First Demonstration of Uncertainty Quantification based on In-situ Entropy Source Achieved by Nonlinear Domain Wall Array

Tianci Cai[#12], Jiahe Li[#1], Honghu Yang[12], Jinghao Jiang[12], Yue Cao[12], Qiqiao Wu[12], Chengshuo yu[1], Keji Zhou[1], Jun Jiang*[1], Jianguo Yang*[12], Qi Liu[1], and Ming Liu[1]

[1]Fudan University, Shanghai, China; [2]Zhangjiang Laboratory, Shanghai, China; *Email: junjiang@fudan.edu.cn; yangjg@zjlab.ac.cn

#equally contribution author

***Abstract*—Analog Compute-in-Memory (ACIM) based on Non-Volatile Memory (NVM) promises efficient inference acceleration for Bayesian Neural Networks (BNNs), but suffers from complex Gaussian Random Number Generator (GRNG) circuits and low weight array utilization. To address these challenges, we propose a novel Dual-mode In-situ Compute Engine (DICE) for Bayesian inference. For the first time, a 16×16 Domain Wall Memory (DWM) array is experimentally demonstrated. The fabricated DWM exhibits a large on/off ratio (>10^4), long retention time (>10^4s at 150°C), and high endurance cycles (>10^9 cycles). Crucially, the proposed DWM array performs dual functions: executing in-memory Matrix-Vector Multiplication (MVM) and serving as an in-situ entropy source. Leveraging the stochastic lateral expansion of ferroelectric domains, we develop an efficient TRNG achieving a throughput of 100 Mbps and an energy efficiency of 0.031 pJ/bit. The proposed DICE system achieves an inference accuracy of 98.1% on the MNIST dataset and demonstrates an Area Under the ROC Curve (AUC) score of 0.95.**

***Keywords*—DWM, BNN, CIM, TRNG**

I. Introduction

Unlike conventional Convolutional Neural Networks (CNNs) that rely on deterministic weights and often suffer from overconfident predictions, BNNs offer the distinct advantages of uncertainty quantification (**Fig. 1a**). While ACIM architectures based on emerging NVMs promise high parallelism and energy efficiency for such workloads, a critical hardware bottleneck emerges in BNN inference: the system must concurrently execute accurate MVM and generate massive stochastic Gaussian weights [1-2]. The intrinsic conflict between required deterministic accuracy and stochastic variation restricts NVM devices to a single operation mode during inference [3]. Consequently, the substantial hardware overhead of GRNGs and the low utilization of NVM arrays in multiple sampling iterations (T) severely compromise system efficiency.

This work resolves this fundamental tradeoff through a comprehensive device-system co-optimization strategy (**Fig. 1b**). At the device level, we first fabricated a 16×16 DWM array. For deterministic operations, DWM exhibits intrinsic nonlinearity, an excellent on/off ratio, and robust retention, making it suitable for precise MVM operations [4]. For stochastic operations, the random dynamics of the ferroelectric domain serves as an ideal entropy source. By dynamically switching between a low read voltage (V_r) for high-precision MVM and a high V_r to trigger stochastic domain lateral expansion, we implement a unified, dual-mode BNN inference platform. At the system level, the DICE architecture allows the NVM weight array to be fully reused as an in-situ entropy source. Coupled with a compact Entropy Extraction Circuit (EEC), the proposed DICE architecture drastically reduces GRNG area overhead. Consequently, the proposed DWM-based design achieves superior dual-mode performance compared to state-of-the-art implementations.

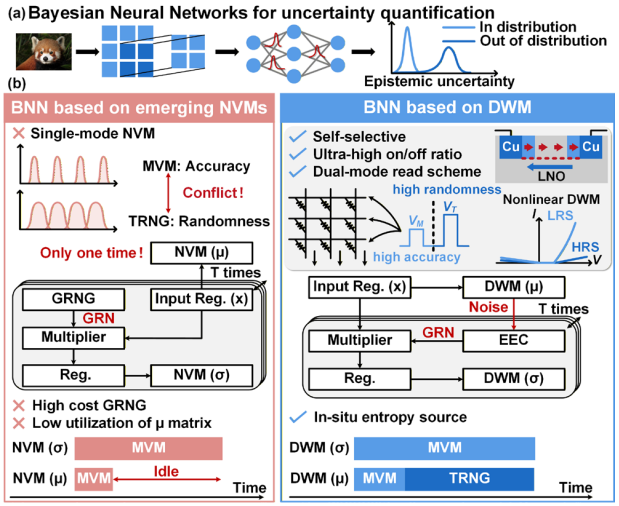


Fig. 1. (a) Applications of Bayesian neural networks for uncertainty quantification via epistemic uncertainty. (b) Challenges of traditional BNN CIM accelerators based on emerging NVMs hardware and features of the proposed BNN architecture based on DWM.

II. The Fabrication and Properties of DWM

A. Device Architecture and Basic Properties

Fig. 3 shows the fabrication process of the DWM array. On the surface of an X-cut LNO sample with a thickness of 500 nm, a 130 nm-thick photoresist was spin-coated. Predefined device patterns were then written using an electron beam lithography machine (EBL JEOL 6300FS). Subsequently, Cr was deposited as an etch mask using a thermal resistance evaporator (NANO 36, Kurt J. Lesker). The excess chromium layer outside the exposed areas was removed via a lift-off process. A reactive ion etching (RIE, SAMCO) process (SF6 45 sccm, Ar 5 sccm, 5 min) was employed to plasma-etch the LNO, forming mesa-like LNO cell between adjacent trenches with an etch depth of ≈100 nm. Wet etching was then performed using an SC-1 solution (H_2O : H_2O_2 : $NH_3 \cdot H_2O$ = 5 : 1 : 1) to remove LNO by-products, followed by the removal of the Cr mask using a ceric ammonium nitrate solution. A 200 nm-thick layer of Cu was deposited by magnetron sputtering (PVD-75, Kurt J. Lesker) to fill all trenches (serving as side electrodes). The excess Cu film outside the trenches on the surface was removed by chemical mechanical polishing (TriboLab CMP, Bruker), thereby forming cells and WL. A 100 nm thick silicon nitride (SiN) layer was subsequently deposited as an isolation layer.

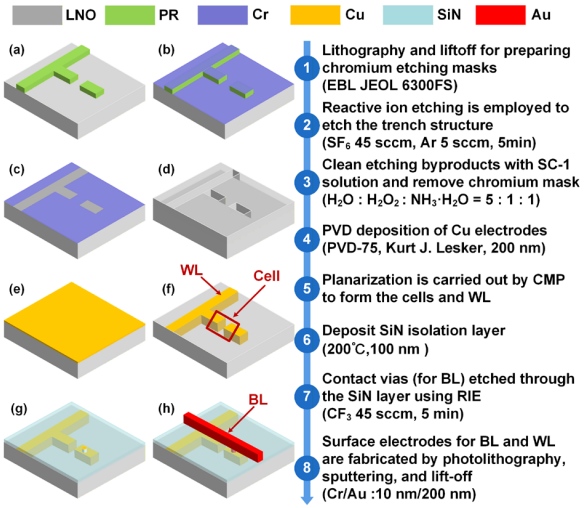


Fig. 2. Process flow of the 16×16 nonlinear LNO domain wall crossbar array fabrication.

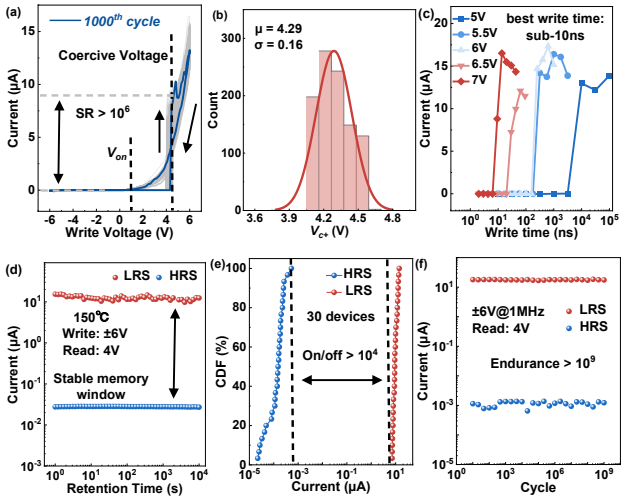


Fig. 3. (a) 1,000 sweeps of *I-V* curves. (b) Statistical positive coercive voltage distribution. (c) Dependence of switching time on write voltage. (d) Retention performance over 10,000s at 150°C. (e) D2D uniformity across 30 devices. (f) Endurance performance under 1 MHz operation cycles.

Contact vias for the BL and WL were prepared through photolithography and RIE. Finally, surface electrodes for the device' s BL and WL are fabricated by photolithography, magnetron sputtering (Cr/Au :10 nm/200 nm), and lift-off.

Fig. 3a displays 1,000 successive DC *I-V* sweep curves. Under a voltage sweep from -6 to +6V, the device exhibits an abrupt transition from the OFF-state to an ON-current of 7.5 μA at a positive coercive voltage (V_{c+}) of 4.3V, achieving a selective ratio (SR) exceeding 10^6. The statistical distributions of V_{c+} is well-fitted by a Gaussian function with a standard deviation of 0.16, confirming the excellent switching uniformity of the fabricated array (**Fig. 3b**). **Fig. 3c** illustrates the dependence of the domain switching time (t_0) on the write voltage (V_w). t_0 exponentially decreases as V_w increases, achieving a remarkable sub-10ns write speed. **Fig. 3d** presents the high-temperature retention characteristics measured at 150 °C. After applying ±6V write pulses, the ON-state read current (monitored at a read voltage of 4V) degrades by only 20% over 10^4s, demonstrating robust non-volatility.

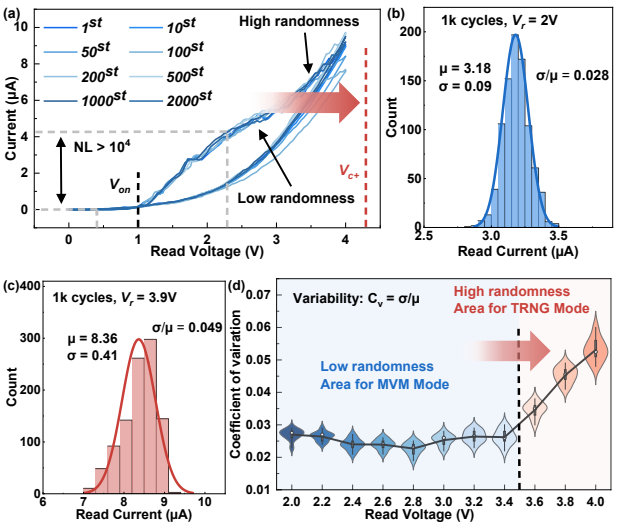


Fig. 4. (a) On-state read currents over 2,000 sweeps, indicating high randomness under high V_r and low randomness under low V_r. (b) (c) Distribution of On-state currents under 2V and 3.9V. (d) Coefficient of variation under different read voltages.

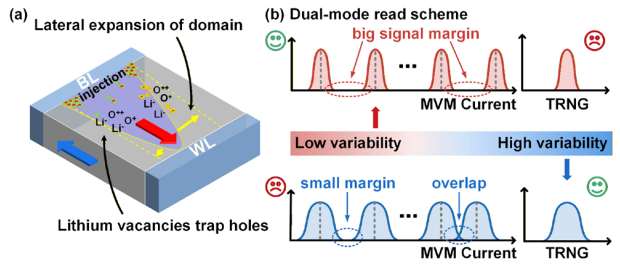


Fig. 5. (a) Schematic illustration of the current variation mechanism under high read voltage. (b) Schematic of DRS based on DWM characteristics.

Furthermore, excellent cell-to-cell uniformity is verified by the tight cumulative probability distribution of states across 30 randomly selected cells (**Fig. 3e**). Notably, the array achieves an outstanding endurance of 10^9 fatigue cycles under 1 MHz bipolar pulses of ±6V without significant degradation.

B. Dual-mode Read Scheme (DRS)

Fig. 4a analyzes the read current variation over 2,000 sweep cycles, revealing a highly voltage-dependent behavior with a nonlinearity (NL) exceeding 10^4. Under a low V_r, the device exhibits deterministic behavior with exceptional current uniformity (**Fig. 4b**). Conversely, as V_r increases, prominent stochasticity emerges (**Fig. 4c**). This transition is quantitatively captured by the coefficient of variation (C_v) in **Fig. 4d**, which clearly separates two distinct operational regimes: a low-variability deterministic regime at lower voltages (ideal for MVM inference) and a high-variability stochastic regime at elevated voltages.

To decode these macroscopic electrical responses, **Fig. 5a** illustrates the underlying micro-physical mechanisms, which conceptually lay the foundation for our proposed DRS (**Fig. 5b**). Fundamentally, the stochasticity of the interfacial conductive Domain Wall (DW) current, triggered upon each application of a high V_r, is a cascaded dynamic process [7]. It originates primarily from two stages. First, during the initial phase, domain nucleation near the interface exhibits profound

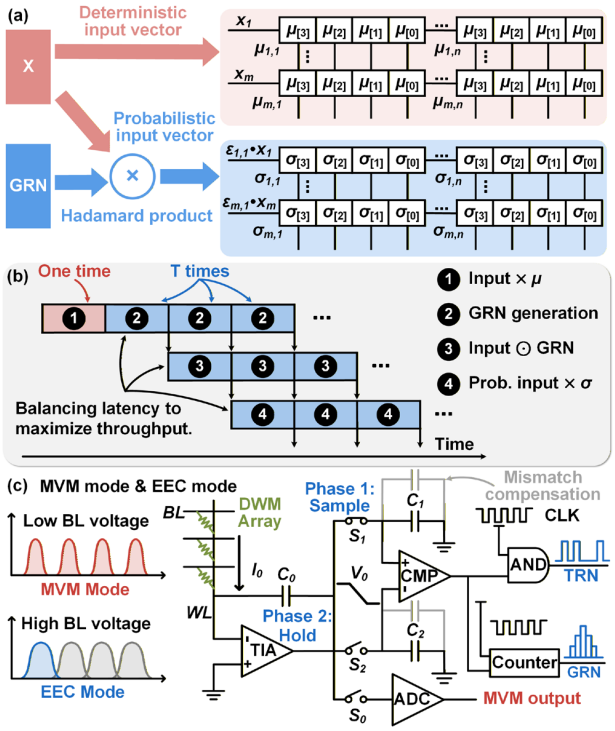


Fig. 6. (a) The proposed Bayesian neural network computation paradigm. (b) The pipeline chart of DICE system. (c) EEC of the TRNG and MVM circuit.

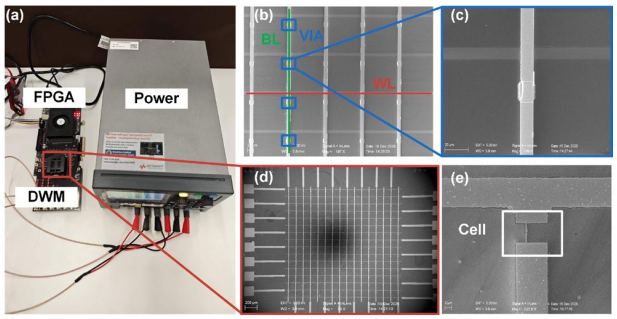
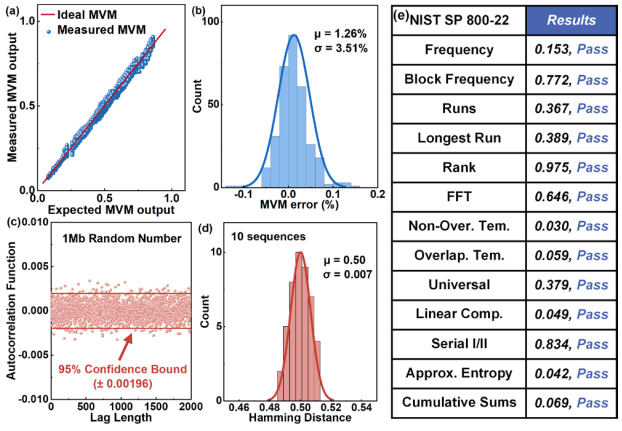


Fig. 7. (a) The hardware test platform of DICE system. (b) (c) (d) (e) SEM image of the fabricated 16×16 DWM array.



(e) NIST SP 800-22	Results
Frequency	0.153, Pass
Block Frequency	0.772, Pass
Runs	0.367, Pass
Longest Run	0.389, Pass
Rank	0.975, Pass
FFT	0.646, Pass
Non-Over. Tem.	0.030, Pass
Overlap. Tem.	0.059, Pass
Universal	0.379, Pass
Linear Comp.	0.049, Pass
Serial I/II	0.834, Pass
Approx. Entropy	0.042, Pass
Cumulative Sums	0.069, Pass

Fig. 8. (a) (b) Measurement results of MVM. (c) ACF of a random sequence at 95% confidence level. (d) The inter-sample Hamming distance among 10 random sequences demonstrates good uniqueness. (e) NIST SP 800-22 randomness test results of the DWM based TRNG.

spatiotemporal stochasticity—fluctuating in both exact spatial location and micro-timing under a given macroscopic electric field. Second, during the propagation phase, the needle-like forward growth of these domains is intrinsically non-deterministic. The domain growth trajectories are highly susceptible to local energy landscapes, where pinning effects induced by randomly distributed oxygen vacancy defects continuously alter and deflect the DW expansion paths.

Consequently, each high-V_r perturbation yields a uniquely random conductive profile, providing an excellent entropy source while preserving low-V_r determinism.

III. Hardware Implementation and Benchmark

A. Dual-mode In-situ Compute Engine

Based on the measured electrical characteristics of the DWM array, we experimentally demonstrate a novel DICE uncertainty quantification system. **Fig. 6a** shows the conceptual principle of the proposed BNN accelerator. By decoupling the random and deterministic components, the weights physically stored within the array remain static. The core computation for Bayesian inference is formulated as follows:

$$y_i^{(t)} = \sum_j \left(\varepsilon_{ij}^{(t)}\sigma_{ij} + \mu_{ij}\right)\cdot x_i, \quad \varepsilon_{ij}^{(t)} \sim N(0,1) \qquad (1)$$

where x_i denotes the input, while μ_{ij} and σ_{ij} represent the mean and standard deviation of the Gaussian weights, respectively. $\varepsilon_{ij}^{(t)}$ is a standard Gaussian random variable. Notably, during the T sampling iterations, only the stochastic term $\varepsilon_{ij}^{(t)}$ is updated T times, whereas the deterministic components (μ_{ij}, σ_{ij}) remain strictly constant.

Fig. 6b details the execution pipeline of the proposed Bayesian inference. Since the MVM between the input vector and the μ weight matrix is executed only once per inference, the μ matrix array is dynamically repurposed as an in-situ entropy source for the subsequent Gaussian Random Number (GRN) generation phase. This architecture significantly maximizes array utilization and entirely eliminate the hardware overhead of external physical entropy source.

Leveraging the highly reliable stochasticity of the aforementioned DRS, we experimentally validated a compact discrete Gaussian sampler circuit. As depicted in **Fig. 6c**, the random bit generation circuit utilizes a trans-impedance amplifier (TIA) to capture the stochastic read current (I_0), followed by a sample-and-hold dual-capacitor stage. A comparator then digitizes these current variations into random binary bits. To effectively mitigate process-induced capacitor mismatches, tunable calibration capacitors are integrated to guarantee an unbiased probability distribtuion ($p = 50\%$).

B. Efficiency Evalutaion and Benchmarks

The experimental setup is shown in **Fig. 7a**, while **Fig. 7b-e** provide Scanning Electron Microscope (SEM) images of the fabricated 16×16 array. As depicted in **Fig. 8a**, the measured 6-bit MVM output currents exhibit a highly linear correlation with the ideal values. To validate the array-level computational fidelity, uniformly distributed random voltage vectors (ranging from 2 to 3.4 V) were applied as inputs. The resulting error distribution (**Fig. 8b**) confirms the high

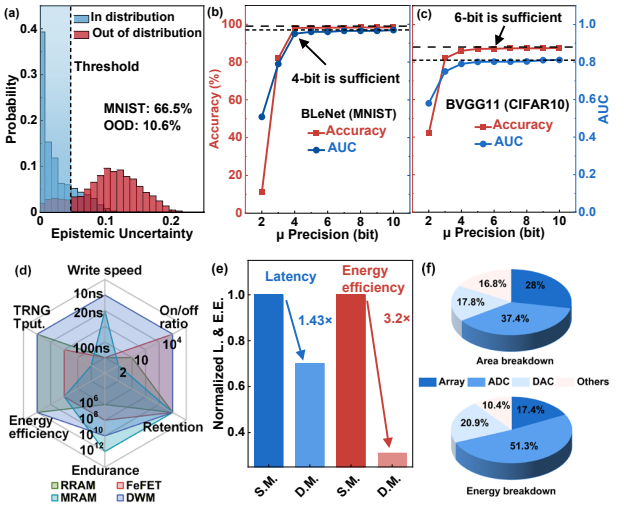


Fig. 9. (a) Distribution of epistemic uncertainty of MNIST and Out-Of-Distribution (OOD) images. (b) (c) Impact of μ precision on classification and OOD detection performance. (d) Benchmark radar chart. (e) System performance comparison. (f) Area and energy breakdown of the proposed DICE architecture.

precision of the hardware implementation. **Fig. 8c** evaluates the AutoCorrelation Function (ACF) of the measured data over a 2,000-step lag on a 1 Mb bitstream, maintaining a tight confidence bound of ±0.00196. The Hamming distance analysis (**Fig. 8d**) demonstrates excellent inter-sequence uniqueness. Furthermore, the DWM-based TRNG successfully passes the NIST SP 800-22 test suite, verifying its high-quality randomness (**Fig. 8e**).

To validate the proposed methods, we evaluated the system with two representative BNNs, Bayesian LeNet5 and Bayesain VGG11, for both classification and OOD detection tasks utilizing the MNIST and CIFAR10 datasets. **Fig. 9a** illustrates the resulting epistemic uncertainty distributions. By evaluating this epistemic uncertainty against a predefined threshold, the Bayesian LeNet model effectively differentiates between in-distribution and OOD data. **Fig. 9b-c** analyze the impact of varing the bit-precision of μ on classification tasks and OOD detection, indicating that 4-bit precision is sufficient for both network architectures. Furthermore, compared to state-of-the-art NVMs, our solution exhibits superior hardware characteristics and provides a highly efficient entropy source (**Fig. 9d**). System performance is compared between dual-mode (D.M.) and single-mode (S.M.) operations for Bayesian inference, demonstrating the great potential of DRS (**Fig. 9e**). **Fig. 9f** illustrates the area and energy breakdown of the proposed DICE with DWM array size of 128×1024.

IV. Conclusion

In this work, for the first time, we implemented a 16×16 DWM array and proposed DICE architecture based on an in-situ entropy source. Leveraging the dual-mode characteristics of DWM devices, our Bayesian inference architecture achieves excellent performance in uncertainty quantification. **Table I** benchmarks the reported system performance against prior designs, showcasing its strong potential.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (NSFC) under grants 62488101.

TABLE I. Performance Comparison With Other Works

Solutions	Device	Entropy Source	NIST Passed	Tput. (Mbps)	Efficiency (pJ/bit)	MNIST Accuracy
IEDM 19[5]	RRAM	Soft Write	800-22 800-90B	1.32 (per cell)	11	N/A
IEDM 20[13]	RRAM	Resistance Perturbation	800-22 800-90B	3.4 (per cell)	3.51	N/A
IEDM 23[1]	MTJ	Telegraphic fluctuation	-	1	0.002	N/A
JSSC 23[2]	CMOS	LFSR	Not a TRNG	7310	1.08	94.9%
IEDM 23[3]	3D 16-layer Fe-diode	Shot noise of Fe-diode	800-22	-	-	95.3%
IEDM 24[6]	FeFET	Domain Disturb	800-22 800-90B	50 (per cell)	6.08	N/A
IEDM 24[12]	MTJ	VCMA	800-22	29.8k	0.089	N/A
This work	DWM	Domain Expansion	800-22	100 * (per cell)	0.031	98.1%

* Up to 16 devices are sampled in parallel

References

[1] Singh, Nihal Sanjay, et al. "Hardware demonstration of feedforward stochastic neural networks with fast MTJ-based p-bits." *2023 International Electron Devices Meeting (IEDM)*. IEEE, 2023.

[2] Dorrance, Richard, et al. "An energy-efficient bayesian neural network accelerator with cim and a time-interleaved hadamard digital grng using 22-nm finfet." *IEEE Journal of Solid-State Circuits* 58.10 (2023): 2826-2838.

[3] Gong, Tiancheng, et al. "First demonstration of a Bayesian machine based on unified memory and random source achieved by 16-layer stacking 3d Fe-diode with high noise density and high area efficiency." *2023 International Electron Devices Meeting (IEDM)*. IEEE, 2023.

[4] Sun, Jie, et al. "Roadmap for ferroelectric domain wall memory." *Microstructures* 4.1 (2024): N-A.

[5] Lin, Bohan, et al. "A high-speed and high-reliability TRNG based on analog RRAM for IoT security application." *2019 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2019.

[6] Shao, Hanyong, et al. "First demonstration of high throughput and reliable homomorphic encryption using FeFET arrays for resource-limited IoT clients." *2024 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2024.

[7] Lu, Haidong, et al. "Electrical tunability of domain wall conductivity in LiNbO3 thin films." *Advanced Materials* 31.48 (2019): 1902890.

[8] Zhang, Wen Jie, et al. "Nonvolatile ferroelectric LiNbO 3 domain wall crossbar memory." *IEEE Electron Device Letters* 44.3 (2023): 420-423.

[9] Hu, Di, et al. "Cryogenic ferroelectric LiNbO 3 domain wall memory." *IEEE Electron Device Letters* 45.3 (2023): 380-383.

[10] Shan, Linbo, et al. "Monolithically 3D integrated memristive bayesian neural network for intelligent motion planning." *2024 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2024.

[11] Fu, Yuyang, et al. "First Demonstration of ReRAM-based In-Memory Searchable Encryption System with Randomness-Controllable Programming." *2025 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2025.

[12] Xiao, Zhihua, et al. "In-memory neural stochastic differential equations with probabilistic differential pair achieved by in-situ p-bit using CMOS integrated voltage-controlled magnetic tunnel junctions." *2024 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2024.

[13] Lin, Bohan, et al. "A high-performance and calibration-free true random number generator based on the resistance perturbation in RRAM array." *2020 IEEE international electron devices meeting (IEDM)*. IEEE, 2020.

[14] Shao, Hanyong, et al. "ML-Resistant and Reliable ChainX PUF Based on FeFET Arrays for Resource-Limited IoT Security." *2025 IEEE European Solid-State Electronics Research Conference (ESSERC)*. IEEE, 2025