SCIENCE CHINA
Information Sciences

• LETTER •

July 2026, Vol. 69, Iss. 7, 179402:1–179402:2
https://doi.org/10.1007/s11432-025-4707-8

VGSOT-BNN: an energy-efficient VGSOT-MRAM-based in-memory computing macro with offset-tolerant sensing for binary neural networks

Chao WANG[1], Qihang GAO[1,2,3,4], Zhongzhen TONG[1,2,3,4], Xianzeng GUO[1,2,3,4], Yulong QIU[1], Zhaohao WANG[1,2,3,4*] & Weisheng ZHAO[1,2,3,4]

[1]*State Key Laboratory of Spintronics, Hangzhou International Innovation Institute, Beihang University, Hangzhou 311115, China*
[2]*Fert Beijing Research Institute, Beihang University, Beijing 100191, China*
[3]*MIIT Key Laboratory of Spintronics, Beihang University, Beijing 100191, China*
[4]*School of Integrated Circuit Science and Engineering, Beihang University, Beijing 100191, China*

Received 29 April 2025/Revised 10 September 2025/Accepted 3 December 2025/Published online 22 June 2026

Citation Wang C, Gao Q H, Tong Z Z, et al. VGSOT-BNN: an energy-efficient VGSOT-MRAM-based in-memory computing macro with offset-tolerant sensing for binary neural networks. Sci China Inf Sci, 2026, 69(7): 179402, https://doi.org/10.1007/s11432-025-4707-8

The growing demand for artificial intelligence (AI) applications on edge devices has driven significant research into energy-efficient neural network architectures, drawing considerable attention to in-memory computing (IMC) for its potential to drastically reduce energy consumption [1]. Spin-transfer torque magnetic random-access memory (STT-MRAM) is a promising candidate for IMC owing to its advantages in non-volatility, speed, density, endurance, and process variation, particularly well-suited for high-efficiency binary neural networks (BNNs) [2–4]. However, STT-MRAM-based IMC for BNN (STT-BNN) faces significant challenges in achieving high computing performance due to the low resistance and limited on/off ratio of magnetic tunnel junction (MTJ).

To overcome these bottlenecks, voltage-gated spin-orbit torque MRAM (VGSOT-MRAM), which facilitates ultra-high device resistance to enhance the energy efficiency, has gained notable attention [5]. Exploiting the structural and performance advantages of this device, we propose an energy-efficient VGSOT-MRAM-based IMC for BNN (VGSOT-BNN), validated via extensive simulations in a 28 nm process. A 3-transistor-2-MTJ (3T2J) bit-cell design integrating memory and computing modes is first introduced, achieving 5.72× higher energy efficiency than conventional STT-BNN. To address the low on/off ratio issue, a 1-transistor-4-MTJ (1T4J) bit-cell supporting complementary sensing is developed, enhancing computing accuracy, reducing bit-cell area and improving speed. Furthermore, an in-array offset trim scheme is proposed to mitigate computing errors without modifying peripheral circuits. With this scheme, the 3T2J and 1T4J designs attain the energy efficiencies of 12472 and 24666 1-bit TOPS/W at 592 and 775 MHz frequencies, respectively, maintaining sufficient computing accuracy.

Conventional STT-BNN architecture. Figure 1(a) illustrates a typical STT-BNN architecture [4]. The bit-cell performs XNOR operations between the input IN and weight W, while the accumulated voltage on the source line (SL) maps the dot product of the N-bit IN and W, which is then quantized to either $+1$ or -1 by a sense amplifier (SA). The resistance of STT-MTJ is typically on the order of 10 kΩ, leading to an excessively low computing-path resistance and consequently high current and energy consumption. Under high computing parallelism, driver circuits incur substantial area overhead, while computing accuracy becomes highly sensitive to the resistance of the wire and transistor.

Proposed VGSOT-BNN architecture. The proposed 3T2J VGSOT-BNN bit-cell design is illustrated in Figure 1(b). During the write operation, the read bit lines RBL and RBLB are biased with V_G to reduce the critical switching current of MTJ1 and MTJ2 through the voltage-controlled magnetic anisotropy (VCMA) effect. The complementary write paths implemented by the layout ensure that MTJ1 and MTJ2 are written with complementary W and WB. For computing operation, RBL and RBLB receive input signals, and the XNOR results from all bit-cells in the same column are accumulated onto V_{SUM} through the SL. Beyond the energy efficiency and accuracy improvements afforded by the high resistance of VGSOT-MRAM, the proposed 3T2J design also optimizes the bit-cell area by sharing a write transistor, reduces SL load capacitance to enhance computing speed, and shortens the RBL length to lower the energy consumption. As a result, it achieves a 5.72× reduction in computational energy compared to STT-BNN.

To further enhance the sensing margin and improve the accuracy, the 1T4J bit-cell design is proposed as illustrated in Figure 1(c). The write operation is performed in two steps: first, BL is set to V_G while BLB is set to GND, significantly reducing the critical switching current for MTJ1 and MTJ2 and enabling their selective writing; second, the bias configurations of BL and BLB are swapped to write complementary data into MTJ3 and

* Corresponding author (email: zhaohao.wang@buaa.edu.cn)

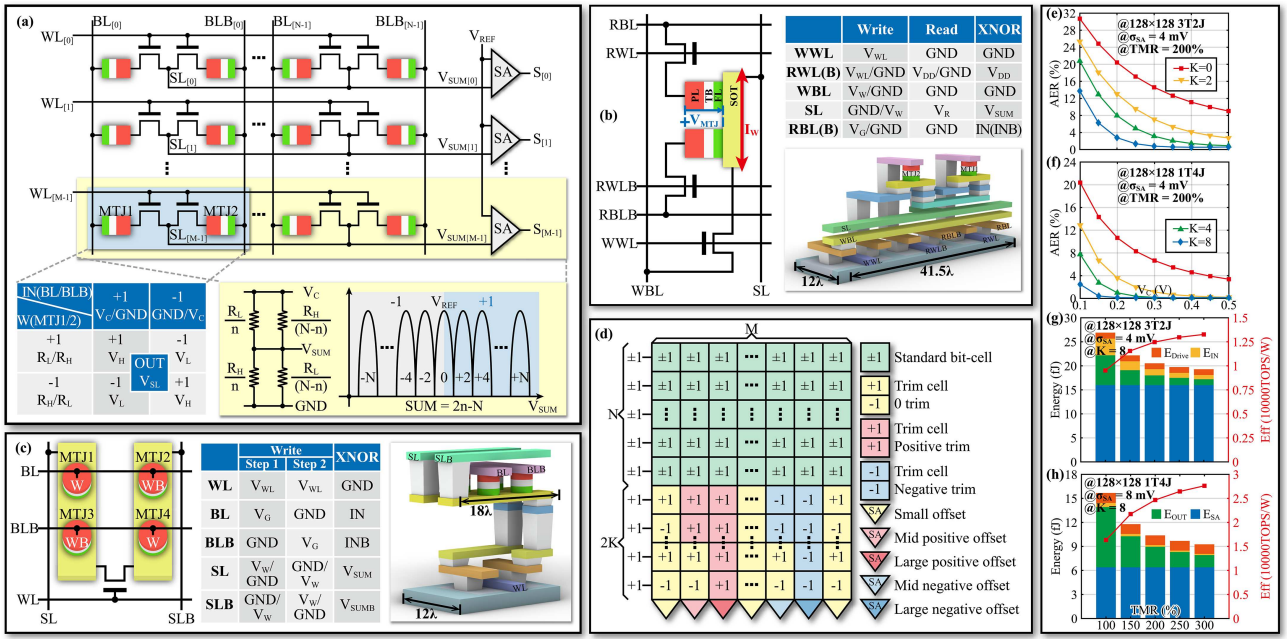

Figure 1 (Color online) (a) Schematics of conventional STT-BNN architecture. Schematics of (b) 3T2J and (c) 1T4J designs: bit-cell, operation control and 3D layout ($\lambda = 14$ nm). (d) Schematics of the proposed in-array offset trim scheme. AER for (e) 3T2J and (f) 1T4J designs as functions of V_C with various K. Computing energy and efficiency for (g) 3T2J and (h) 1T4J designs with varying tunnel magnetoresistance ratio (TMR).

MTJ4, respectively. During the computing operation, complementary voltages V_{SUM} and V_{SUMB} are generated on SL and SLB, respectively, and fed into the SA for comparison. This complementary sensing scheme doubles the sensing margin, thereby substantially improving the accuracy. Moreover, the 1T4J design significantly reduces the bit-cell area and computing latency, with only a minor energy penalty. Nevertheless, the 1T4J design does not support conventional reading and suffers from half-select leakage during writing. Compared to the 3T2J design, which enables standard-memory/compute dual-mode functionality, each design offers distinct advantages and is better suited to different application scenarios.

In-array offset trim scheme. For VGSOT-BNN, the sensing margin can be as low as a few mV, making the SA offset the dominant source of computing error. To address this issue, an in-array offset trim scheme is proposed as shown in Figure 1(d), introducing a small number of trim rows composed of standard bit-cells programmable to "+1" or "+1" to precisely adjust V_{SUM} and thereby trim the SA offset. The proposed scheme incurs only minimal hardware overhead and features non-volatility, making it a highly practical and efficient solution for ensuring robust computing.

Evaluation and analysis. The proposed designs are validated through extensive simulations using a 28 nm process technology for a 128×128 array. As shown in Figures 1(e) and (f), the proposed in-array offset trim scheme significantly reduces the effective average error rate (AER) and substantially reduces the required computing voltage (V_C) to meet the target AER. Both the 3T2J and 1T4J designs enable reliable computation at $V_C < 0.2$ V. Given that energy consumption scales approximately quadratically with V_C, the scheme leads to exceptional energy efficiency. As shown in Figures 1(g) and (h), the 3T2J design achieves an energy efficiency of 12472 TOPS/W at 592 MHz, while the 1T4J design attains an efficiency of 24666 TOPS/W at 775 MHz, both evaluated at TMR of 200%.

Conclusion. We propose a novel VGSOT-BNN architecture that effectively addresses the energy and accuracy challenges of STT-BNN. By leveraging innovative 3T2J and 1T4J bit-cell designs alongside an efficient in-array offset trim scheme, the proposed architecture achieves state-of-the-art performance in both energy efficiency and speed. This work presents a compelling pathway toward ultra-low-power hardware accelerators for BNNs on edge devices.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62401031, 62171013), Postdoctoral Research Funding of Hangzhou International Innovation Institute of Beihang University (Grant No. 2024BKZ004), and National Key Research and Development Program of China (Grant Nos. 2021YFB3601303, 2021YFB3601304, 2021YFB3601300).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Sun Z, Kvatinsky S, Si X, et al. A full spectrum of computing-in-memory technologies. Nat Electron, 2023, 6: 823–835
2 Cai H, Liu B, Chen J T, et al. A survey of in-spin transfer torque MRAM computing. Sci China Inf Sci, 2021, 64: 160402
3 Jung S, Lee H, Myung S, et al. A crossbar array of magnetoresistive memory devices for in-memory computing. Nature, 2022, 601: 211–216
4 Pham T N, Trinh Q K, Chang I J, et al. STT-BNN: a novel STT-MRAM in-memory computing macro for binary neural networks. IEEE J Emerg Sel Top Circuits Syst, 2022, 12: 569–579
5 Doevenspeck J, Garello K, Rao S, et al. Multi-pillar SOT-MRAM for accurate analog in-memory DNN inference. In: Proceedings of IEEE Symp. VLSI Technol. (VLSI), Kyoto, 2021. 1–2