaptic weight.

On the other hand, the STDP learning rule modulates synaptic weights based on the relative timing between presynaptic and postsynaptic spikes [16]. As shown in Fig. 1(b), when a presynaptic spike precedes a postsynaptic spike, the synaptic connection strengthens, known as long-term potentiation (LTP). Conversely, the synaptic weight is weakened, known as long-term depression (LTD). Fig. 1(c) shows a simplified STDP rule [17], which can be expressed as:

$$w = \begin{cases} +1 \text{ level,} & \text{when LTP} \\ -1 \text{ level,} & \text{when LTD} \end{cases} \quad (2)$$

This simplified STDP learning rule enables effective unsupervised synaptic weight updates in hardware, while significantly improving the feasibility and energy efficiency of NVM-based synaptic implementations.

B. TI-MTJ for Neurons and Synapses in SNN

The proposed TI-MTJ is chosen as the backbone for both artificial neurons and synapses. The device stack consists of W(3.5nm)/CoFeB(1.4nm)/MgO(1.3nm)/CoFeB(2.4nm)/Ru(0.8nm)/CoFe(2.5nm)/MnIr(7.5nm), as shown in Fig. 2(a). This MTJ exhibits two stable magnetic states, which can be controlled by SOT generated by the heavy metal layer (W).

A key feature of the TI-MTJ device is that the current density required to reverse the magnetization is temperature-dependent. When a current pulse I_{spike} is applied, Joule

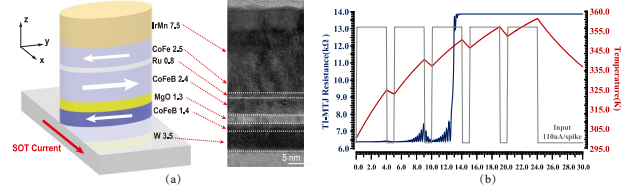


Fig. 2. Thermally interplayed TI-MTJ for emulating artificial neurons and synapses. (a) Schematic and cross-sectional TEM images of the TI-MTJ stack. (b) Complete LIF dynamics driven by SOT current pulse.

heating raises the temperature of the MTJ, thereby reducing the critical switching current I_{th} . This process mimics the integration of V_{mem} in biological neurons. When I_{th} falls below I_{spike} , the magnetization is switched, indicating a firing event. The heating dynamics can be described as:

$$T_{post} = \left(\frac{J_i^2 \alpha}{K} + T_0 \right) (1 - \alpha) + T_{pre} \alpha \quad (3)$$

where $\alpha = e^{-\tau/\tau_0}$, τ is the duration of the applied current and τ_0 is the characteristic thermal time constant of the device. Here, J_i denotes the input current density, K represents a combined specific heat and heat generation factor, T_0 is the room temperature, T_{pre} is the initial temperature before current application, and T_{post} is the temperature after current cessation. Simultaneously, the natural cooling process due to heat dissipation can be described by:

$$T_{post} = T_0(1 - \alpha) + T_{pre} \alpha \quad (4)$$

This cooling phase emulates the leakage behavior of V_{mem} in biological neurons. Note that, potential thermal crosstalk between adjacent devices in a high-density SNN macro must be considered. In this work, thermal crosstalk is effectively suppressed through several approaches. Adjacent TI-MTJs are physically separated and electrically isolated by dielectric layers, which significantly attenuates lateral heat propagation. Heat is primarily dissipated vertically through the substrate. Furthermore, the inherent refractory period of SNN provides sufficient cooling time for the TI-MTJ before the next firing event. The relatively large size of MOS transistors along with the highly localized nature of the SOT current further minimizes thermal coupling between devices.

Owing to the sparse firing activity typical in SNNs, most neurons experience a gradual decrease in TI-MTJ temperature following a spike and V_{mem} falls below V_{th} . The complete LIF emulation process of the TI-MTJ is depicted in Fig. 2(b). With a subsequent write-back current, the free layer gradually recovers to its initial state, completing the neuronal cycle.

C. Overall Architecture of the Proposed TI-MTJ Macro

Fig. 3 illustrates the overall architecture of the proposed thermally interplayed TI-MTJ macro. During SNN operations, input spikes are delivered to the synaptic array to perform analog vector-matrix multiplication and accumulation. The resulting output currents are then integrated by the TI-MTJ neuron array, which leverages its intrinsic thermal dynamics to execute the LIF model and enables real-time spike monitoring. When a neuron fires, the multifunctional inhibitory layer applies both refractory periods and WTA lateral inhibition across relevant neurons and synapses, simultaneously generating a feedback signal. This signal is processed by the write-back driver to update synaptic weights based on a simplified STDP rule. The TI-MTJ macro achieves essential SNN functionalities without external weight memory

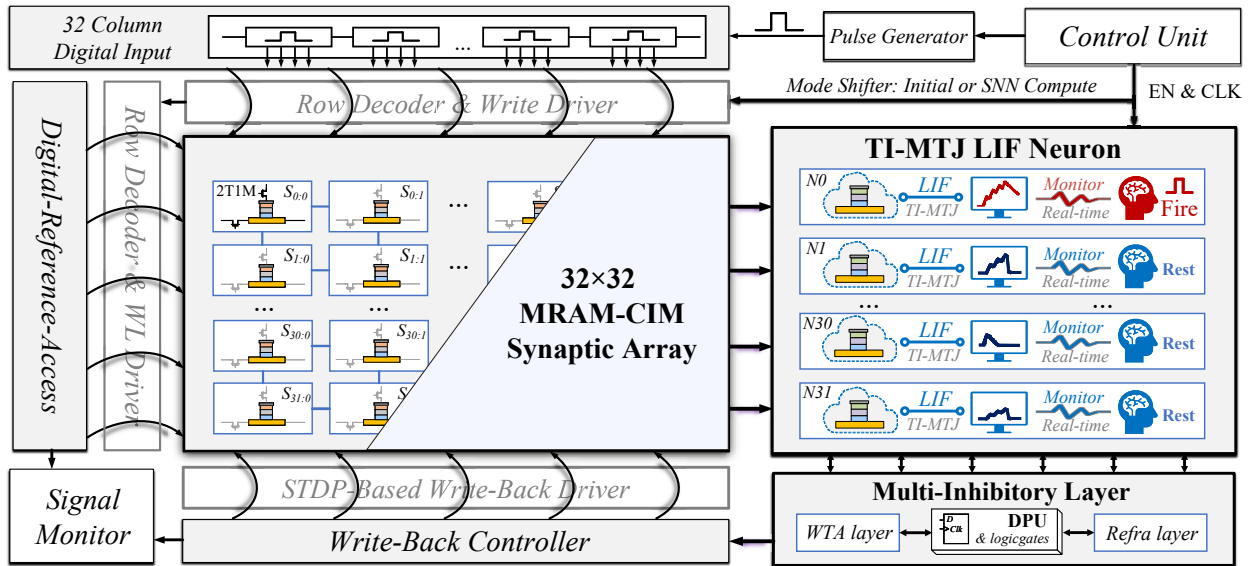


Fig. 3. Overall architecture of the proposed TI-MTJ macro. The macro mainly consists of a 32×32 TI-MTJ synaptic array for SNN computation and STDP-based weight updates, a 32-row TI-MTJ LIF neuron array for real-time spike monitoring, and a multifunctional inhibitory layer for WTA lateral inhibition and refractory period control.

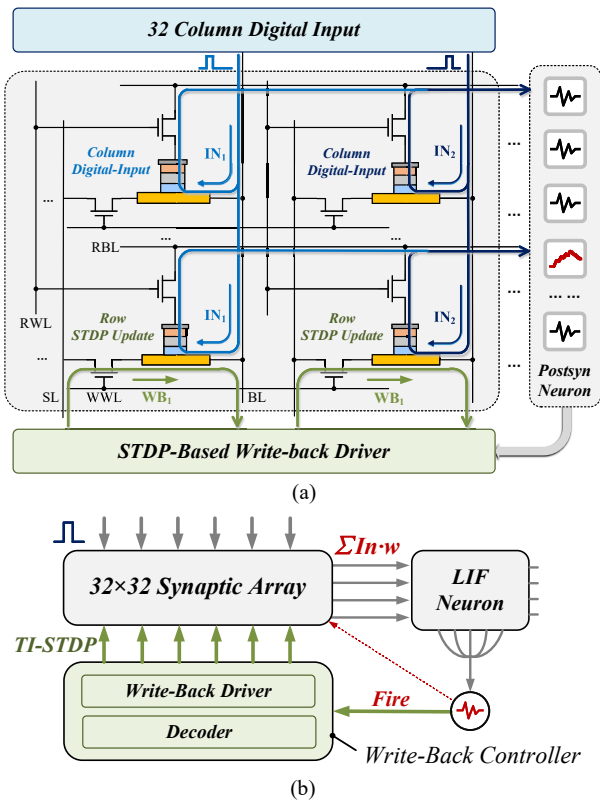


Fig. 4. (a) Principle and (b) workflow of the TI-MTJ synaptic array based on thermal dynamics and STDP-based write-back mechanism.

or explicit capacitive components, significantly improving area and energy efficiency in neuromorphic hardware.

III. PROPOSED TI-MTJ SNN MACRO

A. TI-MTJ Synaptic Array for SNN and STDP

As shown in Fig. 4, the proposed synaptic array employs a 2T-1MTJ cell structure, consisting of one TI-MTJ device and two access transistors. The TI-MTJ synaptic array works in two modes: SNN operation and simplified STDP-based write-back operation. Mode switching is controlled by the

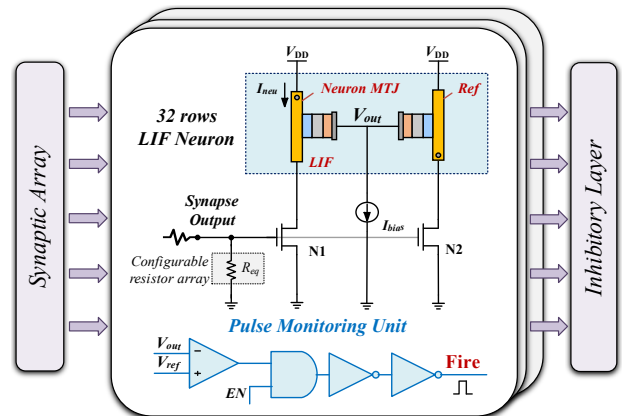


Fig. 5. Schematic of the TI-MTJ LIF neuron.

synaptic array state and the activation of read word line (RWL) and write word line (WWL).

(1) SNN Operation: In this mode, the WWLs remain inactive while the RWLs are enabled. Digital input pulses are applied through the bit lines (BLs) to the TI-MTJ synaptic array. These pulses pass through the activated read path (e.g., IN_1 and IN_2 in Fig. 4), modulated by the conductance (weight) of each synapse. The resulting currents are integrated on the corresponding read bit line (RBL), generating a weighted postsynaptic voltage pulse that is delivered to the connected neuron. Consequently, the stimulus strength presented to the postsynaptic neuron is determined by both the presynaptic digital spike activity and the synaptic weights.

(2) Write-Back Operation: This mode is activated upon postsynaptic neuron firing. Under this condition, SNN computation is suspended and the synaptic array works in the write-back operation mode. Triggered by the firing event, the STDP-based write-back driver generates corresponding signals. The global RWLs are disabled while the WWL connected to firing neuron is active. Concurrently, the driver applies a write-back current pulse to the corresponding row of the TI-MTJ synaptic array to implement spike-timing-dependent weight updates.

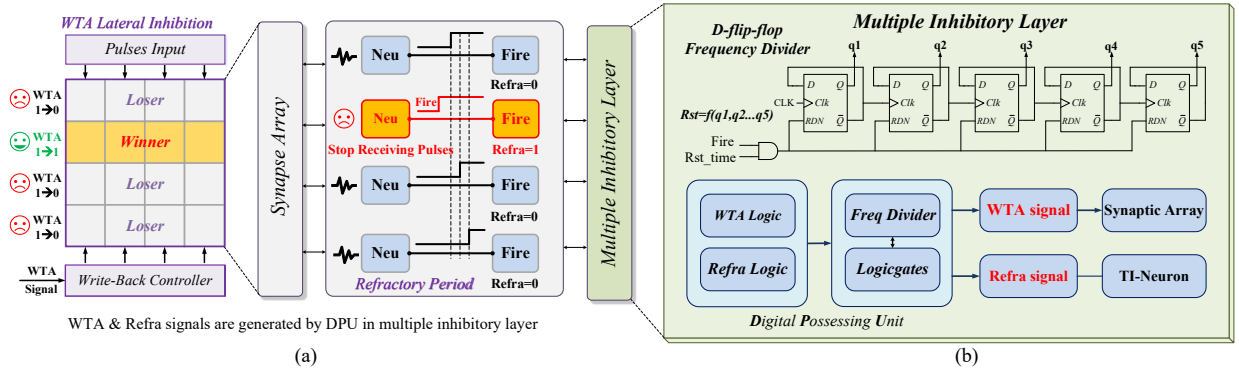


Fig. 6. (a) Architecture framework for refractory period and lateral inhibition. (b) Circuit schematic of the multifunctional inhibitory layer.

(3) **Simplified STDP Rule:** The proposed STDP learning rule leverages the intrinsic thermal properties of the TI-MTJ. When a neuron fires within a time window, the write-back driver applies a current pulse to the corresponding synaptic row. As analyzed in Section II-B, this induces cumulative Joule heating in the TI-MTJs, lowering I_{th} . LTP occurs when I_{th} falls below I_{spike} , triggering a switch in magnetic state. This process effectively correlates postsynaptic activity with recent presynaptic events, emulating a key aspect of STDP. Conversely, low postsynaptic firing frequency results in insufficient heating, leaving synaptic weights unchanged. Additionally, the natural cooling of the device allows I_{th} to recover to its baseline, further mimicking the behavior of LTD under uncorrelated activity. This thermal behavior, combined with the write-back mechanism, implements an effective unsupervised STDP rule. The weight does not automatically revert due to short-term cooling unless being reset by the write-back driver. This nonvolatile characteristic ensures the stability and reliability of learned synaptic weights in the SNN macro.

Thanks to the thermal dynamics of TI-MTJs, the synaptic array supports both SNN computation and unsupervised STDP learning rule. Synaptic weights are encoded as thermal information within the device, eliminating the need for additional charge/discharge circuits. As a result, STDP learning latency is bounded only by the write delay of the TI-MTJ. In brief, the proposed TI-MTJ synaptic array enables real-time, low-power, low-latency, and compact unsupervised STDP learning, offering significant advantages for SNN-based neuromorphic computation.

B. TI-MTJ Neuron for LIF Model

Fig. 5 shows the schematic of the TI-MTJ LIF neuron array, consisting of an input voltage divider, a core TI-MTJ LIF circuit, and a pulse monitoring circuit. The voltage divider receives the aggregated weighted outputs from the synaptic array and applies the resulting voltage to the gate of N_1/N_2 . The equivalent resistance R_{eq} of voltage divider is generated by an adjustable resistor array, enabling tunable integration sensitivity.

The core LIF circuit employs two TI-MTJ devices configured in opposite directions on the heavy-metal layer, ensuring symmetric thermal conditions while exhibiting opposite switching responses. N_2 is identically sized to mirror the current through N_1 . When the synaptic input activates N_1 , the current mirror provides an identical write current I_{neu} to both the neuron MTJ and the reference MTJ. Joule heating from I_{neu} raises the temperature of the neuron MTJ, reducing its switching threshold. Once the threshold is exceeded, the

neuron MTJ is switched, emulating a firing event. Conversely, the reference MTJ remains unchanged and provides a stable reference point benefiting from the opposite magnetic configuration.

The pulse monitoring circuit consists of a dynamic comparator and a bias current source I_{bias} . It detects firing events by continuously sensing the resistance change of the neuron MTJ in real time. When the neuron MTJ switches from a high-resistance to a low-resistance state, the voltage at node V_{out} changes by $\Delta V_{out} = I_{bias}\Delta R$. This voltage shift is captured by the comparator in real time, enabling immediate spike monitoring.

C. Refractory Period and WTA Lateral Inhibition

Based on the TI-MTJ, the proposed neuron array supports real-time LIF computing and spike monitoring. This design eliminates conventional integration capacitors, reducing area and power overhead. It also supports highly efficient and scalable SNN operations.

As shown in Fig. 6(a), the unsupervised learning process requires a refractory period after each neuronal firing to allow sufficient time for synaptic weight updates and thermal recovery. Fig. 6(b) shows a digitally-controlled refractory period. Upon the detection of a spike, the output spike pulse triggers a D-flip-flop-based timing circuit to generate the refractory period signals. During this interval, the fired neuron is isolated from incoming outputs of the synaptic array. Once the refractory period elapses, the neuron is reenabed and ready for the next LIF cycle.

WTA lateral inhibition is another essential feature emulated in this work. As shown in Fig. 6(a), WTA mechanism improves feature extraction by maintaining sparse activation across neurons. It activates only the most responsive neuron while suppresses others in the same layer. The digital delay circuit to realize global WTA lateral inhibition is the same as that for refractory control. When a winning neuron fires, it immediately drives the WTA signals of all other neurons low, inhibiting their activity during the inhibition window. This effectively implements global WTA lateral inhibition across the SNN network.

IV. EXPERIMENTAL RESULT

This section presents the experimental evaluation of the proposed TI-MTJ SNN macro, fabricated in a 110 nm technology node. A comprehensive cross-level simulation framework was employed to validate functionalities across device behavior, circuit level, and network algorithm.

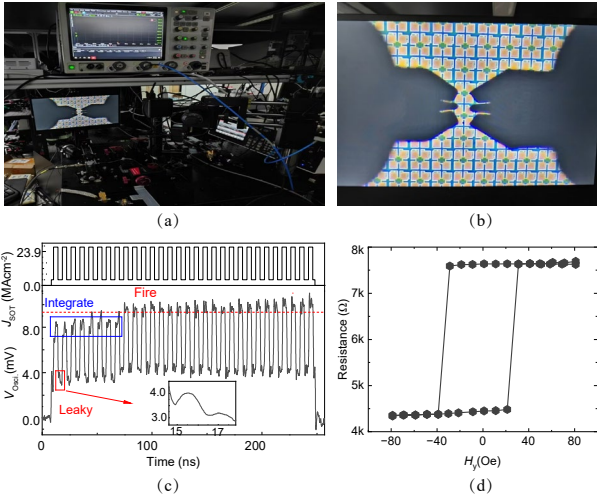


Fig. 7. Device level characterization of the TI-MTJ. (a) Experimental test setup. (b) Electron microscopy image, (c) thermal response, and (d) measured hysteresis loop of the fabricated TI-MTJ device.

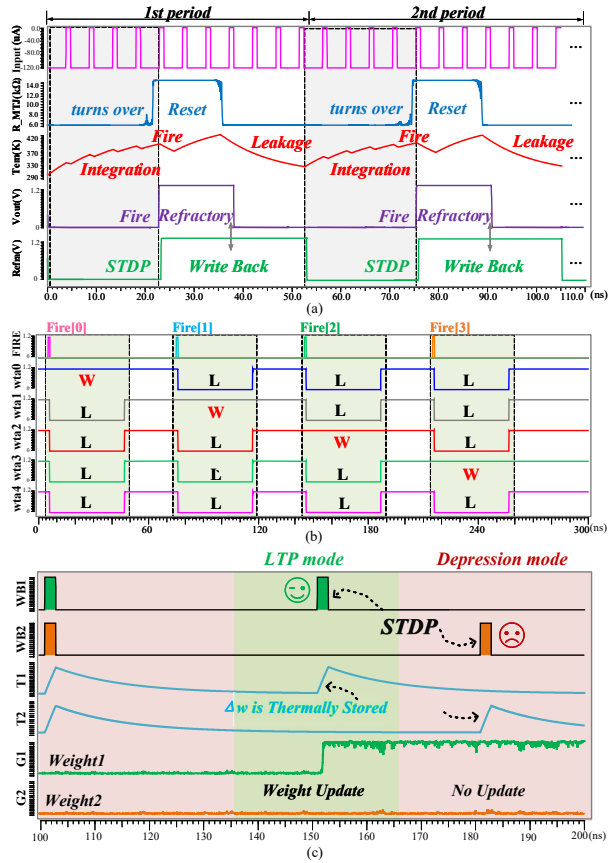


Fig. 8. Functional verification of the TI-MTJ SNN macro. (a) LIF operation with post-firing refractory and write-back control. (b) WTA lateral inhibition mechanism for global suppression. (c) Simplified STDP-based synaptic weight update through thermal accumulation.

A. Function Validation

At the device level, we modeled and characterized the performance of the TI-MTJ device. Fig. 7(a) shows the experimental test setup for device characterization and Fig. 7(b) shows a scanning electron microscopy (SEM) image of the fabricated TI-MTJ device. The measured thermal response of the device under current pulsing is shown in Fig. 7(c). A state transition occurs when the MTJ temperature exceeds the threshold. The behavior is dependent on intrinsic device

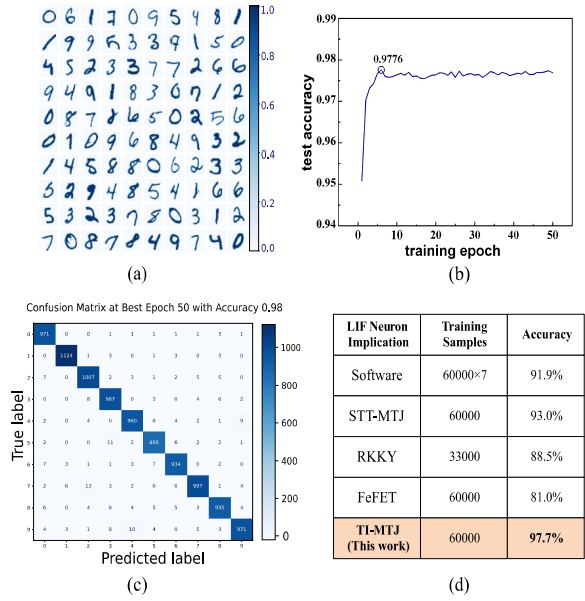


Fig. 9. Accuracy evaluation of the proposed SNN macro. (a) Diagram of the MNIST dataset. (b) Training process, achieving a peak accuracy of 97.76%. (c) Confusion matrix at the optimal epoch. (d) Comparative

properties and external input parameters such as pulse amplitude and frequency. The measured hysteresis loop of the TI-MTJ is shown in Fig. 7(d). By carefully tuning the input signals and physical parameters, the TI-MTJ can serve as a synapse and a neuron, providing a versatile and biologically plausible foundation for neuromorphic hardware.

At the circuit level, we performed simulations to validate the functionalities for SNN. Fig. 8(a) demonstrates the LIF behavior of the neuron. The temperature of the neuron TI-MTJ rises under input current, emulating membrane integration. And it decays naturally in the absence of input, mimicking biological leakage. Upon reaching the threshold temperature, a spike is generated while the refractory (Refra) and write-back signals are asserted to initiate the post-firing reset and weight update processes. During this phase, the RWL is disabled, isolating the neuron and allowing it to cool. The temporal mismatch between “Refra” and “Write Back” signals stems from delays in the digital processing unit.

Fig. 8(b) shows the WTA lateral inhibition mechanism. Initially, all neurons are active (“W”). When a neuron fires (e.g., Fire[0]), it is recorded as the “winner” and its WTA signal (WTA[0]) is set high, whereas all others (WTA[1-4]) are set to GND, recording them as “losers” (“L”). This lateral inhibition persists for an adjusted interval determined by the digital processing unit, after which all neurons are reset to compete in the next cycle.

Fig. 8(c) illustrates the synaptic weight update mechanism. The mechanism employs a simplified STDP rule, where synaptic weights are thermally encoded within the TI-MTJ devices. When an LIF neuron fires, the write-back driver applies a current pulse (WBI/WB2) to the corresponding row of synaptic devices. Each write-back pulse induces Joule heating (T1/T2), raising the synapse temperature and increasing its probability of switching. In the absence of write-back pulses, the device cools naturally through thermal dissipation. The conductance curves (G1/G2) reflect this interplay between pulsed heating and natural cooling. For the highly active neuron1, frequent firing within the LTP window causes cumulative heating sufficient to temporarily reduce I_{th}

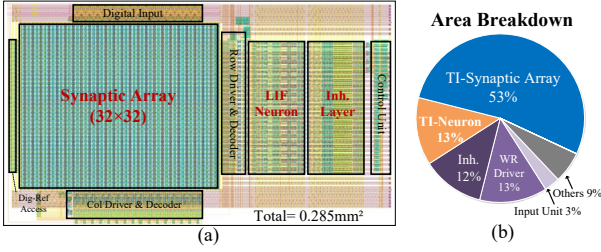


Fig. 10. (a) Chip layout in 110 nm. (b) Area breakdown.

TABLE I. CHIP SPECIFICATIONS

Technology node	110nm
Apply voltage	1.2V
Array size	32 × 32
Synapse/Neuron	TI-MTJ
Dataset	MNIST
Accuracy	97.76%
Read latency	1.9ns
STDP latency	10.5ns
Energy/Synapse	6.07fJ
Energy efficiency	22.8TOPS/W

below I_{spike} , triggering the magnetization switching and synaptic weight update. In contrast, the less active neuron2 remains unchanged. Once switched, the TI-MTJ synapse retains its state nonvolatility and does not revert the state due to natural cooling, ensuring long-term retention of learned associations.

B. Accuracy Performance

As shown in Fig. 9(a), we employed the MNIST dataset to train and evaluate the network, consisting of 60,000 training and 10,000 test samples. The network architecture includes 784 input neurons, 196 hidden neurons, and 10 output neurons, with full connectivity and training via backpropagation. Fig. 9(b) shows that a test accuracy of 97.76% was achieved after 50 training epochs. The confusion matrix in Fig. 9(c) summarizes the classification performance across all 10 categories using the 10,000 test samples at the optimal epoch. The results indicate that most images were correctly classified within 50 epochs, demonstrating the capability of the proposed macro in handling complex classification tasks. As shown in Fig. 9(d), the macro achieves higher accuracy compared to other designs.

C. Energy and Area Breakdown

Fig. 10(a) shows the layout and Fig. 10(b) shows the area breakdown of the proposed TI-MTJ macro. The macro features a highly compact architecture, with the TI-MTJ synaptic array occupying 53% of the total area. The neuron array and read/write drivers account for only 13%. This area compactness stems from the elimination of conventional capacitor-based CMOS neuron circuits, as the LIF behavior is intrinsically implemented within the TI-MTJ devices. Furthermore, a simplified STDP learning rule is embedded locally within the synaptic array, requiring only a minimal write-back driver. By integrating memory and computation at the device level, the work avoids external storage or separate processing modules, significantly reducing area overhead and improving scalability for large-scale SNN implementation.

TABLE II. PERFORMANCE COMPARISON

	[22]	[23]	[24]	[25]	This work
Synapse	Memristor	S-MTJ	MTJ-HM	STI-MTJ	TI-MTJ
Neuron	Cap-based	Digital	MTJ-HM	Cap-based	TI-MTJ
Technology	90nm	28nm	45nm	65nm	110nm
Model	LIF	LIF	LIF	LIF	LIF
STDP	✓	×	✓ (cap-based)	×	✓
Δw stored	Subarray*	Latch*	Switching probability	Subarray*	Thermal response
Energy/Syn.	20fJ	8.87fJ	29.23pJ	5.48fJ	6.07fJ
Dataset	MNIST	MNIST	MNIST	MNIST/CIFAR-10	MNIST
Accuracy	95%	91.5%	< 80%	97.9%	97.76%

*: Additional memory subarrays or circuits are required to store Δw .

Table I summarizes the chip specification of the TI-MTJ macro. The macro achieves an energy efficiency of 22.8TOPS/W at 1.2V supply voltage. Each TI-MTJ synaptic operation consumes 6.07fJ per spike during inference and the network reaches 97.76% accuracy on the MNIST dataset. The TI-MTJ devices exhibit significantly faster switching compared to conventional SOT devices, yielding a read latency of 1.9 ns and an STDP update latency of 10.5 ns. These characteristics contribute to lower overall latency, benefiting high-speed SNN implementation.

Table II compares the performance of the proposed macro with state-of-the-art SNN designs using the same dataset. The results highlight a notable advantage in energy efficiency per synaptic operation, while achieving competitive inference accuracy. This improvement is mainly due to the intrinsic thermal dynamics of TI-MTJ devices, which natively support in-memory STDP learning by encoding synaptic weight changes (Δw) through localized temperature accumulation. As a result, the mechanism eliminates the need for auxiliary storage subarrays, latches, or specialized analog circuits typically used to implement plasticity, enabling efficient full-array STDP updates without extra memory or computational overhead.

V. CONCLUSION

In this paper, we propose a thermally interplayed TI-MTJ macro to emulate SNN functionalities. By leveraging intrinsic thermal dynamics, both biologically plausible LIF neurons and a simplified STDP learning rule are directly achieved within the device. In this way, the proposed macro enables efficient and biologically inspired SNN operations. The fabricated macro achieves high energy efficiency of 22.8TOPS/W, with each synaptic operation consuming only 6.07fJ per spike during inference and achieving 97.76% accuracy on the MNIST dataset. By embedding core SNN operations into the device, the architecture eliminates the need for separate memory arrays and dedicated computational modules, significantly improving energy efficiency and scalability compared to prior designs.

ACKNOWLEDGMENT

This work was supported by National Key Research and Development Program of China (No. 2024YFA1408804), Beijing Nova Program (20250484807), Research Funding of Hangzhou International Innovation Institute of Beihang University (2024KQ157), Zhejiang Provincial Key RD Program (2025C01071), and National Natural Science Foundation of China (62301019 and 62274008).

REFERENCES

- [1] W. Maass, "Networks of spiking neurons: the third generation of neural network models," *Neural networks*, vol. 10, no. 9, pp. 1659–1671, 1997.
- [2] G. Verma, A. Nisar, S. Dhull and B. K. Kaushik, "Neuromorphic accelerator for spiking neural network using SOT-MRAM crossbar array," *IEEE Transactions on Electron Devices*, vol. 70, no. 11, pp. 6012–6020, 2023.
- [3] J. H. Wijekoon and P. Dudek, "Compact silicon neuron circuit with spiking and bursting behaviour," *Neural Networks*, vol. 21, no. 2-3, pp. 524–534, 2008.
- [4] P. A. Merolla et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014.
- [5] J. Pei et al., "Towards artificial general intelligence with hybrid tianjic chip architecture," *Nature*, vol. 572, no. 7767, pp. 106–111, 2019.
- [6] G. Indiveri et al., "Neuromorphic silicon neuron circuits," *Frontiers in Neuroscience*, vol. 5, pp. 73, 2011.
- [7] T. Tuma, A. Pantazi, M. Le Gallo, A. Sebastian, and E. Eleftheriou, "Stochastic phase-change neurons," *Nature nanotechnology*, vol. 11, no. 8, pp. 693–699, 2016.
- [8] D. Fan, Y. Shim, A. Raghunathan, and K. Roy, "STT-SNN: A spin-transfer-torque based soft-limiting non-linear neuron for low-power Artificial Neural Networks," *IEEE Transactions on Nanotechnology*, vol. 14, no. 6, pp. 1013–1023, 2015.
- [9] X. Zhao et al., "Tailoring skyrmion motion dynamics via magnetoelectric coupling: Toward highly energy-efficient and reliable non-volatile memory applications," *Journal of Applied Physics*, vol. 132, no. 8, pp. 084902, 2022.
- [10] Y.-C. Chen, C.-C. Lin, and Y.-F. Chang, "Post-moore memory technology: Sneak path current (SPC) phenomena on RRAM crossbar array and solutions," *Micromachines*, vol. 12, no. 1, pp. 50, 2021.
- [11] F. Xiong, A. D. Liao, D. Estrada, and E. Pop, "Low-power switching of phase-change materials with carbon nanotube electrodes," *Science*, vol. 332, no. 6029, pp. 568–570, 2011.
- [12] S. R. Kulkarni and B. Rajendran, "Spiking neural networks for handwritten digit recognition—supervised learning and network optimization," *Neural Networks*, vol. 103, pp. 118–127, 2018.
- [13] X. Xu et al., "Full reliability characterization of three-terminal SOT-MTJ devices and corresponding arrays," in *IEEE International Reliability Physics Symposium (IRPS)*, pp. 1–6, 2023.
- [14] A. Sengupta, A. Banerjee, and K. Roy, "Hybrid spintronic-CMOS spiking neural network with on-chip learning: Devices, circuits, and systems," *Physical Review Applied*, vol. 6, no. 6, p. 064003, 2016.
- [15] M. G. Johnson and S. Chartier, "Spike neural models part II: Abstract neural models," *The Quantitative Methods for Psychology*, vol. 14, no. 1, pp. 1–16, 2018.
- [16] M. Gilson, T. Fukai, and A. N. Burkitt, "Spectral analysis of input spike trains by spike-timing-dependent plasticity," *PLoS Computational Biology*, vol. 8, no. 7, pp. e1002584, 2012.
- [17] G. Kim, K. Kim, S. Choi, H. J. Jang and S. -O. Jung, "Area- and energy-efficient STDP learning algorithm for spiking neural network SoC," *IEEE Access*, vol. 8, pp. 216922–216932, 2020.
- [18] P. U. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Frontiers in Computational Neuroscience*, vol. 9, pp. 99, 2015.
- [19] A. Jaiswal, S. Roy, G. Srinivasan, and K. Roy, "Proposal for a leaky-integrate-fire spiking neuron based on magnetoelectric switching of ferromagnets," *IEEE Transactions on Electron Devices*, vol. 64, no. 4, pp. 1818–1824, 2017.
- [20] A. Agrawal and K. Roy, "Mimicking leaky-integrate-fire spiking neuron using automation of domain walls for energy-efficient brain-inspired computing," *IEEE Transactions on Magnetics*, vol. 55, no. 1, pp. 1–7, 2019.
- [21] D. Wang et al., "Spintronic leaky-integrate-fire spiking neurons with self-reset and winner-takes-all for neuromorphic computing," *Nature Communications*, vol. 14, no. 1, p. 1068, 2023.
- [22] A. E. Arrassi, A. Gebregiorgis, A. E. Haddadi, and S. Hamdioui, "Energy-efficient SNN implementation using RRAM-based computation in-memory (CIM)," in *IFIP/IEEE 30th International Conference on Very Large Scale Integration (VLSI-SoC)*, pp. 1–6, 2022.
- [23] S. N. Pagliarini, S. Bhuin, M. M. Isgenc, A. K. Biswas, and L. Pileggi, "A probabilistic synapse with strained MTJs for spiking neural networks," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 4, pp. 1113–1123, 2020.
- [24] G. Srinivasan, A. Sengupta, and K. Roy, "Magnetic tunnel junction based long-term short-term stochastic synapse for a spiking neural network with on-chip STDP Learning," *Scientific Reports*, vol. 6, pp. 29545, 2016.
- [25] V. -T. Nguyen, Q. -K. Trinh, R. Zhang, and Y. Nakashima, "STT-BSNN: An in-memory deep binary spiking neural network based on STT-MRAM," *IEEE Access*, vol. 9, pp. 151373–151385, 2021.