# Cryogenic In-Memory Computing Circuits with Giant Anomalous Hall Current in Magnetic Topological Insulators for Quantum Control

K. Qian[1*], A. Lee[2*#], Z. Xiao[1,3*], H. He[4], S. Cheung[1], Y. Liu[5], F. P. Nugraha[1] and Q. Shao[1,3#]
[1]HKUST, Hong Kong, China; [2]InstonTech, Suzhou, China; [3]AI Chip Center for Emerging Smart Systems, Hong Kong; [4]UCLA, CA, USA; [5]Harbin Institute of Technology, Shenzhen, China *Equal contribution; [#]Email: eeqshao@ust.hk

***Abstract*—**Cryogenic in-memory computing, operating at temperatures below 4.2 K, offers high performance and energy efficiency in computation-intensive environments, especially for quantum controls. Spintronic memristors based on anomalous Hall effect (AHE) can be utilized with advantages of multi-bit bipolar weights and nearly unlimited endurance. However, there exists two major challenges in realizing an AHE-based neural network (NN): first, anomalous Hall resistances $R_H$ typically range in the order of a few ohms in conventional ferromagnets; second, signal-to-noise ratio of summation and read disturbance can lead to errors in NN operation. In this work, we demonstrate large $R_H$ of 12 kΩ at 2 K using a magnetic topological insulator (MTI) utilizing the quantum AHE, and for the first time, propose and experimentally verify the multiply-and-accumulate operation of a transverse-read Hall-current based neural network (Hall NN). Simulation of the proposed MTI Hall NN achieves matching accuracy to a full-precision network in the noisy qubit state preparation task. Compared with the MRAM NN, the MTI NN features a 90% lower write energy and 10 times higher TOPS/W showing the promise of MTI Hall NNs.

## I. Introduction

The advent of AI has brought forth new opportunities for hardware accelerators built upon memristor crossbars which promise efficient and high-speed matrix multiplications [1-3]. In the quantum computing system, deep reinforcement learning (RL) has been applied for tasks such as quantum error correction [4], real-time quantum feedback [5] and state preparation [6]. RL can outperform commonly used conventional algorithms in quantum control tasks [7]. Until now, quantum systems are cooled to cryogenic temperatures, while RL agents are controlled by classical hosts such as GPUs and TPUs at room temperature (Fig. 1). The development of cryogenic in-memory computing (IMC) is a solution for quantum control with improved computing speed and energy efficiency [8]. Among memristor candidates, magnetic devices have advantages of high speed, low power, and practically unlimited endurance. The majority of magnetic neural network (NN) designs have focused on magnetoresistive random-access memories (MRAMs), and much less around Hall-based devices [9-10]. The Hall resistance $R_H$ has the unique ability to implement positive, negative, and zero weights which would otherwise require differential devices [11]. While traditional ferromagnets (FMs) usually have $R_H$ in the range of several ohms, magnetic topological insulators (MTIs) can exhibit increased $R_H$ with decreasing temperature and be quantized at 25.6 kΩ (named quantum anomalous Hall effect) in ultralow temperatures, making them ideal for cryogenic IMC. However, due to the four-way conducting nature of Hall devices, there has yet to be experimental demonstration of a functioning Hall NN. Prior designs attempting to form multiply-and-accumulate (MAC) functions suffer from signal shorting when attempting to sum Hall voltages $V_H$ and resulting in low SNR [12]. In this work, we propose and for the first time experimentally demonstrate the basic operations of an energy-efficient MTI Hall NN architecture with low read disturbance using a transverse-read Hall-current-based MAC. Our algorithm and circuit-level simulations of an MTI Hall NN array inference show accuracy matching that of full-precision networks for qubit state control and lower energy consumption compared with existing MRAM and CMOS technologies.

## II. Device Characterization

MTI and FM (reference) Hall devices were fabricated for the Hall NN design verification. The FM Hall device is composed of Ta(3)/$Co_{20}Fe_{60}B_{20}$(1)/MgO(2)/$TaO_x$(3) on a Si/$SiO_2$ wafer deposited via magnetron sputtering. The MTI is made of molecular beam epitaxy-grown single-crystalline $(Cr_{0.15}Bi_{0.26}Sb_{0.59})_2Te_3$ thin films with $AlO_x$ capping on GaAs substrates. The optical image and fabrication procedure of the MTI Hall device are shown in Fig. 2. As shown in Figs. 4a-b, $R_H$ of FM and MTI Hall devices are 4 Ω and 12 kΩ, respectively. When a longitudinal current $I_x$ flows through the device, $V_H$ proportional to the device magnetization $m_z$ and $I_x$ can be observed on the transverse channels (Figs. 4c-d). The ratio between $V_H$ and $I_x$ is characterized as $R_H = R_{yx} = V_H/I_x$. If the transverse Y channel is clamped at an additional voltage $V_{clamp}$ as shown in Fig. 3b, the current in the Y direction can be attributed to both AHE and $V_{clamp}$, which is expressed as $I_y = (V_{clamp} - I_x R_H)/R_{YY}$, where $R_{YY}$ is the transverse-channel resistance. The measured $I_y$ with $V_{clamp}$ are shown in Figs. 4e-f. $V_{clamp}$ on the device is crucial to ensure each cell operates independently regardless of states of other devices during MAC operations. $R_H$ can be electrically programmed via current-induced spin-orbit torque under an assisting magnetic field along the current direction. As shown in Fig. 5b, our critical switching current density $J_C$ is 4.0 x $10^5$ A/$cm^2$, an order of magnitude lower than that of FM [11]. The switching behavior is modeled as shown in Fig.3a.

## III. Hall Neural Network

An NN accelerator requires the efficient implementation of the MAC function. Previous works have proposed the accumulation/summation of $V_H$ through connecting the channels for $V_H$ in series, as shown in Fig. 6a [12]. However, the resulting total voltage $V_{total}$ of such a design is not always equal to the sum ($V_{sum}$) of the individual $V_H$, because the four-way conductive Hall-cross ports short the $V_H$ generated by neighboring cells. A physical simulation of a series $V_H$ summation shown in Fig. 6b and e, clearly shows a difference

between $V_{\text{total}}$ and $V_{\text{sum}}$. Another challenge in previous Hall NNs is the read disturbance, as both write and read driving currents share the same paths.

To overcome these challenges, we propose a Hall NN based upon two design principles. Firstly, MAC is conducted in current mode with clamped channels as shown in Fig. 6c. A physical simulation of electric potential distribution in Fig. 6d and f shows successful MAC operations where $I_{\text{total}}$ is equal to the sum ($I_{\text{sum}}$) of the individual $I_{\text{H}}$. Secondly, we separate write and read driving current paths. Identical current readouts can be generated in both the transverse and lateral channels as shown in Fig. 5a, yet disturbance is significantly suppressed as $I_{\text{y}}$ cannot induce magnetization switching with $B_{\text{x}}$ (Fig. 5b).

The schematic of our Hall NN is presented in Fig. 7a. Three transistors are introduced to each Hall device to form a memory cell. Transistor $T_{\text{T}}$, controlled by $WWL$, connects longitudinal node $X+$ of the device to bus $SL$. Transistors $T_L$ and $T_R$, both controlled by the $RWL$, connect the transverse nodes $Y+$ and $Y-$ to bus $BL$ and $BLB$, respectively. The longitudinal node $X-$ of the device is connected to $SLB$ directly. Cells in the same row share $WWL$, $RWL$, and $SLB$ buses, while cells of the same column share the same $SL$, $BLB$, and $BL$. The perpendicular $SL$ and $SLB$ orientation is necessary to allow both read and NN operations in the same array. The operation waveform and the conditions of each device during operations are displayed in Figs. 7b-c. During a memory write, the $WWL$ of the selected row is activated, while the other $WWLs$ and all $RWLs$ are grounded. $SL$ and $SLB$ are biased to the write conditions to switch the device to the desired state, e.g., $V_{\text{SL}} = V_{\text{W}}$; $V_{\text{SLB}} = V_{\text{VGND}}$ for each row. During a memory read, both the $WWL$ and the $RWL$ of the selected row are activated, all $BLs$ and $BLBs$ are biased to the read driving voltage $V_{\text{R}}$, and all $SLs$ and $SLBs$ are clamped to the read clamp voltage $V_{\text{clamp}}$ with the center node of the Hall device virtually grounded, e.g., $V_{\text{BL,BLB}} = V_{\text{VGND}} \pm \frac{V_{\text{R}}}{2}$; $V_{\text{SL,SLB}} = V_{\text{VGND}} \pm \frac{V_{\text{CLAMP}}}{2}$. The state of the selected cell is confirmed by the Hall current on each $SL$. The virtual ground $V_{\text{VGND}}$ design avoids sneaky paths in the Hall NN. During an NN operation, all $WWLs$ and $RWLs$ are turned on. The NN inputs $V_{\text{in}}$ are applied to the $BLs$ and $BLBs$, and $V_{\text{clamp}}$ in reverse are applied to the $SLs$ and $SLBs$, e.g., $V_{\text{BL,BLB}}[\text{i}] = V_{\text{VGND}} \pm \frac{V_{\text{in}}[\text{i}]}{2}$; $V_{\text{SL,SLB}} = V_{\text{VGND}} \mp \frac{V_{\text{CLAMP}}}{2}$. The NN operation results are determined by the currents on each $SLB$.

## IV. Results

The setup and experimental results for the Hall voltage-based and current-based MAC operation are shown in Figs. 8a-d, which are consistent with simulations in Fig. 6. We then consider a 3-input NN operation where the inputs and weights are binarized, e.g., $V_{\text{VGND}} \pm V_{\text{in}}$ as inputs and $m_z = \pm 1$ as weights. The total current $I_{\text{total}}$ as a function of the MAC result across different combinations of $I_{\text{in}}$ and weights is displayed in Fig. 8e. The linear relation between $I_{\text{total}}$ and the MAC result confirms the correct operation of the current-based Hall NN.

The Hall NN is benchmarked using a 22 nm process design kit. The device is modelled in Verilog-A using a distributed conductance-transconductance model (Fig. 3) with parameters extracted from the experiments. Simulation waveforms of a 3 × 3 Hall NN is presented in Fig. 9a. In cycles 1 to 3, the array is programmed to the pattern $[[-1,+1,-1],[+1,-1,-1],[+1,+1,+1]]$. Read verification is carried out in cycles 4 to 6, where there are distinct currents for the two states (1.33 uA and 2.26 uA). In cycle 7, the NN operation of the stored weights with the input pattern of $[+1,-1,+1]$ is computed. The currents of 4.33 uA, 6.27 uA, and 5.32 uA on each $SLB$ correspond to the MAC results of $-1, 3, 1$, respectively, confirming the NN function. Note that while any value of $V_{\text{clamp}}$ and $V_{\text{R}}$ can be used, we set both to 10 mV to ensure that the readout current always flows in the same direction to allow simple current-mirror readout.

To show the relevance of cryogenic IMC for quantum computing, the performance of the MTI NN is then analyzed in circuit-level and RL training for qubit preparation. From layout and parasitic extraction for a 512 × 512 MTI NN with a 50 × 50 nm device size, the network achieves a 90% lower write energy and 10 times higher TOPS/W compared with MRAM (Table 1), where an MTI device with multi-states is considered as 4 bits [13]. For qubit preparation in Figs. 9b-d, the task aims to control the state of k=8 serially coupled spins via magnetic flux pulses and drive them from an initial state to a target state with a measured 2% write and read noise [7]. We compared the RL methods on software and MTI with the conventional Krotov method. The result shows that the MTI NN and full-precision network using RL are more robust to the magnetic flux noise. The bipolar MTI NN shows software-level accuracy, while the fidelity of unipolar NN is low.

## V. Conclusion

This paper for the first time experimentally demonstrates the basic operations and proposes a scalable circuit architecture of an MTI Hall-based NN accelerator utilizing Hall current summation and transverse read scheme. Giant $\pm R_{\text{H}}$ and low switching current promise efficient NN operations at 2 K. The MTI NN features a 90% lower write energy and 10 times higher TOPS/W than MRAM NNs in the circuit level simulation. The RL agent model using MTI NN for qubit preparation achieved comparable performance with full-precision software. Overall, this work provides an efficient and accurate Hall NN solution for cryogenic IMC.


## Acknowledgements

The authors thank Dr. Kang Wang and Dr. Peng Zhang of UCLA for providing the MTI thin films. This research was supported by National Key R&D Program of China (Grants No.2021YFA1401500), NSFC youth program (Grant No. 12304137), AI Chip Center for Emerging Smart Systems, and the State Key Laboratory of Advanced Displays and Optoelectronics Technologies.


## References


[1] C. Xue et. al., JSSC, 2019; [2] W. Khwa et. al., ISSCC, 2022; [3] Q. Dong et. al., ISSCC, 2020; [4] V.V. Sivak et al., Nature 616, 50–55, 2023; [5] K. Reuer et al., Nat Commun 14, 7138, 2023; [6] V. V. Sivaket. et. al., Phys. Rev. X 12, 011059, 2022; [7] X.M. Zhang et.al., npj Quantum Information, 5:85, 2019; [8] S. Alam et.al., Nat Electron, 6, 185–198, 2023; [9] J. Doevenspeck et al., Symp. VLSI, 2021; [10] N. Xu et. al., IEDM, 2018; [11] Q. Shao et al., IEDM, 2019; [12] X. Lan et al., Advanced Intelligent Systems, 2021. [13] Y. Liu, et al., arXiv preprint arXiv:2209.09443, 2022.

## Cryogenic In-Memory Computing with Magnetic Topological Insulators

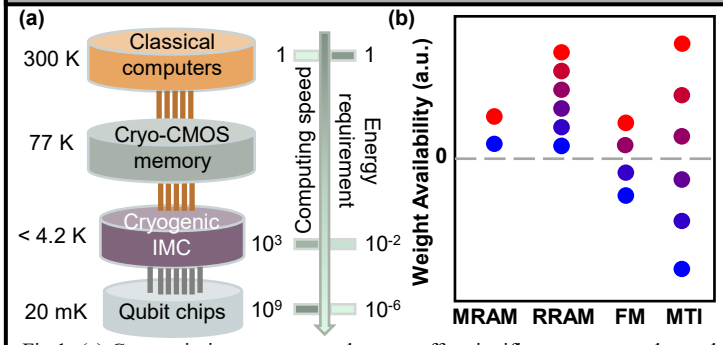


Fig.1. (a) Cryogenic in-memory accelerators offer significant energy and speed benefits in computation-intensive environments, especially for quantum computers [8]. (b) Achievable weight ranges of different memristor technologies. Hall devices of the ferromagnet (FM) and magnetic topological insulator (MTI) provide large and unique bipolar weight range.

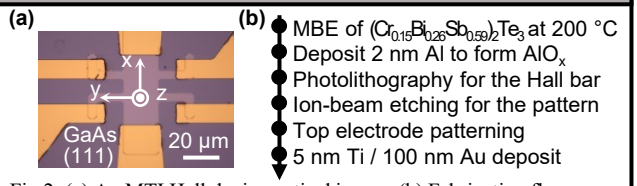


Fig.2. (a) An MTI Hall device optical image. (b) Fabrication flow.

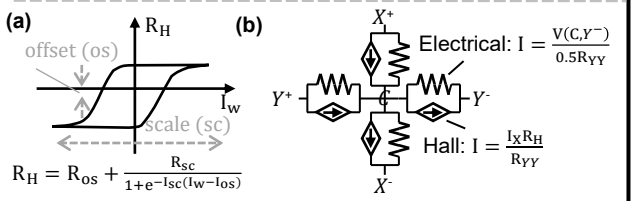


Fig.3. Models for (a) the current-induced switching loop, and (b) Anomalous Hall resistance with voltage $V_C$ on the center node.

## Experimental Characteristics of Hall Devices

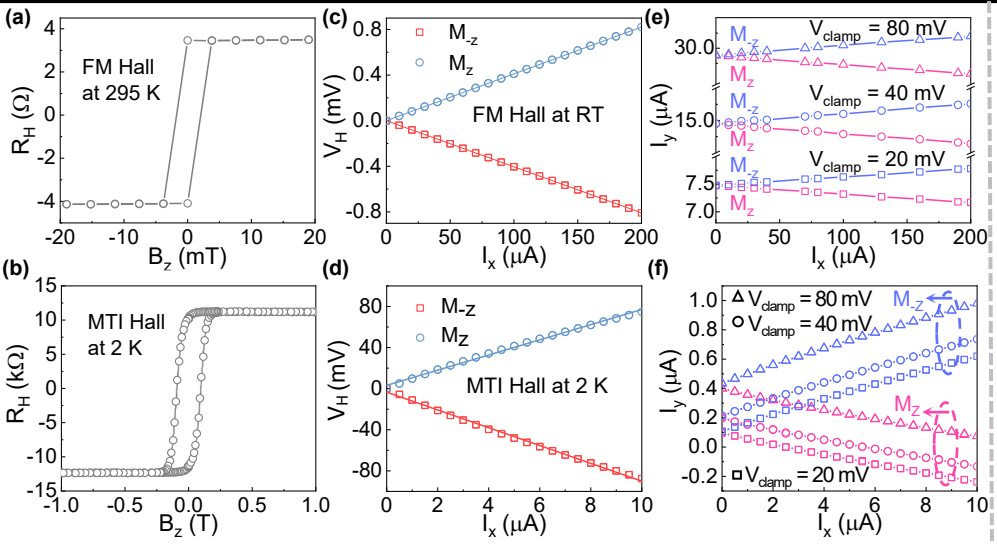


Fig.4. Measured Anomalous Hall resistance $R_H$ with out-of-plane magnetic field $B_z$ of (a) the FM Hall device (Ta/CoFeB/MgO/$TaO_x$) and (b) the MTI Hall device. Measured Hall voltage vs input current of (c) the FM Hall device and (d) the MTI Hall device. Measured Hall current with different clamp voltages $V_{clamp}$ of (e) the FM Hall device and (f) the MTI Hall device. While the FM device achieves an Anomalous Hall resistance $R_H$ of ~ 4 Ω, the MTI device achieves a giant $R_H$ of ~12 kΩ because of near the quantum Anomalous Hall regime at 2 K.

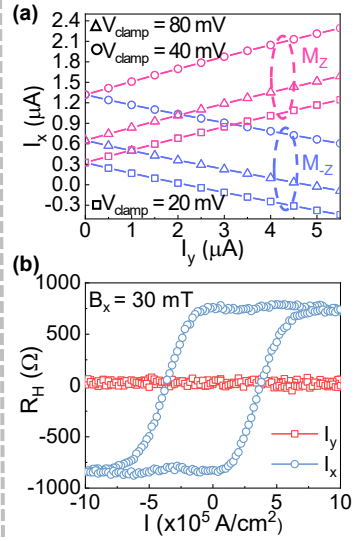


Fig.5. Vertical read to reduce read disturbance by applying input $I_y$ instead of $I_x$ in the MTI. (a) The readout Hall current and (b) current-induced switching with $B_x$.

## Prior Voltage-Based MAC vs. Proposed Current-based MAC

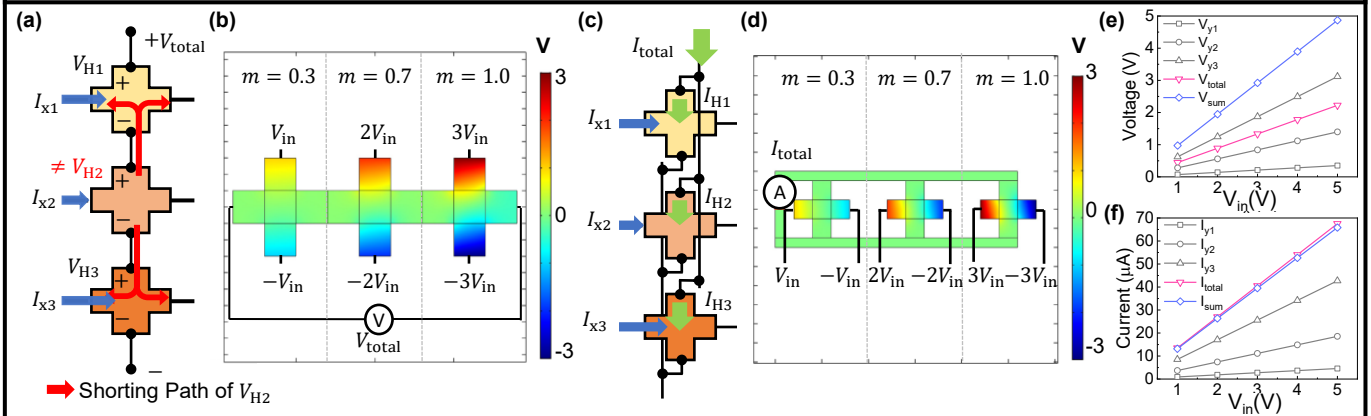


Fig.6. (a) Schematic of voltage-based MAC, where the $V_H$ of individual cells are connected in series (red: shorting paths). (b) Physical simulation of voltage distribution along each cell; (c) Schematic of current-based MAC; where the Hall current $I_H$ of individual cells are connected in parallel and (d) simulated voltage distribution; (e) Simulated $V_{total}$ as a function of $V_{in}$, where there exists a clear difference between the sum of the $V_H$ and $V_{total}$, mainly due to the shorted center cells not contributing to $V_{total}$. (f) Simulated $I_{total}$ as a function of $V_{in}$, which shows similar results between sum of $I_H$ and $I_{total}$.

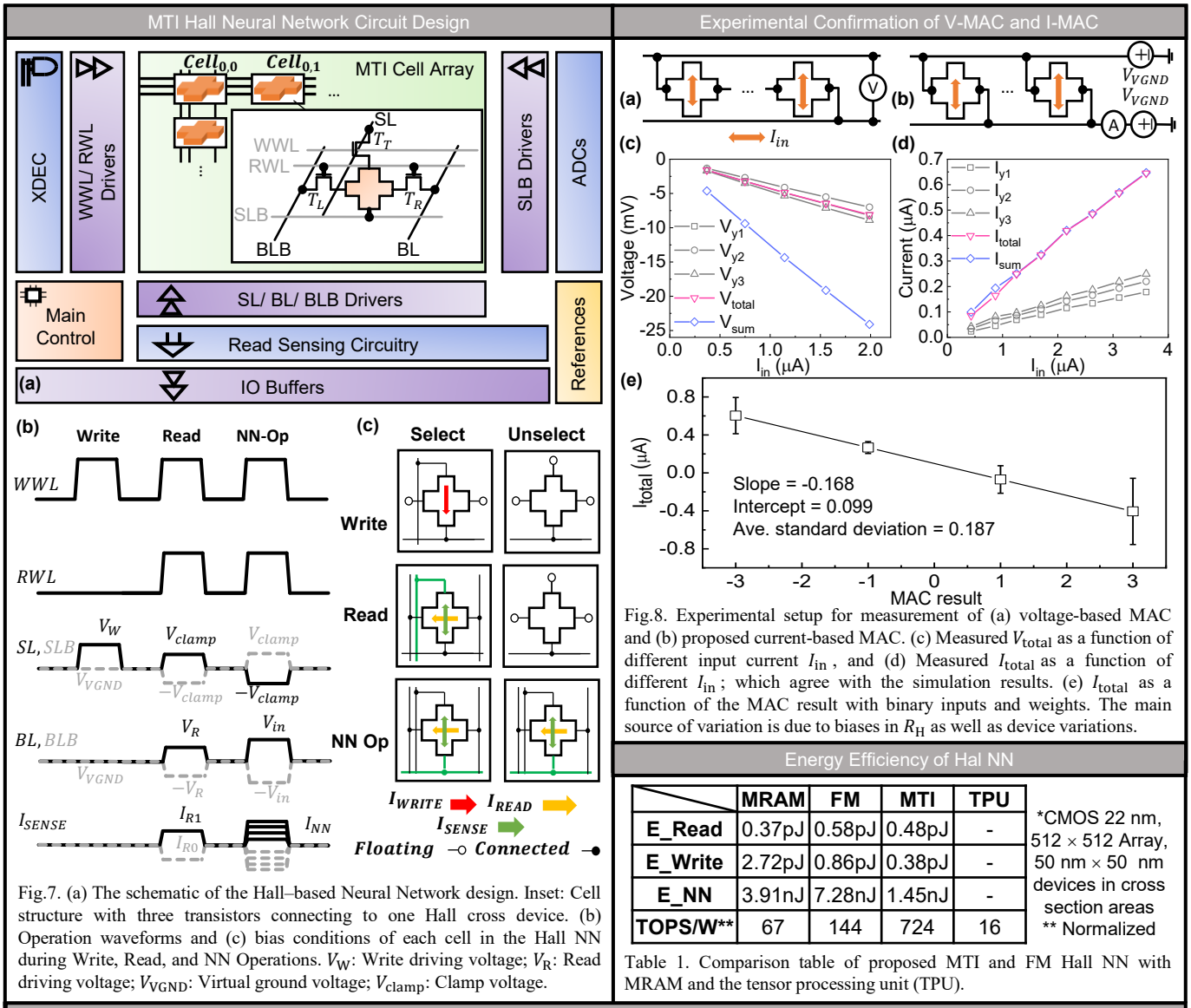


Fig.7. (a) The schematic of the Hall–based Neural Network design. Inset: Cell structure with three transistors connecting to one Hall cross device. (b) Operation waveforms and (c) bias conditions of each cell in the Hall NN during Write, Read, and NN Operations. $V_W$: Write driving voltage; $V_R$: Read driving voltage; $V_{VGND}$: Virtual ground voltage; $V_{clamp}$: Clamp voltage.




Fig.8. Experimental setup for measurement of (a) voltage-based MAC and (b) proposed current-based MAC. (c) Measured $V_{total}$ as a function of different input current $I_{in}$, and (d) Measured $I_{total}$ as a function of different $I_{in}$; which agree with the simulation results. (e) $I_{total}$ as a function of the MAC result with binary inputs and weights. The main source of variation is due to biases in $R_H$ as well as device variations.

Energy Efficiency of Hal NN

| | MRAM | FM | MTI | TPU |
|---|---|---|---|---|
| E_Read | 0.37pJ | 0.58pJ | 0.48pJ | - |
| E_Write | 2.72pJ | 0.86pJ | 0.38pJ | - |
| E_NN | 3.91nJ | 7.28nJ | 1.45nJ | - |
| TOPS/W** | 67 | 144 | 724 | 16 |

*CMOS 22 nm, 512 × 512 Array, 50 nm × 50 nm devices in cross section areas
** Normalized

Table 1. Comparison table of proposed MTI and FM Hall NN with MRAM and the tensor processing unit (TPU).

Circuit Simulation Results of Neural Network Operations with I-MAC

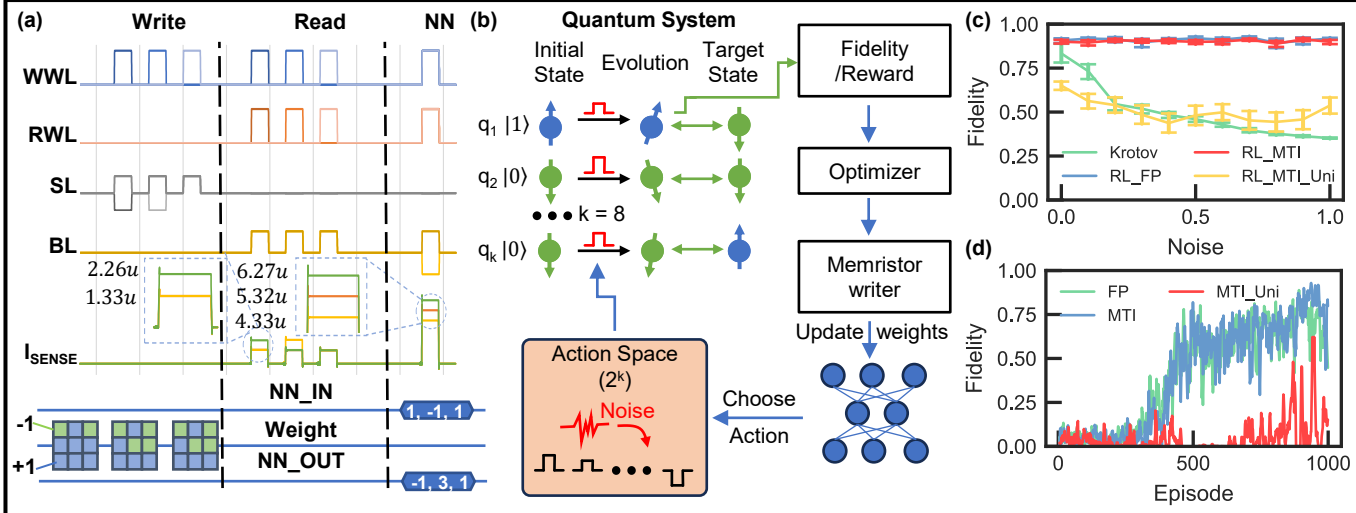


Fig.9. (a) Simulated waveforms of the Hall NN using a 22nm PDK: Write, Read, followed by NN Operation. (b) Preparation multi-qubit states using algorithms to control magnetic field pulse with noise in MTI Hall neural network. (c) Final fidelity of the Krotov method and RL method in full-precision (FP), unipolar and MTI bipolar NNs. (d) Fidelity in each episode of FP NN, unipolar NN and MTI bipolar NN with noise = 0.5.