In-memory Subnet Computation for Area and Energy Efficient AI

Shi Zhao[a], **Yue Li**[a], **Jieming Pan**[a], **Evgeny Zamburg**[a, b], **Aaron Voon-Yew Thean**[a, b]

[a] *Department of Electrical and Computer Engineering, National University of Singapore, 117608 Singapore*
zhao_shi@u.nus.edu, e1374399@u.nus.edu, jm.pan@nus.edu.sg

[b] *Singapore Hybrid-Integrated Next-Generation μ-Electronics (SHINE) Centre, National University of Singapore*
zamburg@nus.edu.sg, aaron.thean@nus.edu.sg

1. Introduction

Continual learning on edge demands energy-efficient hardware capable of dynamic adaptation through few-shot data acquisition, presenting fundamental challenges to conventional von Neumann architectures. This work addresses this imperative by proposing an in-memory computing architecture utilizing our customised dual-gate back-end-of-line (BEOL) Ferroelectric Field-Effect Transistors (FeFETs). Integrated with state-of-the-art subnet systems, our computation architecture combines energy-efficient MAC operation with concurrent subnet masking for rapid and resource-efficient continual learning. A single 400×100 FeFET crossbar array supports simultaneous MAC and masking operation, achieving at least 2.42× improvement in energy efficiency. Leveraging this architecture, we achieve 90.6% average accuracy over 10 incremental tasks using the Permuted-MNIST (PMNIST) benchmark dataset, comparable to GPU baselines at 91.9%. With model size scaling, our architecture consistently demonstrates at least 1.39× area efficiency and at least 2.13× faster computation speed compared to state-of-the-art approaches. These results underscore the critical role of unconventional computing architectures in overcoming the energy, latency, and scalability barriers, and thereby advancing hardware-realizable continual learning for edge intelligence.

2. Methodology

2.1 Dual-gate Ferroelectric Field-Effect Transistor

Here we introduce a dual-gate Field-Effect Transistor (DG-FeFET) with Ferroelectric gate insulator (HZO) at the top side of channel material and normal high-k (HfO) at the bottom side. The fabrication process flow is shown in **Fig.1**. Our DG-FeFET features top of the class memory window of 2.1V, high on/off ratio of 10^6 without hysteresis, an on current of 10 μA/μm at drain voltage of 0.1V, and long retention of 10^4s as shown in **Fig.2(a)**. Most importantly, the top and bottom gate can work independently through the same channel to support MAC and masking operation simultaneously. **Fig.2(b)** demonstrates highly responsive channel conductance with gate voltage control, providing effective mask control.

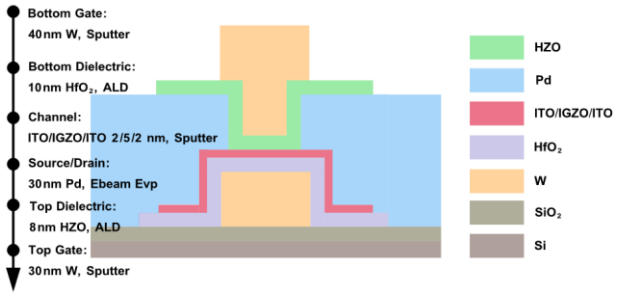

Fig.1: Structure of Dual-gate FeFET device [1].

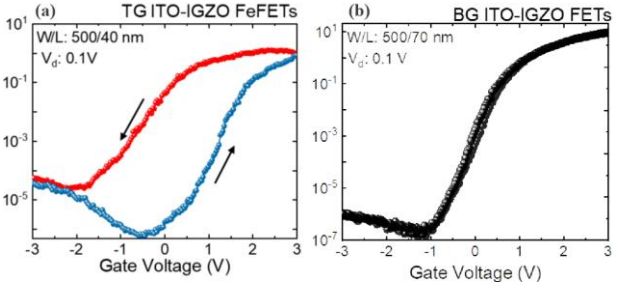

Fig.2: I_d-V_g of TG-FeFET and BG-FET.

2.2 DG-FeFETs Crossbar Array Architecture

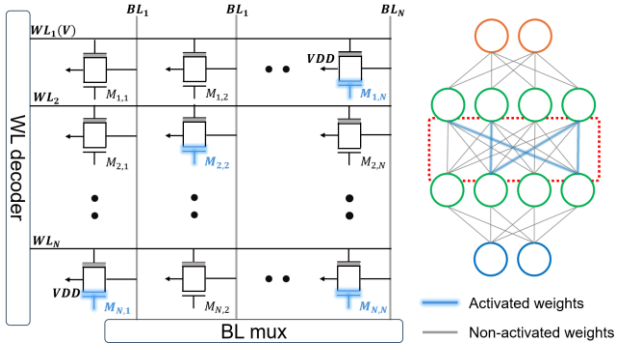

Fig.3: FeFET crossbar array architecture.

As shown in **Fig.3**, each computing element in the crossbar array comprises a single DG-FeFET. The top gate performs compute-in-memory (CIM) like [2], [3] whereas the bottom gate serves as the binary subnetwork mask M that disables computation for weights not used during inference following $y = M \odot W \times X$ [4]. The overall workflow is illustrated in **Fig.4**.

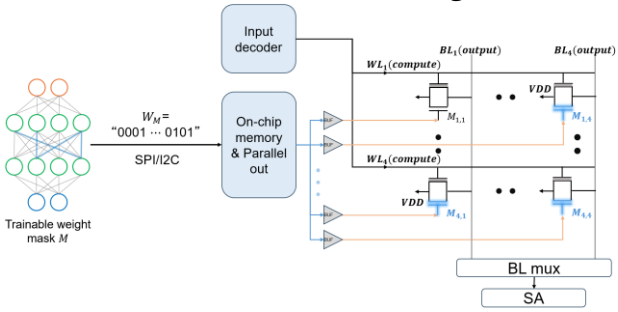

Fig.4: Flowchart of FeFET crossbar array.

In-memory Subnet Computation for Area and Energy Efficient AI

Shi Zhao[a], **Yue Li**[a], **Jieming Pan**[a], **Evgeny Zamburg**[a, b], **Aaron Voon-Yew Thean**[a, b]

[a] *Department of Electrical and Computer Engineering, National University of Singapore, 117608 Singapore*
zhao_shi@u.nus.edu, e1374399@u.nus.edu, jm.pan@nus.edu.sg

[b] *Singapore Hybrid-Integrated Next-Generation μ-Electronics (SHINE) Centre, National University of Singapore*
zamburg@nus.edu.sg, aaron.thean@nus.edu.sg

2.3 Experimental Setup

Experiments were conducted on a consistent two-layered subnetwork-based deep neural network (DNN) with a 400-100-10 architecture using the PMNIST dataset [5]. In our continual learning model, each permuted dataset is learned incrementally with a dedicated subnetwork mask before the inference process. A total of 5 epochs and 20% subnetwork mask sparsity are used throughout the evaluation.

3. Results and discussion

3.1 Graphic Processing Unit (GPU) vs DG-FeFET

By encoding input signals as analog voltages and representing weights as the conductance of FeFETs, our CIM approach overcomes the Von-Neumann bottleneck and effectively increases information density compared to the digital counterparts in conventional GPU platforms. As shown in **Table 1**, this leads to more than 30× improvement in TOPS/W.

Table 1: Benchmark between GPU and our work (*: estimated value based on energy efficiency).

Metric	GPU [6]	This work
Average accuracy	90.6%	91.9%
Computation time (s)	3.2×10^{-5}*	2.47×10^{-4}
Energy (J)	1.57×10^{-4}*	3.75×10^{-7}
TOPS/W	10	518

Leveraging our DG-FeFET memory window, high on-off ratio, and long retention, our array achieved an average accuracy of 90.6% on the PMNIST dataset (derived based on **Fig.5**), closely matching the 91.9% accuracy obtained with a conventional GPU-based implementation. We achieve an excellent 518 TOPs/W (**Appendix A**), ~30× of conventional GPU implementation, through the integration of weight masking at the bottom gate. This approach eliminates the need for additional switching transistors and separate dot product computations.

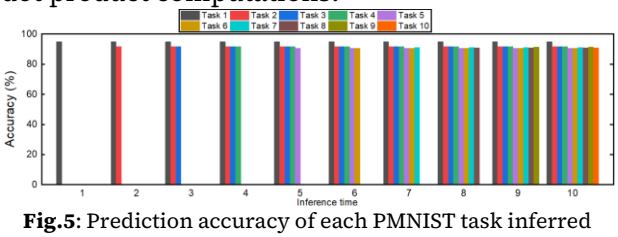

Fig.5: Prediction accuracy of each PMNIST task inferred once a new task is introduced.

3.2 Evaluation Against State-Of-The-Art CIM

Compared to other state-of-the-art FeFET-based CIM architecture with an extra switch transistor for mask control, our DG-FeFET approach integrates mask control and computation in a single transistor. **Table 2** highlights the characteristics of our DG-FeFET and 1T1FeFET.

Table 2: Device parameters for DG-FeFET and TG-FeFET.

Devices	Max Conductance	Min Conductance	Size (W/L)	Off current
DG FeFET	1×10^{-12}	5×10^{-6}	500/70 nm	1×10^{-13}
TG FeFET	4.5×10^{-11}	4.5×10^{-5}	500/40nm	2×10^{-12}

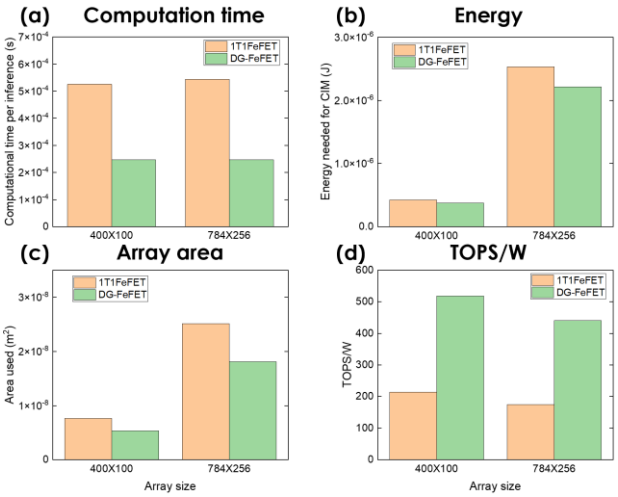

Fig.6: Comparison of (a) computation speed, (b) energy consumption, (c) array area, and (d) energy efficiency between 1T1FeFET and DG-FeFET [7].

Despite the smaller device size of the additional switching transistor, our DG-FeFET design reduces the overall footprint by ~30% (**Fig.6c**). By reducing the switching time in subnetwork computation, DG-FeFETs also reduce the computation time by at least 2.13× (**Fig.6a**) and improve the overall energy efficiency by ~2.5× (**Fig.6d**).

4. Conclusion

Through device-algorithm co-design, novel algorithms call for unconventional computing to maximize performance and efficiency. Our demonstration highlights the potential for ~30× and ~2.5× improvement in computation and energy efficiency to conventional GPU-based and 1T1FeFET implementations respectively. These results underscore the critical role of unconventional computing architectures in overcoming the energy, latency, and scalability barriers, and thereby advancing hardware-realizable continual learning for edge intelligence.

In-memory Subnet Computation for Area and Energy Efficient AI

Shi Zhao[a], **Yue Li**[a], **Jieming Pan**[a], **Evgeny Zamburg**[a, b], **Aaron Voon-Yew Thean**[a, b]

[a] *Department of Electrical and Computer Engineering, National University of Singapore, 117608 Singapore*
zhao_shi@u.nus.edu, e1374399@u.nus.edu, jm.pan@nus.edu.sg

[b] *Singapore Hybrid-Integrated Next-Generation μ-Electronics (SHINE) Centre, National University of Singapore*
zamburg@nus.edu.sg, aaron.thean@nus.edu.sg

Acknowledgments

This work is supported in part by Agency for Science, Technology and Research (A*STAR), Singapore, under its AME Programmatic Funds (A1892B0026 & A18A1B0045), the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Science Program, as well as NRF's Medium Sized Centre: Singapore Hybrid-Integrated Next-Generation μ-Electronics (SHINE) Centre funding program.

References

[1] C. CHUN-KUEI, "HIGH-PERFORMANCE OXIDE-BASED FERROELECTRIC FETS AND FETS FOR NON-VOLATILE LOGIC APPLICATIONS," 2024.

[2] T. Soliman *et al.*, "First demonstration of in-memory computing crossbar using multi-level Cell FeFET," *Nat. Commun.*, vol. 14, no. 1, p. 6348, Oct. 2023, doi: 10.1038/s41467-023-42110-y.

[3] E. Yu, G. K. K, U. Saxena, and K. Roy, "Ferroelectric capacitors and field-effect transistors as in-memory computing elements for machine learning workloads," *Sci. Rep.*, vol. 14, no. 1, p. 9426, Apr. 2024, doi: 10.1038/s41598-024-59298-8.

[4] Y. Cheng *et al.*, "Photonic neuromorphic architecture for tens-of-task lifelong learning," *Light Sci. Appl.*, vol. 13, no. 1, p. 56, Feb. 2024, doi: 10.1038/s41377-024-01395-4.

[5] H. Kang *et al.*, "Forget-free Continual Learning with Winning Subnetworks".

[6] "NVIDIA H100 Tensor Core GPU," NVIDIA. Accessed: Feb. 27, 2025. [Online]. Available: https://www.nvidia.com/en-sg/data-center/h100/

[7] Y. Luo, X. Peng, and S. Yu, "MLP+NeuroSimV3.0: Improving On-chip Learning Performance with Device to Algorithm Optimizations," in *Proceedings of the International Conference on Neuromorphic Systems*, in ICONS '19. New York, NY, USA: Association for Computing Machinery, Jul. 2019, pp. 1–7. doi: 10.1145/3354265.3354266.

Appendix A. Energy Efficiency Calculation

To calculate TOPS/W for 1T1FeFET and DG-FeFET solutions, total number of operations (TOP) are calculated as:

$$TOP = Row \times Col + sparsity \times Row \times Col$$

, where *Row* and *Col* refers to the row and column size of CIM array and *sparsity* is the percentage of subnetwork weight used for inference of each task (20% in this case). The former term calculates the total number of dot products, and the latter term finds the total number of MAC for sparse matrix multiplication.

Appendix B. Computation Time and Energy Estimation

To estimate the computation time for subnetwork-based matrix multiplication, the total number of computations is estimated following **Appendix A,** and the operating frequency of GPU is chosen as 1.5GHz as a typical value. Then, the energy is estimated by dividing TOPS/W from TOPS estimated.

Appendix C. Simulation Platform For 1T1FeFET and DG-FeFET CIM Array

In this work, NuroSim is used as a standard open-source benchmark platform. Characteristics of single-gate and double-gate FeFET fabricated are updated in NeuroSim to ensure a consistent comparison. The additional switch transistor for 1T1FeFET approach is modelled using standard cell in 32 nm technology node. Computation speed takes into account of gate capacitance switching time of all transistors used in CIM array.

Appendix D. Hyperparameter For Continual Learning Training

Dataset	PMNIST	
Network Configuration	400-100-10	784-256-10
Batch Size	256	
Weight and Mask Initialization	Kaiming Uniform method	
Activation Function	ReLU	
Number Of Epochs	5	
Training Framework	Pytorch with SGD	