DESIGN OF FERROELECTRIC FET-BASED SYMMETRIC COMPUTING-IN-MEMORY ARRAY WITH SIMULTANEOUS WEIGHT GRADIENT CALCULATION AND WEIGHT UPDATE FOR ON-CHIP LEARNING

Yuxin Lin[1], Zerui Chen[1], Jin Luo[2], Qianqian Huang[2,3*] and Ru Huang[2,3*]*

[1]School of Software & Microelectronics, Peking University, Beijing 102600, China
[2]School of Integrated Circuits, Peking University, Beijing 100871, China
[3]Beijing Advanced Innovation Center for Integrated Circuits, Beijing 100871, China
[*]Corresponding Author's Email: luoj@pku.edu.cn, hqq@pku.edu.cn, ruhuang@pku.edu.cn

Poster number: 1-76

In this work, a novel ferroelectric FET (FeFET) based **symmetric computing-in-memory (CIM) array** enabling **simultaneous weight gradient (ΔW) calculation and weight update** is proposed for on-chip learning with reduced latency. By utilizing **stochastic computing (SC) scheme with novel bipolar pulse streams** for ΔW calculation and utilizing the output pulses of SC for update of FeFET weights, the ΔW calculation and weight update of backpropagation (BP) learning are in-situ realized simultaneously for the whole array of CIM cell with special designed tri-state XNOR gate. Moreover, through the design of symmetric differential input and readout based on reference FeFET, four-quadrant vector-matrix multiplication (VMM) of **configurable forward and backward propagation** are implemented in the FeFET-based CIM array. Based on the design, image recognition's neural network is demonstrated with a high accuracy of 95.33% and reduced hardware cost, showing great potential for on-chip learning.

Introduction

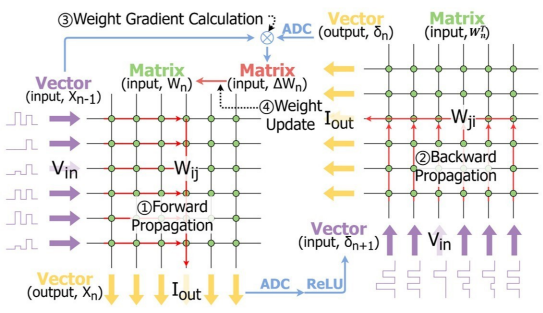
- **Non-volatile memory (NVM) based CIM (nvCIM)** for on-chip learning: high area-efficiency and high energy-efficiency
- **On-chip learning of DNN using backpropagation algorithm** based on CIM architecture:

(1) Forward and backward propagation:

$$x_n = W_n x_{n-1}$$
$$\delta_{n-1} = W_n^T \delta_n$$

(2) Weight calculation and weight update:

$$\Delta W_N = x_{n-1}\delta_n^T$$
$$W_n \leftarrow W_n + \Delta W_n$$



- **FeFET-based nvCIM:** non-volatile multilevel weight storage, low energy consumption and high CMOS compatibility
- **Current approaches** for CIM-based on-chip learning need extra peripheral circuits for weight gradient calculation and row-wise writing of weight update, **suffering from high hardware cost and latency**

FeFET-based Symmetric CIM

Architecture of FeFET-based CIM for on-chip learning

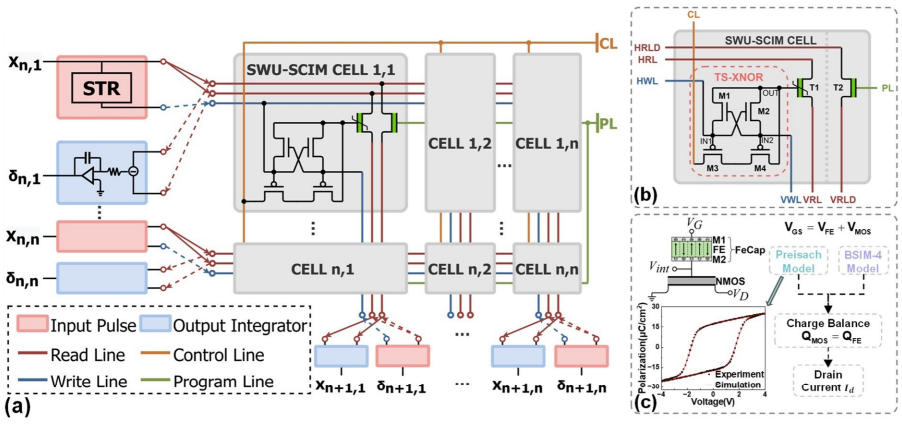
Our proposed FeFET-based CIM architecture: a symmetric CIM array with simultaneous weight calculation and update, configurable peripheral circuits managing multiple modes of on-chip learning

➢ ***Stochastic computing (SC) based weight update (SWU-SCIM) cell***

- T1 (variable-conductance FeFET) and T2 (fixed-conductance FeFET) constitute a **differential pair for signed weight storage**
- M1~M4 constitute a **tri-state output XNOR (TS-XNOR) for bipolar SC** based weight gradient calculation and weight update

➢ ***Configurable peripheral circuits for multiple modes***

- Stochastic translators (STRs): convert inputs into random bitstreams for SC based weight calculation and update mode
- Alternating subtractors and integrators arranged in the row and column: configure forward or backward propagation mode



Modelling framework of FeFET

The FeFET device model is developed through the charge-balance principle between BSIM-4 MOSFET model and multi-domain Preisach ferroelectric model

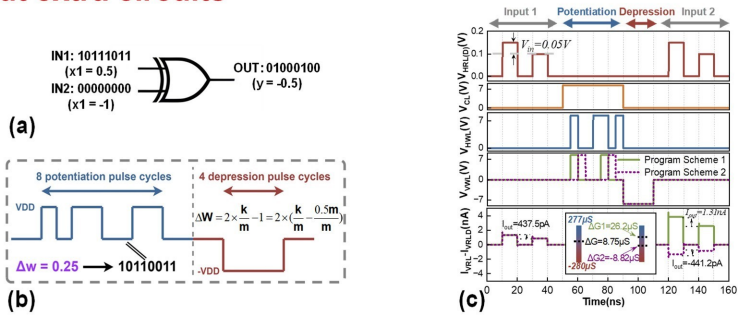
Weight Gradient Calculation and Update

➢ **Stochastic computing based weight gradient calculation**

- SC scheme uses pulse-modulated probabilistic bitstreams to **simplify arithmetic operations with basic logic gates**
- XNOR logic is used for multiplication of bipolar data

➢ **In-situ simultaneous weight calculation and weight update**

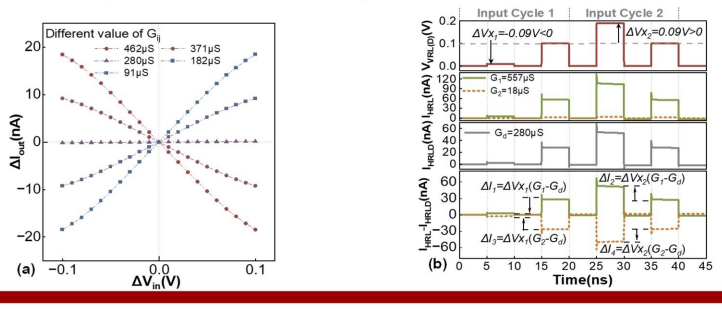
- **Bipolar pulses program**: in SWU-SCIMs, pulse-width modulated outputs of TS-XNOR are directly applied to program FeFET, **realizing simultaneous weight update for the whole array without extra circuits**



Forward and Backward Propagation

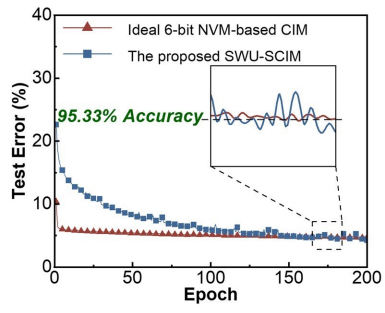
➢ **Four-quadrant VMM for forward and backward propagation**

- ***Signed weight:*** differential weight device pair (T1 & T2) are programmed with W_{ij} and $W_{max}/2$, and outputs are read differentially
- ***Signed input:*** differential voltage V_x and $V_{max}/2$ applied over two cycles, and outputs from the two cycles are subtracted



Evaluation

- Proposed FeFET-based symmetric CIM architecture with simultaneous weight calculation and update enables **high accuracy for on-chip learning**
- It supports both forward and backward propagation while enabling simultaneous weight calculation and weight update for the whole array with **reduced hardware cost**



	Nature [1]	IEDM [2]	IEDM [3]	JXCDC [4]	This work
Weight Type	1T-1R	21T-1C	1T-1FeFET	5FeFET	**4T-2FeFET**
Weight Precision	Analog	9	7	1	**6**
Forward Propagation	√	√	√	√	√
Backward Propagation	×	√	×	√	√
Error Calculation	√	×	×	√	√
Weight Update	Only Last Layer	√	√	√	√
Synchronous Update	Only Last Layer	√	×	×	√
Learning Accuracy	96.20%	92.70%	94.10%	-	**95.33%**

[1] P. Yao et al., Nature 577, 2020. [2] Y. Kohda et al., IEEE IEDM, 2020. [3] K. A. Aabrar et al., IEEE IEDM, 2021. [4] W. Shim, et al., IEEE JXCDC, 7(1), 2021.

Summary: This work reports a novel **FeFET-based symmetric CIM array** for **on-chip learning**, enabling **simultaneous weight gradient calculation and weight update** through **bipolar pulses based stochastic computing** design. Moreover, **configurable forward and backward propagations** with a four-quadrant range are achieved through a symmetrical array design with differential input and readout. Based on the design, on-chip learning of DNN for image recognition is demonstrated with high accuracy, highlighting its strong potential for edge AI.

Acknowledgements: This work was supported by NSFC (62404008, 61927901, 62374009), 111 Project (B18001), and China Postdoctoral Science Foundation (2024M750098, GZC20240026)