2025 IEEE 68th International Midwest Symposium on Circuits and Systems (MWSCAS)
August 10-13, 2025. Lansing, Michigan, USA

2025 IEEE 68th International Midwest Symposium on Circuits and Systems (MWSCAS) | 979-8-3315-8934-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/MWSCAS53549.2025.11244385

Toward Variation-Tolerant Ferroelectric Neural Computing

Special Session Paper

Arnob Saha
School of EECS
Penn State University
University Park, PA 16802, USA
afs6507@psu.edu

Md Zesun Ahmed Mia
School of EECS
Penn State University
University Park, PA 16802, USA
zesun.ahmed@psu.edu

Jiahui Duan
EE Department
University of Notre Dame
Notre Dame, IN 46556, USA
jduan3@nd.edu

Kai Ni
EE Department
University of Notre Dame
Notre Dame, IN 46556, USA
kni@nd.edu

Abhronil Sengupta
School of EECS
Penn State University
University Park, PA 16802, USA
sengupta@psu.edu

***Abstract*—Ferroelectric field-effect transistors (FeFETs) are emerging to be a promising candidate for low-power in-memory compute based platform design for accelerating AI workloads. However, device-circuit non-idealities and variations due to device scaling pose significant challenges. The paper reviews recent developments in this field related to in-depth characterization of such variabilities in industry standard FeFETs and discusses the interplay of non-linearities in programmable conductance states, device-to-device and cycle-to-cycle variations with device scaling. We outline the design of variation-tolerant neuromorphic hardware accelerators through hardware-software-neuroscience co-design strategies that considers such non-idealities during the training process.**

***Index Terms*—Neuromorphic Computing, Ferroelectric Field Effect Transistors, Algorithm-Hardware-Neuroscience Co-design, Variation-Tolerant Design, In-Memory Computing**

I. Introduction

Ever-increasing computational requirements of neural networks to solve complex AI workloads has driven the quest for alternatives beyond conventional CMOS based von-Neumann computing architectures which remain energy and memory bandwidth limited due to repeated bidirectional data transfer between the memory and compute [1]. Inspired by the in-situ memory embedded compute occurring in the brain, recent advances in neuromorphic computing [2], [3] leverage emerging non-volatile memory (eNVM) devices arranged in a cross-array fashion to implement Compute-In-Memory (CIM) systems for parallel evaluation of the dot product computing kernel necessary for hardware acceleration of neural network workloads. While different eNVM devices are being currently explored like resistive random access memory (RRAM), phase-change memory (PCM), magnetic devices, among others [4]; Hafnia (HfO_2)-based ferroelectric field-effect transistors (FeFETs) have shown great promise by dint of its back-end-of-line (BEOL) CMOS compatibility, scalability, ultra-low energy consumption and disengaged read-write operation [5], [6]. Especially, multi-state programming capability in a single FeFET device by gradually modulating its polarization makes it an attractive candidate for analog synaptic applications [7]–[11]. However, inherent device-circuit non-idealities, especially for aggressively scaled devices, challenges its feasibility as a CIM device alternative. Thus, an intertwined hardware-algorithm analysis is critical to understand the non-linear impact of FeFET device variability and operating circuit conditions on computational and algorithm performance under device scaling. Moreover, to counter the non-ideal variations, taking measures for effective variation-tolerant approaches during the training process is necessary to ensure reliable operation. Prior works have considered simulating different random process variations (ferroelectric phase variation, metal work function variation, line-edge roughness effect) to investigate their impact on threshold voltage variations [12] and current variations due to continual reading operations [13]. Moreover, effect of ferroelectric thickness on FeFET device-circuit non-idealities and system accuracy has been reported [14]. Additionally, while there exists prior works on variability-aware algorithmic approaches to design reliable generic memristor technology-based IMC systems, they do not consider technology specific interplay effects of FeFET variations with device properties [15]–[18]. This work reviews recent advancements on extensive characterization analysis of industry-standard FeFET devices for $28nm$ high-K metal gate (HKMG) technology to understand the interplay of device non-idealities (bit precision, non-linearities in programming, device-to-device (D2D) variations, cycle-to-cycle (C2C) variations, data retention, among others) with device dimensions and operating voltage conditions. We review various algo-

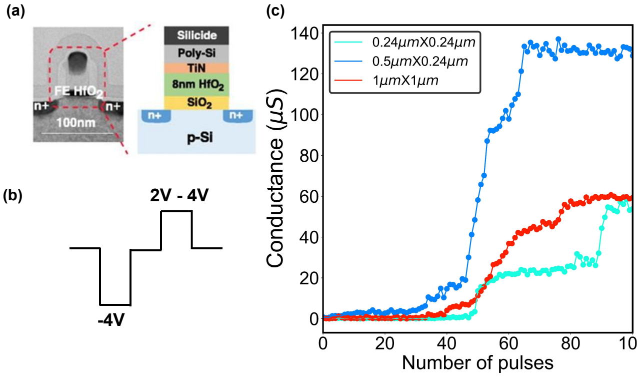

Fig. 1: (a) TEM cross-section and schematic diagram of the FeFET device with $8nm$ thick doped HfO_2 and native oxide layer of $1nm$ thick SiO_2, (b) Applied programming pulse scheme with a reset pulse of $-4V$ and a programming pulse varied from $2V$ to $4V$ with $20mV$ step, (c) Programmed conductance modulation of single $0.24\mu m \times 0.24\mu m$, $0.5\mu m \times 0.24\mu m$ and $1\mu m \times 1\mu m$ devices respectively at a read gate voltage of $1.2V$ [19].

rithmic techniques to combat variations by leveraging recent advancements in deep learning [19], [20]. We conclude with a forward-looking perspective of autonomous self-repair of non-ideal neuromorphic hardware by drawing inspiration from the self-repair role of glial cells in the brain [21].

II. FeFET Non-Ideality Characterization and Algorithm Impact

Hafnia-based FeFET is structurally the same as traditional MOSFETs where the gate stack comprises of doped HfO_2 as ferroelectric layer and SiO_2 as native oxide interlayer. Fig. 1(a) shows the TEM cross-sectional image and schematic portrayal of the device [22], [23]. Three different device sizes (noted in the figure inset) are considered [19]. Fig. 1(b) demonstrates the applied voltage pulse scheme to the gate terminal of the FeFET. The devices are subjected to a high negative reset pulse of $-4V$ to ensure negative polarization state of all the domains resulting in minimum conductance state of the device followed by a programming pulse ranging from $2V - 4V$ with a step size of $20mV$ progressively applied to the gate terminal. Immediately after every programming event, the drain current has been read out to analyze the conductance profile of the devices at a particular operating voltage, as shown in Fig. 1(c), where gradual change in conductance can be observed in larger devices due to the presence of more number of domains, hence higher number of programmable states. On the contrary, scaled devices experience stochastic, discrete and non-linear changes in conductance due to presence of lower number of domains in the ferroelectric layer [24].

FeFETs exhibit probabilistic conductance switching behavior due to inherent stochastic nucleation process that dominates over polarization switching dynamics [25] (see Fig. 2). Fig. 2(a) plots the severity of variations with respect to the applied programming pulses [19], [20]. Lower operating voltage significantly increases conductance variations, especially in scaled devices. As represented in Fig. 2(b), significant C2C variations is present in the device programming profile although it does not seem to depend strongly on device scaling. Furthermore, different identically sized FeFET devices have different orientation of the domains with varying coercive field voltages resulting in D2D variations [26]. As illustrated in Fig. 2(c), such variations increase significantly with device scaling. Device state retention data is shown in Fig. 2(d) where state retention does not seem to be a concern for FeFETs, although there may be minor drifts for some conductance states. In addition to such device-level non-idealities, other types of issues like stuck-at faults can also be a concern where the synaptic elements are permanently fixed at their minimum/maximum conductance state, failing to respond to programming pulses. Such faults can originate from various device-level defects, including trapped charges, defects within the ferroelectric layer, or imperfections introduced during the fabrication process [27].

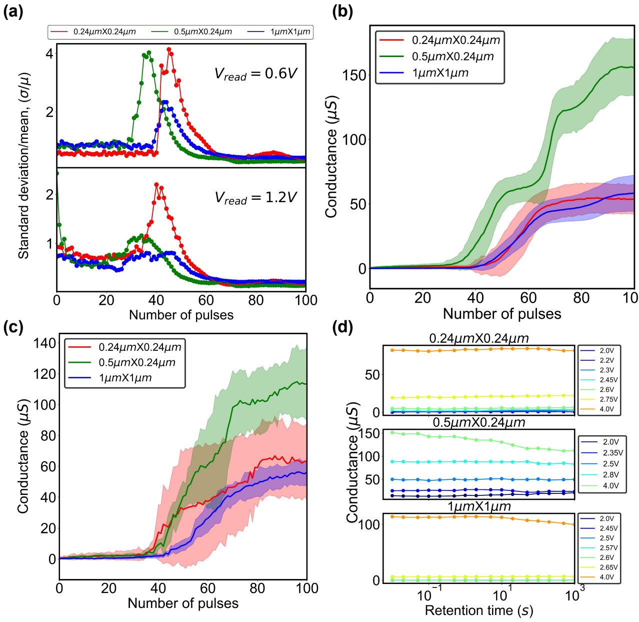

Fig. 2: (a) Severity of variations (σ/μ) as a function of applied programming pulses, (b) Cycle-to-cycle variations for varying device sizes. The solid lines depict the mean variations while the shaded regions represent variations among 50 repeated measurements of a singular device, (c) Device-to-device variations for different device dimensions. The shaded regions indicate the deviations from the mean (solid lines) upon measuring 3 different devices of each dimension, (d) Retention study of few intermediate states for different device sizes [19], [20].

Since there are multiple sources of non-idealities, many of which impacts algorithm-level performance in an intertwined fashion, it is important to assess which factors influence the system-level accuracy to the maximum extent. Recent quantitative analysis [19] performed using NeuroSim [28], a popular CIM-based circuit-level macro model to analyze inference performance, has revealed that combined C2C and D2D variations are the main source of accuracy degradation. For example, as shown in Fig. 3, hardware-constrained performance evaluation of LeNet-5 on MNIST dataset [29] achieves near ideal software-based accuracy considering state discretization effects and programming non-linearities but excluding C2C

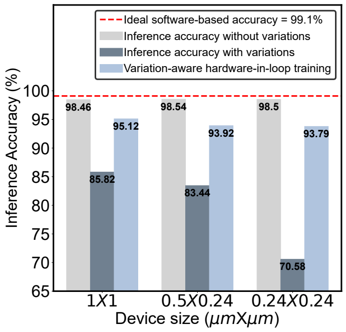

Fig. 3: Comparison between inference performances of the CIM architecture with the presence and absence of device-level variations along with improvement of accuracy achieved by employing variation-aware hardware-in-loop training to mitigate the effect of variations [19].

and D2D variation effects. But, the inference accuracy inclusive of device-level combined C2C-D2D variations drops by $\sim 13\% - 28\%$ for varying device sizes [19]. Typically hardware-in-loop training can be used to combat such accuracy degradation effects where every iteration of the training process considers the noise originating from the hardware in the feedforward propagation of the inputs. As represented in Fig. 3, variation-aware hardware-in-loop training approach can improve the inference accuracy by $\sim 9\% - 23\%$ with maximum accuracy recovered for larger device sizes.

III. Probabilistic Approach to Robust System Design

While hardware-in-the-loop training is a good starting point for combating variations, it is still computationally expensive as the backpropagation updates need to be implemented on-chip in an iterative fashion. In order to adopt a one-time variation-aware training strategy, one can leverage insights from probabilistic deep learning. Bayesian deep learning [30], [31] is an emerging computing paradigm in this landscape wherein a neural network (see Fig. 4) is trained with probabilistic weights with the objective of making the posterior probability distribution as close as possible to the prior. Such Bayesian Neural Networks (BNNs) find use in confidence-critical applications like self-driving cars where some sort of uncertainty quantification from the network is essential. Algorithmic researchers typically consider Gaussian priors. However, in this scenario, one could utilize the characterized device variation data as the intrinsic prior considering interplay effects with device scaling and operating voltages (see Fig. 2(a)) rather than using fixed variation values. Given the prior knowledge, the approximated posterior distributions of weights are learnt iteratively through the "Bayes by Backprop" method [35]. As the training process completes, the mean values of the optimized posterior distributions are considered as the variation-aware trained weights which is then mapped as FeFET conductance in the CIM crossbar system for inference performance evaluation. Thus, the maximization of accuracy and minimization of FeFET reliability issues can be simultaneously handled in the BNN training approach. As shown in Fig. 5, for different operating voltages, inference performance has been evaluated for AlexNet on CIFAR10 dataset. Inference accuracy using variation-aware BNN training approach performs better than the non-Bayesian training framework. Moreover, the inference accuracy using the Bayesian framework considering the intrinsic device-level variation characteristics is comparable to the ideal software accuracy of 85.4% [20] and it outperforms the performance obtained using fixed variations (without considering interplay effects with device size, programmed conductance state or operating voltages).

IV. A Neuroscience-Inspired Outlook for Variation Tolerant Hardware Design

While variation-aware training methods effectively build resilience during the system's training and configuration phase, an exciting and perhaps more profound long-term vision involves equipping neuromorphic systems with the ability to autonomously adapt and repair themselves during ongoing operation. This moves beyond conventional deep learning approaches and delves into largely untapped insights from neuroscience. Intriguingly, the brain's own glial cells, particularly astrocytes—often overlooked in neuromorphic design—are believed to possess innate self-repair capabilities. Recent studies suggest these remarkable cells perform a crucial function akin to activity normalization [21], [32]–[34]. By monitoring and modulating synaptic activity, astrocytes strive to maintain baseline neuronal firing rates, ensuring stable network function even when faced with perturbations like synaptic failures or variations [21].

This biological principle resonates strongly with the challenges associated with device-circuit non-idealities in IMC system design. As detailed in Section II, such hardware is susceptible to variations and faults, causing synaptic elements (the FeFETs) to hold incorrect values. These deviations from intended weights can corrupt the in-memory computation, leading to erroneous system-level results. Herein lies the transformative potential: could we engineer astromorphic principles into these computing systems? The goal would be to

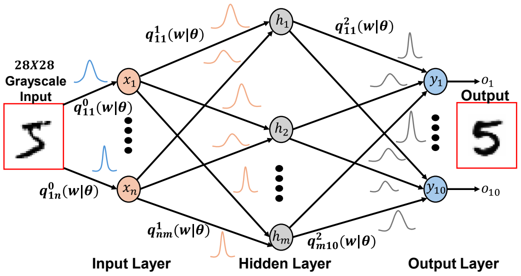

Fig. 4: Bayesian neural network where each weight value has a posterior distribution, $q(w|\theta)$, trained using "Bayes by Backprop" method for given prior weight distribution, $P(w)$ [35]. For a given dataset, D, variational parameter, θ, for each weight posterior of the BNN is updated by descending towards the gradients of the objective function, $\eta(D,\theta) = -E_{q(w|\theta)}[\log P(D|w)] + KL[q(w|\theta)||P(w)]$, where the data-dependent first term refers to the likelihood loss and prior-dependent second term refers to Kullback-Leibler (KL) divergence loss.

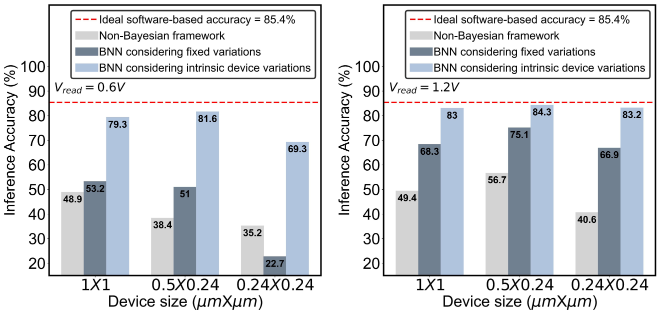

Fig. 5: Inference performance analysis of AlexNet on CIFAR10 dataset for non-Bayesian and Bayesian framework, wherein the Bayesian framework performance has been evaluated for fixed variations and device-level intrinsic variations profile at operating voltages of (a) $0.6V$, (b) $1.2V$ [20].

create autonomous self-repair mechanisms that, inspired by biology, dynamically counteracts the effects of hardware faults [21]. Crucially, the key functions of these complex biological processes can be translated into practical hardware implementations. Computational neuroscience modeling has shown that the core of astrocyte-mediated temporal dynamics and their normalization effect can be distilled into effective, local learning rules [21]. Such local rules, where updates depend only on directly related signals (e.g., local synaptic state, neuronal activity), are exceptionally well-suited for efficient implementation within the physical constraints of in-memory crosspoint architectures, avoiding the need for complex global communication or external calculation [21]. Recent work applying these astromorphic principles, specifically through Astrocyte-augmented Spike Timing Dependent Plasticity (A-STDP: a form of bio-inspired local learning), has demonstrated significantly better accuracy and repair convergence speed for unsupervised learning tasks [21]. In summary, astromorphic self-repair represents a compelling, future-driven outlook for achieving truly robust neuromorphic computing. While current demonstrations have primarily focused on mitigating faults like synaptic failures and potential drift effects [21], the underlying principles of bio-inspired normalization and local adaptation hold potential for broader applicability and can pave the way for next-generation AI hardware with autonomous self-repair capabilities.

V. Conclusions

FeFETs offer significant promise for efficient in-memory neuromorphic computing, yet realizing their potential is challenged due to inherent device-circuit variations that impact reliability, particularly at scaled dimensions. This review highlighted that overcoming these hurdles demands a co-design approach - integrating extensive device characterization with sophisticated algorithmic solutions ranging from variation-aware training strategies to efficient Bayesian methods incorporating device variation statistics. Furthermore, computational neuroscience insights offers a compelling future direction through astromorphic self-repair, leveraging bio-inspired normalization principles via local, hardware-compatible learning rules to potentially achieve dynamic resilience during operation. Ultimately, the successful deployment of robust FeFET-based neuromorphic systems hinges on synergistically combining these strategies across the computing stack from devices to circuits, systems and algorithms.

Acknowledgments

The authors would like to acknowledge GlobalFoundries, Dresden, Germany, for providing FeFET testing devices. This material is based upon work supported primarily by the National Science Foundation under Grant No. CNS 2137259 - Center for Advanced Electronics through Machine Learning (CAEML) and its industry members. The characterization of FeFET device non-idealities is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences Energy Frontier Research Centers program under Award No. DE-SC0021118 and the National Science Foundation under Grant No. 2347024.

References

[1] Zou, X., Xu, S., Chen, X., Yan, L. and Han, Y., 2021. Breaking the von Neumann bottleneck: architecture-level processing-in-memory technology. Science China Information Sciences, 64(6), p.160404.

[2] I. Chakraborty, A. Jaiswal, A. K. Saha, S. K. Gupta, and K. Roy, "Pathways to efficient neuromorphic computing with non-volatile memory technologies," *Appl. Phys. Lett.*, vol. 7, no. 2, pp. 021308(1–30), Jun. 2020, doi: 10.1063/1.5113536.

[3] W. Haensch, A. Raghunathan, K. Roy, B. Chakrabarti, C. M. Phatak, C. Wang, and S. Guha, "Compute in-Memory with Non-Volatile Elements for Neural Networks: A Review from a Co-Design Perspective," *Adv. Mat.*, vol. 35, no. 37, pp. 2204944(1–32), Sep. 2023, doi: 10.1002/adma.202204944.

[4] Upadhyay, N.K., Jiang, H., Wang, Z., Asapu, S., Xia, Q. and Joshua Yang, J., 2019. Emerging memory devices for neuromorphic computing. Advanced Materials Technologies, 4(4), p.1800589.

[5] Khan, A.I., Keshavarzi, A. and Datta, S., 2020. The future of ferroelectric field-effect transistor technology. Nature Electronics, 3(10), pp.588-597.

[6] Yu, S., Hur, J., Luo, Y.C., Shim, W., Choe, G. and Wang, P., 2021. Ferroelectric HfO2-based synaptic devices: recent trends and prospects. Semiconductor Science and Technology, 36(10), p.104001.

[7] Saha, A., Islam, A.N.M., Zhao, Z., Deng, S., Ni, K. and Sengupta, A., 2021. Intrinsic synaptic plasticity of ferroelectric field effect transistors for online learning. Applied Physics Letters, 119(13).

[8] Islam, A.N.M., Saha, A., Jiang, Z., Ni, K. and Sengupta, A., 2023. Hybrid stochastic synapses enabled by scaled ferroelectric field-effect transistors. Applied Physics Letters, 122(12).

[9] Jerry, M., Chen, P.Y., Zhang, J., Sharma, P., Ni, K., Yu, S. and Datta, S., 2017, December. Ferroelectric FET analog synapse for acceleration of deep neural network training. In 2017 IEEE international electron devices meeting (IEDM) (pp. 6-2). IEEE.

[10] Mulaosmanovic, H., Ocker, J., Müller, S., Noack, M., Müller, J., Polakowski, P., Mikolajick, T. and Slesazeck, S., 2017, June. Novel ferroelectric FET based synapse for neuromorphic systems. In 2017 Symposium on VLSI Technology (pp. T176-T177). IEEE.

[11] Sun, X., Wang, P., Ni, K., Datta, S. and Yu, S., 2018, December. Exploiting hybrid precision for training and inference: A 2T-1FeFET based analog synaptic weight cell. In 2018 IEEE international electron devices meeting (IEDM) (pp. 3-1). IEEE.

[12] Miller, N.E., Wang, Z., Dash, S., Khan, A.I. and Mukhopadhyay, S., 2021, June. Characterization of Drain Current Variations in FeFETs for PIM-based DNN Accelerators. In 2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS) (pp. 1-4). IEEE.

[13] Choe, G. and Yu, S., 2021. Variability study of ferroelectric field-effect transistors towards 7nm technology node. IEEE Journal of the Electron Devices Society, 9, pp.1131-1136.

[14] Wang, C., Victor, J., Saha, A.K., Chen, X., Si, M., Sharma, T., Roy, K., Ye, P.D. and Gupta, S.K., 2023, June. FeFET-based synaptic cross-bar arrays for deep neural networks: Impact of ferroelectric thickness on device-circuit non-idealities and system accuracy. In 2023 Device Research Conference (DRC) (pp. 1-2). IEEE.

[15] Y. Long, X. She, and S. Mukhopadhyay, “Design of reliable DNN accelerator with un-reliable ReRAM,” in Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE), 2019, pp. 1769–1774.

[16] Y. Bi, Q. Xu, H. Geng, S. Chen, and Y. Kang, “Resist: Robust network training for memristive crossbar-based neuromorphic computing systems,” IEEE Trans. Circuits Syst. II, Exp. Briefs, vol. 70, no. 6, pp. 2221–2225, Jan. 2023.

[17] D. Gao et al., “Bayesian inference based robust computing on memristor crossbar,” in Proc. 58th ACM/IEEE Design Autom. Conf. (DAC), Dec. 2021, pp. 121–126.

[18] D. Gao et al., “BRoCoM: A Bayesian framework for robust computing on memristor crossbar,” IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 42, no. 7, pp. 2136–2148, Oct. 2022.

[19] Saha, A., Manna, B., Lu, S., Jiang, Z., Ni, K. and Sengupta, A., 2024, April. Device Feasibility Analysis of Multi-level FeFETs for Neuromorphic Computing. In 2024 IEEE 6th International Conference on AI Circuits and Systems (AICAS) (pp. 327-331). IEEE.

[20] Manna, B., Saha, A., Jiang, Z., Ni, K. and Sengupta, A., 2024. Variation-Resilient FeFET-Based In-Memory Computing Leveraging Probabilistic Deep Learning. IEEE Transactions on Electron Devices.

[21] Han, Z., Islam, A.N. and Sengupta, A., 2023, June. Astromorphic self-repair of neuromorphic hardware systems. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 6, pp. 7821-7829).

[22] Trentzsch, M., Flachowsky, S., Richter, R., Paul, J., Reimer, B., Utess, D., Jansen, S., Mulaosmanovic, H., Müller, S., Slesazeck, S. and Ocker, J., 2016, December. A 28nm HKMG super low power embedded NVM technology based on ferroelectric FETs. In 2016 IEEE International Electron Devices Meeting (IEDM) (pp. 11-5). IEEE.

[23] C. Li, F. Müller, T. Ali, R. Olivo, M. Imani, S. Deng, C. Zhuo, T. Kämpfe, X. Yin, and K. Ni, “A scalable design of multi-bit ferroelectric content addressable memory for data-centric computing,” *IEEE International Electron Devices Meeting (IEDM)*, Dec 2020, pp. 29–3, doi: 10.1109/IEDM13553.2020.9372119.

[24] Islam, A.N.M., Saha, A., Jiang, Z., Ni, K. and Sengupta, A., 2023. Hybrid stochastic synapses enabled by scaled ferroelectric field-effect transistors. Applied Physics Letters, 122(12).

[25] Dutta, S., Schafer, C., Gomez, J., Ni, K., Joshi, S. and Datta, S., 2020. Supervised learning in all FeFET-based spiking neural network: Opportunities and challenges. Frontiers in neuroscience, 14, p.634.

[26] Ni, K., Chakraborty, W., Smith, J., Grisafe, B. and Datta, S., 2019, June. Fundamental understanding and control of device-to-device variation in deeply scaled ferroelectric FETs. In 2019 Symposium on VLSI Technology (pp. T40-T41). IEEE.

[27] Xia, L., Huangfu, W., Tang, T., Yin, X., Chakrabarty, K., Xie, Y., Wang, Y. and Yang, H., 2017. Stuck-at fault tolerance in RRAM computing systems. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 8(1), pp.102-115.

[28] Peng, X., Huang, S., Luo, Y., Sun, X. and Yu, S., 2019, December. DNN+ NeuroSim: An end-to-end benchmarking framework for compute-in-memory accelerators with versatile device technologies. In 2019 IEEE international electron devices meeting (IEDM) (pp. 32-5). IEEE.

[29] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), pp.2278-2324.

[30] Graves, A., 2011. Practical variational inference for neural networks. Advances in neural information processing systems, 24.

[31] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: A review for statisticians,” J. Amer. Stat. Assoc., vol. 112, no. 518,pp. 859–877, 2017.

[32] Wade, J., McDaid, L.J., Harkin, J., Crunelli, V. and Kelso, S., 2012. Self-repair in a bidirectionally coupled astrocyte-neuron (AN) system based on retrograde signaling. Frontiers in computational neuroscience, 6, p.26578.

[33] Liu, J., Harkin, J., Maguire, L.P., McDaid, L.J. and Wade, J.J., 2017. SPANNER: A self-repairing spiking neural network hardware architecture. IEEE transactions on neural networks and learning systems, 29(4), pp.1287-1300.

[34] Rastogi, M., Lu, S., Islam, N. and Sengupta, A., 2021. On the self-repair role of astrocytes in STDP enabled unsupervised SNNs. Frontiers in Neuroscience, 14, p.603796.

[35] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, “Weight uncertainty in neural network,” *in Proc. 32nd International Conference on Machine Learning*, Jun 2015, pp. 1613–1622.