Ultra-Fast CV-ECRAM-based Analog PIM with Dynamic Retention Compensation Techniques

Seungkun Kim[1*], Jinho Byun[1*], Jung Gyu Min[2*], Jeonghoon Son[1], Jaehee Kim[2], Youngjoo Lee[3], and Seyoung Kim[1]
[1]Department of Materials Science and Engineering, POSTECH, Pohang, Korea; [2]Department of Electrical Engineering, POSTECH, Pohang, Korea; [3]School of Electrical Engineering, KAIST, Daejeon, Korea
E-mail : youngjoo@kaist.ac.kr, kimseyoung@postech.ac.kr

***Abstract*—We demonstrate an analog processing-in-memory (a-PIM)-based AI system with 100× reduced latency by achieving high-speed cup-shaped vertical electrochemical random-access memory (CV-ECRAM), novel device retention compensation algorithm, and tailored RISC-V-based processor for a-PIM. The CV-ECRAM exhibits extremely low cycle-to-cycle variation of 0.014 and device-to-device variation of 0.046, and ultra-fast update is verified at 50 ns pulses. Furthermore, our proposed dynamic max-transfer (DMT) algorithm improves the retention tolerance by more than 25×. Additionally, the RISC-V processor with novel stochastic pulse generation hardware and 8b-parallel vector instruction set achieves 100× latency reduction. Finally, we achieve a training accuracy recovery of 84.9% by the proposed end-to-end a-PIM system.**

***Keywords*—Processing-In-Memory, Electrochemical Random-Access Memory, Tiki-Taka Algorithm, RISC-V Processor**

I. Introduction

Resistive cross-point array-based analog processing-in-memory (a-PIM) architecture has been proposed to accelerate AI training by leveraging analog operations and local data to enable fully parallel deep neural network training without massive data movement between memory and processing unit, achieving a significant reduction in data processing time compared to conventional von Neumann architecture [1]. However, device non-idealities of a-PIM cause significant errors in weight read and update processes, resulting in training failure. To overcome the intrinsic asymmetry of analog devices, one of the major issues, as shown in Fig. 1-a), the Tiki-Taka (TT) algorithm achieves superior training performance as an advanced hardware-aware training method [2].

However, a-PIM remains a critical challenge in the feature vanishing problem due to the large update latency (= decay rate (γ_{decay})) compared to the limited data retention time of practical devices, which may be aggravated by the TT algorithm, as shown in Fig. 1-b). To address this problem, we construct an integrated a-PIM system with three novel strategies, leading to dramatically reduced update latency, as shown in Fig. 1-c). Firstly, we fabricate cup-shaped vertical ECRAM (CV-ECRAM), and demonstrate reliable synaptic characteristics with ultra-fast operation. Furthermore, we introduce a novel retention compensation algorithm using the Hadamard matrix-based transfer method, successfully validated in experimental demonstration. Finally, by designing an advanced RISC-V processor on FPGA, we significantly reduce the system latency. Each advanced scheme significantly complements the corresponding challenge, successfully restoring the training accuracy in total.

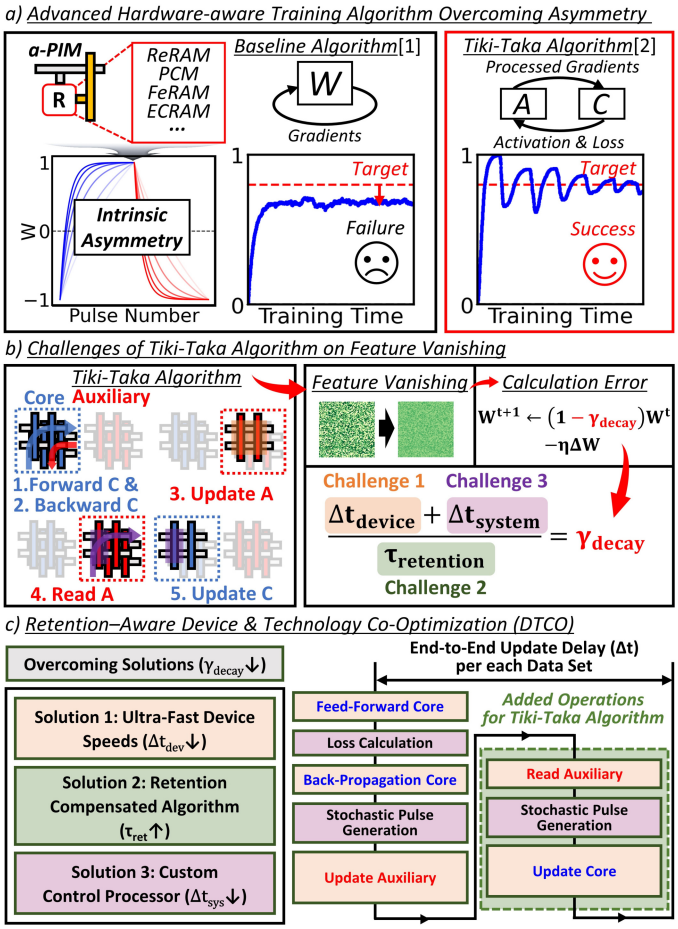

Fig. 1. Challenges and solutions of the advanced training algorithm. a) Tiki-Taka algorithm to overcome intrinsic asymmetry of analog devices. b) Schematic and challenges of Tiki-Taka algorithm on feature vanishing. c) Device retention and total system delay-induced limitations and overcoming solutions through Device & Technology Co-Optimization (DTCO).

*Seungkun Kim, Jinho Byun, and Jung Gyu Min equally contributed to this work. This work was supported by K-CHIPS(Korea Collaborative & High-tech Initiative for Prospective Semiconductor Research) (2410000230, 20024796, 23011-15TC / 2410000283, 20024760, 23008-45FC) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

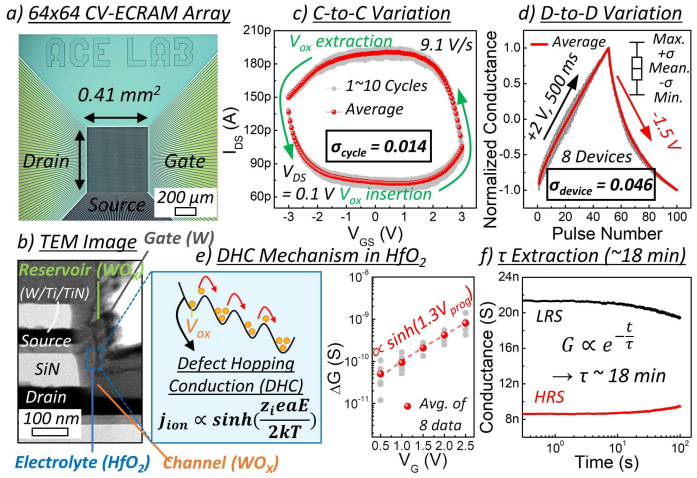

Fig. 2. Reliable characteristics of cup-shaped vertical ECRAM (CV-ECRAM). a) OM image of 64×64 CV-ECRAM array. b) TEM image. c) Transfer curves over 10 cycles. d) Switching curves of 8 devices. e) DHC mechanism. f) Retention characteristics.

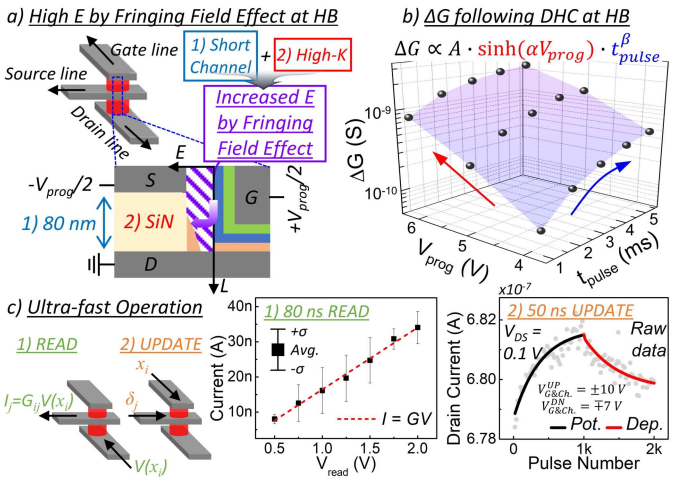

Fig. 3. Fringing field effect & ultra-fast operation speeds. a) Fringing field effect within CV-ECRAM. b) ΔG trend depending on V_{prog} and t_{pulse} at HB scheme. c) Ultra-fast READ & UPDATE speeds.

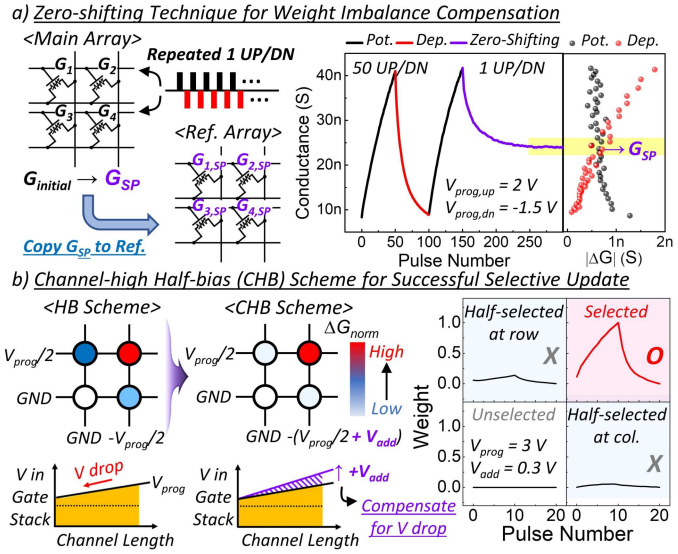

Fig. 4. Novel techniques for hardware array operation. a) Zero-shifting technique for array operation. b) Channel-high half-bias scheme for successful selective update during the outer product computation.

II. Cup-shaped Vertical ECRAM

Fig. 2-a) shows an OM image of the $4F^2$ layout 64×64 CV-ECRAM array with 500 nm hole size and 0.41 mm^2 area. The gate stack is composed of channel (WO_X), electrolyte (HfO_2), and reservoir (WO_X). The short channel length (80 nm) is defined by the distance between source and drain multilayers (W/Ti/TiN), which are separated by the SiN inter-metal layer (IML), as shown in Fig. 2-b). Fig. 2-c) illustrates the transfer characteristics over 10 cycles with ultra-low cycle-to-cycle variation (σ_{cycle} = 0.014), which is attributed to the stable structure of the CV-ECRAM. The CV-ECRAM exhibits anti-clockwise hysteresis due to the migration of oxygen vacancies (V_{OX}) within the channel. Furthermore, uniform switching behavior is demonstrated across 8 individual devices (σ_{device} = 0.046), as depicted in Fig. 2-d). Additionally, the defect hopping conduction mechanism within the HfO_2 electrolyte is verified by conductance change (ΔG) as a function of gate voltage (V_G), as shown in Fig. 2-e). Fig. 2-f) illustrates the retention characteristics with a decay constant (τ) of over 18 minutes. In the CV-ECRAM, as described in Fig. 3-a), under the half-bias (HB) scheme, the fringing field effect is induced by the extremely short channel length (80 nm) and high-k (SiN∼8) IML [3]. Therefore, the ΔG trend at the HB scheme, depending on V_{prog} and pulse width (t_{pulse}), is empirically confirmed to follow the $\sinh(\alpha V_{prog})$ function and a power law (t_{pulse}^{β}), as shown in Fig. 3-b). Additionally, ultra-fast READ (80 ns) and UPDATE (50 ns) operations are experimentally verified, as shown in Fig. 3-c). To demonstrate hardware array operation, the following novel techniques were utilized. Fig. 4-a) describes the zero-shifting technique, which copies the symmetry point (G_{SP}) of main array to reference array for weight imbalance compensation [4]. Furthermore, as described in Fig. 4-b), the channel-high half-bias (CHB) scheme is implemented to achieve successful selective updates during the outer product computation in neural networks by applying an additional voltage (V_{add}) to the channel [5].

III. Dynamic Max-Transfer (DMT) Algorithm

Fig. 5-a) shows the training performance degradation as γ_{decay} increases [6]. For $\gamma_{decay} = 10^{-5}$, weight distributions are highly concentrated around 0 compared to $\gamma_{decay} = 10^{-6}$, confirming that core array's weight decay to 0 impacts training performance. Fig. 5-b) highlights this issue, showing that even at $\gamma_{decay} = 10^{-6}$, increasing the transfer period results in degraded accuracy, whereas at $\gamma_{decay} = 10^{-5}$, increasing the number of transfers improves performance. Transferring along rows shows improved performance than that of columns due to the frequent transfer caused by its shorter length, as shown in 5-c). Fig. 6 introduces the DMT algorithm, addressing decay issues by triggering more frequent updates without additional time cost delay compared to unit-vector based transfer in the original algorithm [2]. In the Tiki-Taka algorithm, transferring the accumulated weight gradients from auxiliary array to core

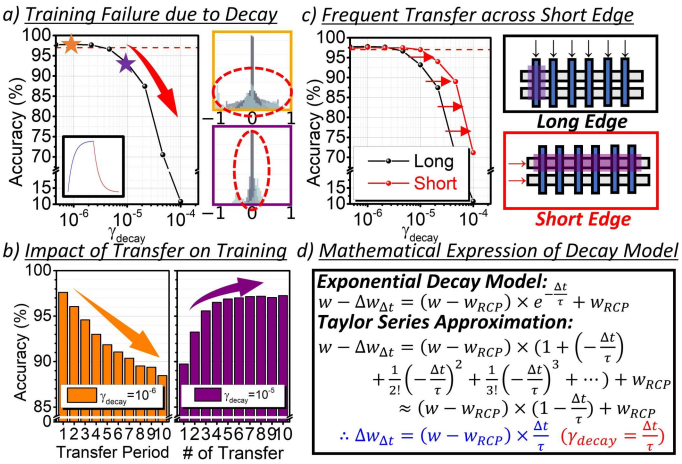

Fig. 5. Analysis of retention-induced training failure. a) Training failure occurs as increasing the decay rate and weight distributions with decay rate of 10^{-6} and 10^{-5}. b) Training failure with decay rate of 10^{-6} occurs as increasing the transfer period and accuracy with decay rate of 10^{-5} recovered as increasing the number of transfers. c) More frequent transfer method considering the weight matrix size and the training results sweeping the decay rates. d) Mathematical expression of the weight decay during training.

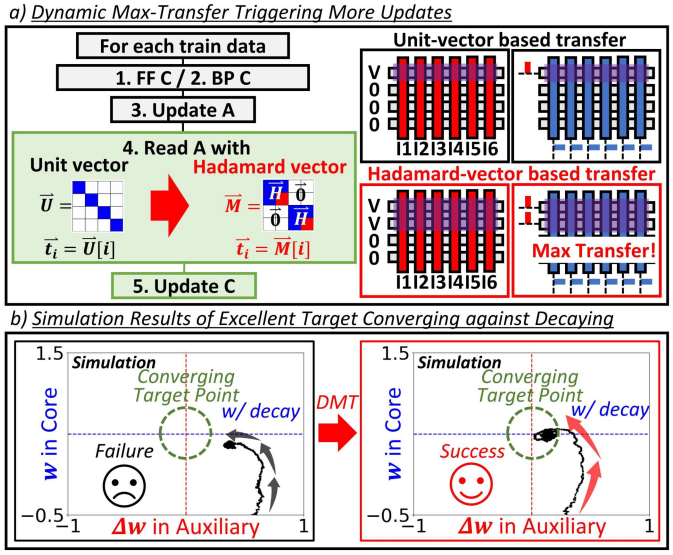

Fig. 6. Dynamic matrix-transfer (DMT) algorithm. a) Schematics of original and advanced transfer methods. b) Target convergence simulations of matrix regression when considering the decay.

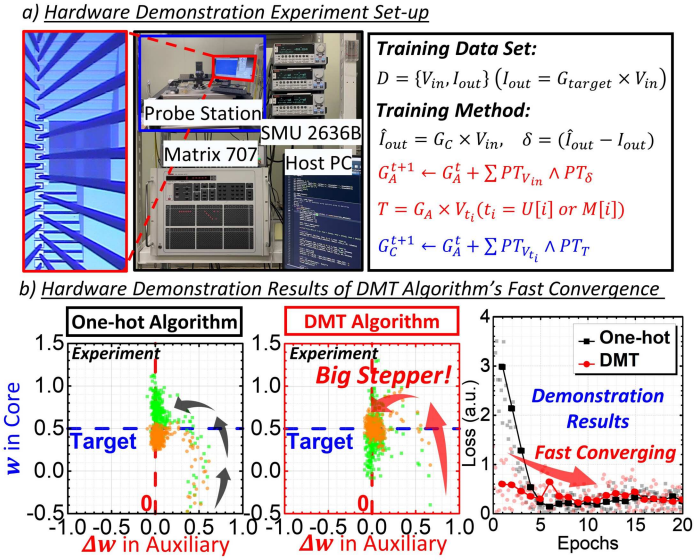

Fig. 7. Hardware demonstration experiment of DMT algorithm. a) Measurement set-up for hardware demonstration. b) Experimental results of one-hot algorithm and DMT algorithm.

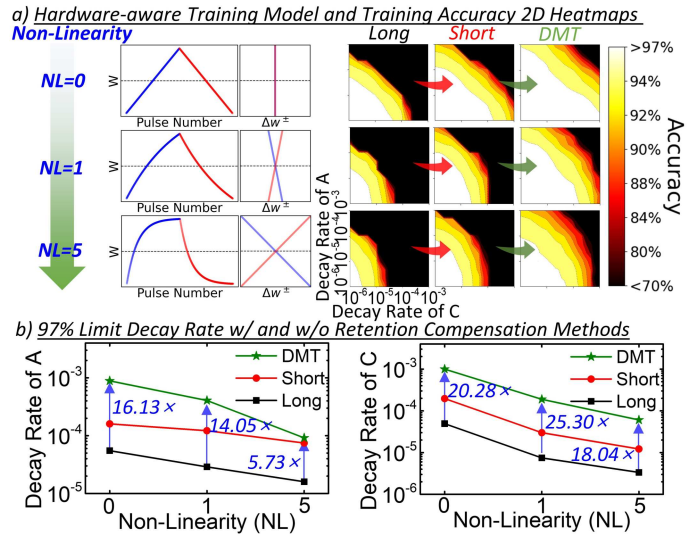

Fig. 8. Large-scale verification of retention compensation effect. a) Training results of different synaptic characteristics and transfer methods. b) 97% limit decay rate of A and C with and without retention compensation methods, assuming the other is fixed at 10^{-6}.

array relies on a predefined transfer vector matrix. While the original method adopts the identity matrix for this purpose, we instead employ block-diagonal Hadamard matrices, which enable more efficient and frequent transfers while preserving gradient information. Target convergence simulations confirm the robustness of the DMT algorithm against weight decay. Fig. 7 validates this algorithm's fast convergence through hardware experiments with CV-ECRAM array, controlling training with software at complex system level. Fig. 8 compares various transfer methods under different asymmetry. The 2D heatmap depicts training accuracy while sweeping γ_{decay} of A and C from 10^{-6} to 10^{-3}. We show that up to 25.30x higher retention tolerance is achieved with the DMT algorithm.

IV. End-to-End System Implementation

Fig. 9-a) depicts the end-to-end AI accelerator, including a RISC-V-based a-PIM control system on Xilinx KC705 FPGA connected with a 64×64 CV-ECRAM array. The stochastic pulse-generation hardware (SPG-HW) allows the simultaneous generation of all update pulses, significantly saving the pulse-generation time compared to the conventional SW-based approach [7]. Note that we share a single linear-feedback shift register (LFSR) to generate parallel random sequences, relaxing the hardware complexity by 88.3% compared to the naïve architecture using parallel LFSRs. As shown in Fig. 9-b), the Boolean AND-operation test with probability 0.5 for both inputs for 100k test set [8] reveals that the proposed method still offers sufficient randomness similar to the SW

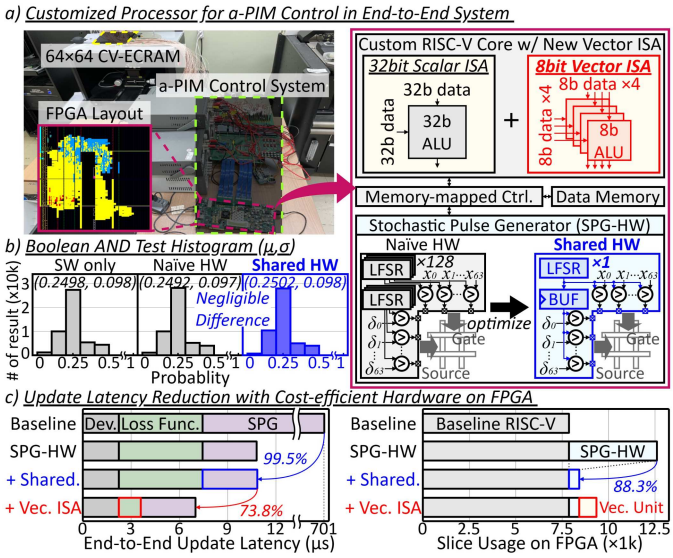

Fig. 9. End-to-End AI acceleration system implementation.

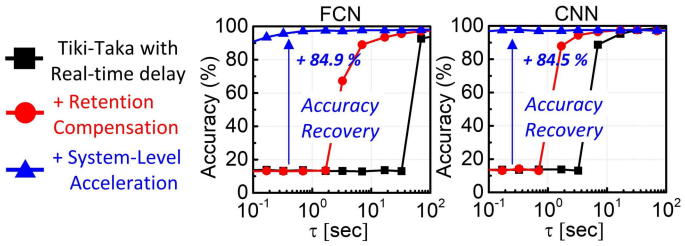

Fig. 10. Excellent accuracy recovery by proposed works.

baseline. In addition, the RISC-V core is newly customized to reduce the processing delay of control signal generation. More specifically, we define custom instructions for 8b-parallel vector processing, reducing the loss calculation latency by 73.8%. As a result, the proposed a-PIM control system reduces the end-to-end update latency by 100 times with minimal hardware overheads shown in Fig. 9-c).

V. Performance Improvement

We evaluated the large-scale neural network training performance using the Modified National Institute of Standards and Technology dataset [9], utilizing IBM's open-source library [6]. The learning rates for the Tiki-taka algorithm were adjusted to 0.01 from C to A and 0.005 from A to C. To verify the operability across multiple models, we evaluated the performance improvement of an FCN model configured as 784-256-128-10 and a CNN model consisting of two convolutional layers with 5×5 kernels, followed by a fully connected structure of 512-128-10. Fig. 10 shows the accuracy recovery throughout each step, providing up to 84.9% recovery on given τ, verifying our work's outstanding performance. Table. I compares the proposed work to the reference systems, showing our work's advance in a-PIM system construction.

VI. Conclusion

In this paper, we present an end-to-end a-PIM AI system that achieves a 100× reduction in latency through three key innovations: high-speed CV-ECRAM, a retention-aware DMT algorithm, and a customized RISC-V-based processor. The CV-ECRAM exhibits excellent reliability, with minimal σ_{cycle} (0.014) and σ_{device} (0.046), and supports ultra-fast updates down to 50 ns. Our DMT algorithm improves retention tolerance by more than 25×. The processor incorporates stochastic pulse generation and an 8-bit parallel vector instruction set to further reduce latency. Ultimately, the proposed system enables a training accuracy recovery up to 84.9%.

TABLE I. Comparison of the proposed a-PIM system with previous works.

	[10]	[11]	[12]	**This work**
Device Structure	O-ion (Vertical)	Li-ion (Vertical)	H-ion (Planar)	**O-ion (Vertical)**
Operation Speeds	10 μs	10 ms	20 ms	**50 ns**
Array Size	1 x 2	32 x 32	2 x 3	**64 x 64**
Retention Tolerance	–	–	–	**25.3x improvement**
System Integration	–	–	Perf. Not reported	**100x latency reduction**

References

[1] T. Gokmen and Y. Vlasov, "Acceleration of deep neural network training with resistive cross-point devices: Design considerations," *Frontiers in neuroscience*, vol. 10, p. 333, 2016.

[2] Z. Wu, T. Gokmen, M. Rasch, and T. Chen, "Towards exact gradient-based training on analog in-memory computing," *Advances in Neural Information Processing Systems*, vol. 37, pp. 37 264–37 304, 2024.

[3] Y.-H. Hsiao, H.-T. Lue, T.-H. Hsu, K.-Y. Hsieh, and C.-Y. Lu, "A critical examination of 3d stackable nand flash memory architectures by simulation study of the scaling capability," in *2010 IEEE International Memory Workshop*. IEEE, 2010, pp. 1–4.

[4] K. Noh, H. Kwak, J. Son, S. Kim, M. Um, M. Kang, D. Kim, W. Ji, J. Lee, H. Jo *et al.*, "Retention-aware zero-shifting technique for tiki-taka algorithm-based analog deep learning accelerator," *Science Advances*, vol. 10, no. 24, p. eadl3350, 2024.

[5] S. Kim, J. Son, H. Kwak, and S. Kim, "Accurate weight update in an electrochemical random-access memory based cross-point array using channel-high half-bias scheme for deep learning accelerator," *Advanced Electronic Materials*, vol. 9, no. 12, p. 2300476, 2023.

[6] M. J. Rasch, D. Moreda, T. Gokmen, M. Le Gallo, F. Carta, C. Goldberg, K. El Maghraoui, A. Sebastian, and V. Narayanan, "A flexible and fast pytorch toolkit for simulating training and inference on analog crossbar arrays," in *2021 IEEE 3rd international conference on artificial intelligence circuits and systems (AICAS)*. IEEE, 2021, pp. 1–4.

[7] F. Aguirre, A. Sebastian, M. Le Gallo, W. Song, T. Wang, J. J. Yang, W. Lu, M.-F. Chang, D. Ielmini, Y. Yang *et al.*, "Hardware implementation of memristor-based artificial neural networks," *Nature communications*, vol. 15, no. 1, p. 1974, 2024.

[8] S. Heo, D. Kim, W. Choi, S. Ban, O. Kwon, and H. Hwang, "Experimental demonstration of probabilistic-bit (p-bit) utilizing stochastic oscillation of threshold switch device," in *IEEE Symp. on VLSI Technology and Circuits*. IEEE, 2023, pp. 1–2.

[9] L. Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE signal processing magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[10] C. Lee, W. Choi, M. Kwak, S. Kim, and H. Hwang, "Excellent synapse characteristics of 50 nm vertical transistor with wox channel for high density neuromorphic system," in *Symp. on VLSI Technology*, 2021, pp. 1–2.

[11] J. Lee, R. D. Nikam, D. Kim, and H. Hwang, "Highly scalable (30 nm) and ultra-low-energy (~5fj/pulse) vertical sensing ecram with ideal synaptic characteristics using ion-permeable graphene electrodes," in *2022 International Electron Devices Meeting (IEDM)*, 2022, pp. 2.2.1–2.2.4.

[12] E. R. Van Doremaele, T. Stevens, S. Ringeling, S. Spolaor, M. Fattori, and Y. van de Burgt, "Hardware implementation of backpropagation using progressive gradient descent for in situ training of multilayer neural networks," *Science Advances*, vol. 10, no. 28, p. eado8999, 2024.