Scalable and Energy-Efficient On-Device SNNs Enabled by Magnetic Tunnel Junctions

Sai Li[1], Ziyi Teng[1], Shuncheng Jia[2], Zhaohao Wang[1], Kaihua Cao[3], Hongchao Zhang[3], Hongxi Liu[3], Tielin Zhang[2],Weisheng Zhao[1] *Fellow, IEEE*

[1]National Key Laboratorycof Spintronics, School of Integrated Circuit Science and Engineering, Beihang University, Beijing, 100191, China
[2]Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China
[3]Truth Memory Corporation, Beijing 100088, China

To address the critical need for efficient, biologically-inspired neuromorphic chip exhibiting flexible learning and robustness against noise and catastrophic forgetting, We present a neuromorphic computing engine that leverages the intrinsic stochasticity and fast switching characteristics of spin-orbit torque magnetic tunnel junctions (SOT-MTJs). Its architecture innovatively integrates multi-port SOT switching to implement a 9-type spike-timing-dependent plasticity (STDP) rule, enhanced by meta-learning, while achieving excitation-inhibition balance of the spiking neural networks (SNNs) via sparsity-aware input encoding. This integrated approach effectively realizes key functions of flexible synaptic plasticity and selective neuronal activation. The results demonstrate over 96% classification accuracy on both MNIST and TIDigits datasets. Crucially, it showcases robust continuous learning capabilities by sequentially training all 10 MNIST classes in a single pass, thereby substantially reducing catastrophic forgetting with minimal computational overhead. This work demonstrates that MTJs can enable on-device SNNs, paving the way for the development of scalable and energy-efficient neuromorphic computing chips.

***Index Terms*—Spintronic Devices, Neuromorphic Computing, Synaptic Plasticity, Continuous Learning**

I. Introduction

The remarkable capability of the human brain is to continuously learn from dynamic environments, where different regions such as the occipital and temporal lobes govern distinct cognitive modalities and are modulated by various neurotransmitters (Figure 1a). Excitatory and inhibitory neurons coexist within each region, with inhibitory neurons blocking neurotransmission to help maintain network stability. Higher-order learning is further regulated by metaplasticity, a neuromodulation-driven mechanism that dynamically tunes the threshold and polarity of synaptic plasticity in response to environmental context. Key among these are two mechanisms: neuronal selection, which allows the brain to focus on relevant information, and synaptic modulation, which underpins learning and memory (Figure 1b). Specifically, neuronal selection are achieved through a delicate Excitation-Inhibition (E-I) balance, which dictates whether neurons fire intensively, sparsely, or remain silent in response to inputs (Figure 1c) [1]. On the other hand, neuromodulatory mechanisms—particularly those involving dopamine—adjust synaptic efficacy beyond classical Hebbian rules [2]. These effects, known as metaplasticity, regulate the amplitude, polarity, and temporal dynamics of long-term potentiation (LTP) and depression (LTD), which can manifest as 9 distinct forms of STDP (Figure 1d) [3].

However, existing neuromorphic hardware often supports only simplified functionalities, hindering the realization of complex learning mechanisms on energy-efficient chips [4]. To overcome this limitation, we introduce a novel neuromorphic engine (Figure 1(e)) leveraging the intrinsic stochasticity and rapid switching characteristics of SOT-MTJs [5]. This engine employs multi-port SOT switching to directly implement both neuronal selection and synaptic modulation within on-device SNNs for learning via weight updates and controlling neuron spiking rates.

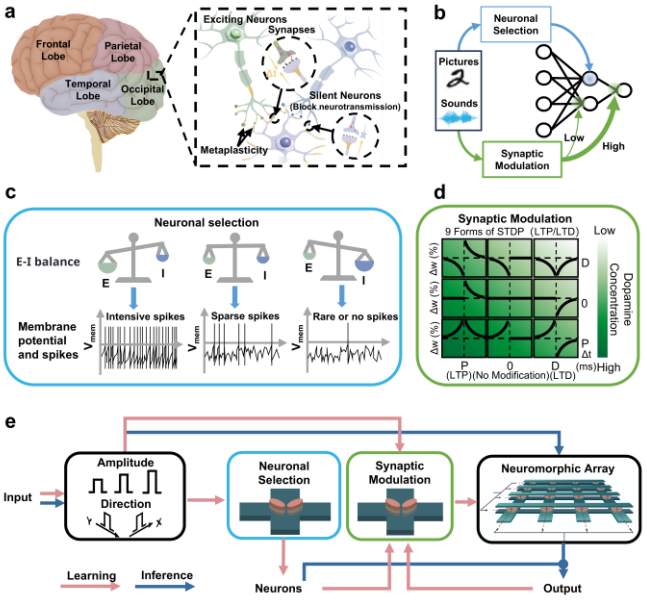

Fig. 1. Biologically inspired architecture and its SOT-MTJ implementation. (a) Brain regions and fundamental neural components (excitatory/silent neurons, synapses, metaplasticity). (b) Schematic of selective information processing in a SNN, where input stimuli (e.g., images, audio) are handled by neuron populations based on neuronal selection and synaptic modulation. (c) Neuronal selection via E-I balance, controlling spike rates. (d) Synaptic modulation is governed by dopamine-dependent metaplasticity, enabling dynamic combinations of LTP and LTD and realizing nine distinct STDP learning rules. (e) Architecture of the proposed neuromorphic engine using SOT-MTJs, integrating neuronal selection and synaptic modulation blocks within a neuromorphic array for input processing, learning, and inference.

II. Methods

We fabricated an in-plane type-Y SOT-MTJ pair, whose multilayer stack and device layout are illustrated in Fig. 2a. The four-terminal structure features a heavy-metal cross acting as the SOT current channels, with two MTJs patterned on its orthogonal arms, yielding a 90°tilt between their easy axes. Current-induced switching was examined with the circuit shown in Fig. 2b. Applying current along the X-direction induces synchronous switching of the two MTJs, whereas Y-direction excitation causes them to switch anti-synchronously. This inherent capability to toggle between correlated and anti-correlated states provides the hardware basis for the synaptic functions discussed subsequently.

Given that SOT devices offer highly reproducible, pulse-controlled switching, we focused on their single-pulse response rather than pursuing conventional multi-level conductances. Using the previously determined threshold voltages as a guide, rectangular voltage pulses of varying amplitude were applied to the SOT-MTJs at room temperature without external magnetic field. For the measurements shown in Fig. 2c, the pulse amplitude was swept from –3.94 V to –3.66 V in 0.02 V steps. The switching probability was calculated as $P_{SW} = N_{flip}/N_{tot}$ and plotted against pulse amplitude, where each probability was measured 50 times. Therefore, we could obtain sigmoid fits of the switching probability curve under different SOT currents. Repeating the protocol for the opposite polarity and for the left/right MTJ produces four characteristic curves (AP to P and P to AP for both junctions), which form the quantitative basis for the pulse-based synaptic modes.

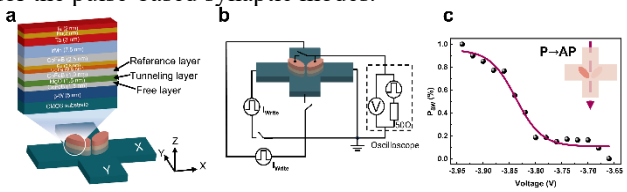

Fig. 2. Structure, measurement set-up, and single-pulse switching statistics of the SOT-MTJ. (a) Multilayer stack and geometry of the four-terminal device. (b) Electrical measurement setup used for SOT switching. (c) Probability switching characterization for the left MTJ under –Y excitation.

III. Result

Exploiting the direction-dependent switching of the two MTJs—synchronous under ±Y excitation and anti-synchronous under ±X excitation—the SOT-MTJ pair implements 9-type STDP rules. A local selector chooses which junction is monitored, while the sign of its conductance change, ΔG, defines long-term potentiation (LTP, $\Delta G > 0$, $\Delta W = +1$) or depression (LTD, $\Delta G < 0$, $\Delta W = -1$); the weight remains unchanged when no switch occurs.

For instance, when a pre-synaptic pulse($-Y$) followed by a post-synaptic pulse($+Y$) the left MTJ switches from AP to P, generating a LTP mode ($\Delta t > 0$); reversing the Y direction order, it produces a LTD mode ($\Delta t < 0$). Selecting the right MTJ could invert these updates because its SOT Y-polarity is opposite. Currents along $\pm X$ drive the two MTJs synchronously, giving pure LTD (DD), no change (00) or pure LTP (PP), irrespective of Δt. Formally, the update obeys $\Delta W = f_{local} \times LTP/LTD$.

Neuronal selection is implemented by the same stochastic response. A sigmoid switching curve $G(\cdot)$ of the SOT-MTJ pair is used to generate a Mask. When the device is in the AP state, Mask ≠ 0 and the leaky-integrate-and-fire (LIF) neuron behaves normally (“Fire”); when it is in the P state, Mask = 0 and the neuron is silenced even if its membrane potential reaches threshold (“don’t Fire”).

We simulated these synaptic-modulation and neuron-selection schemes in a pulse-based spiking neural network. The resulting accuracy reached 96.51 % on MNIST and 96.54 % on TIDigits. Furthermore, the architecture demonstrates effective continuous learning capability by sequentially training on all 10 MNIST classes without iteration, significantly reducing catastrophic forgetting while maintaining low computational overhead.

References

[1] J. Liang, Z. Yang, and C. Zhou, “Excitation–Inhibition Balance, Neural Criticality, and Activities in Neuronal Circuits,” *The Neuroscientist*, vol. 31, no. 1, pp. 31–46, Feb. 2025.

[2] G. Bi and M. Poo, “Synaptic modification by correlated activity: Hebb’s postulate revisited,” *Annu. Rev. Neurosci.*, vol. 24, pp. 139–166, Mar. 2001.

[3] S. B. Flagel, J. J. Clark, T. E. Robinson, L. Mayo, A. Czuj, I. Willuhn, C. A. Akers, S. M. Clinton, PE. M. Phillips, and H. Akil, “A selective role for dopamine in stimulus-reward learning,” *Nature*, vol. 469, no. 7328, pp. 53–57, Jan. 2011.

[4] D. Kudithipudi, C. Schuman, and S. Furber, “Neuromorphic computing at scale,” *Nature*, vol. 637, no. 8047, pp. 801–812, Jan. 2025.

[5] Z. Guo, J. Yin, Y. Bai, D. Zhu, K. Shi, G. Wang, K. Cao, and W. Zhao, “Spintronics for energy-efficient computing: An overview and outlook,” *Proc. IEEE*, vol. 109, no. 8, pp. 1398–1416, Aug. 2021.