POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Monolithic 4K ECRAM-Based Analog AI Chip for Energy-Efficient On-Chip Tranining

Hyunjeong Kwak*

AI & Neuromorphic Device Lab

Department of Materials Science and Engineering, POSTECH

**Email: hjkwak9524@postech.ac.kr*

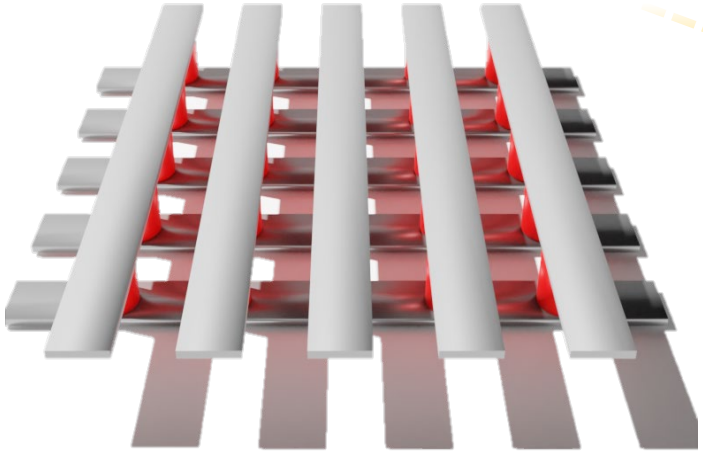
Major Contributions

I. **The first fully-integrated ECRAM-based analog AI accelerator chip**, achieving energy efficiency (6.17 TOPS/W) and low power consumption (11.08 mW)
II. A BEOL-compatible fabrication process
III. **The largest selector-free in-situ neural network training demonstration**, leveraging superior device characteristics

Ⅰ. Introduction

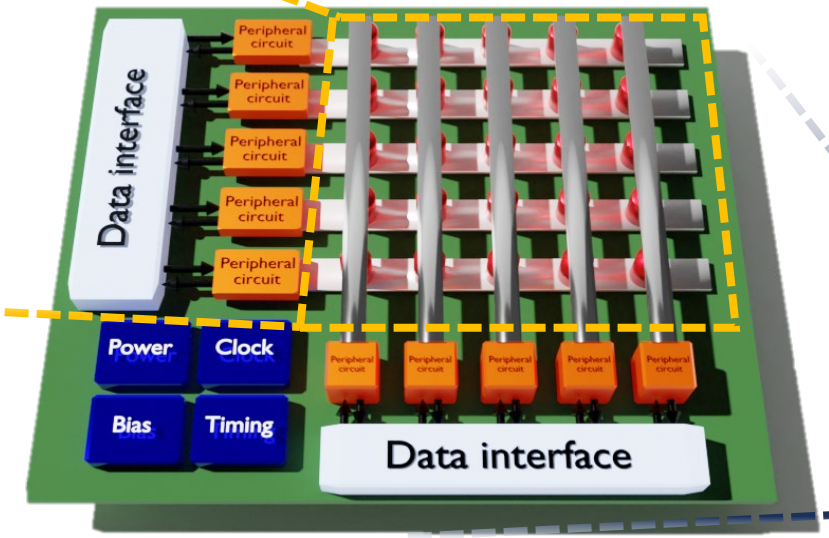
❖System Hierarchy for Analog Accelerator Chip

Resistive Cross-point array

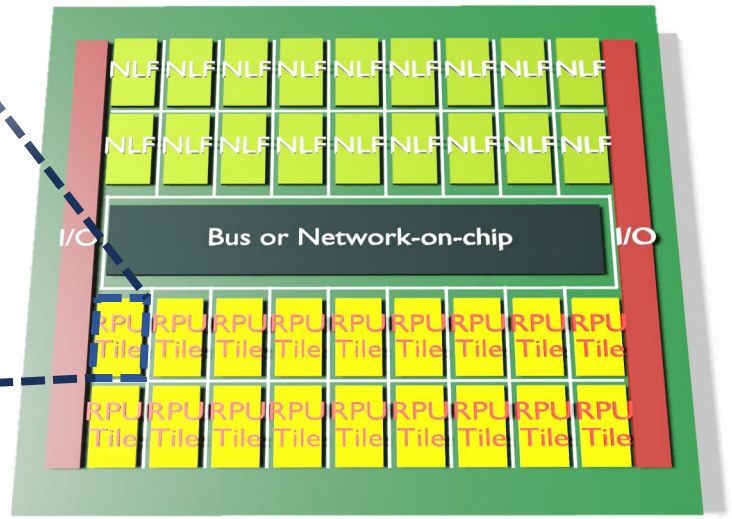


$$I_j = \sum V_i \, g_{ij}$$

RPU Tile w/ Peripheral Circuits



Analog AI Accelerator



*RPU = Resistive processing unit, NLF = Non-linear function, NoC = Network-on-chip

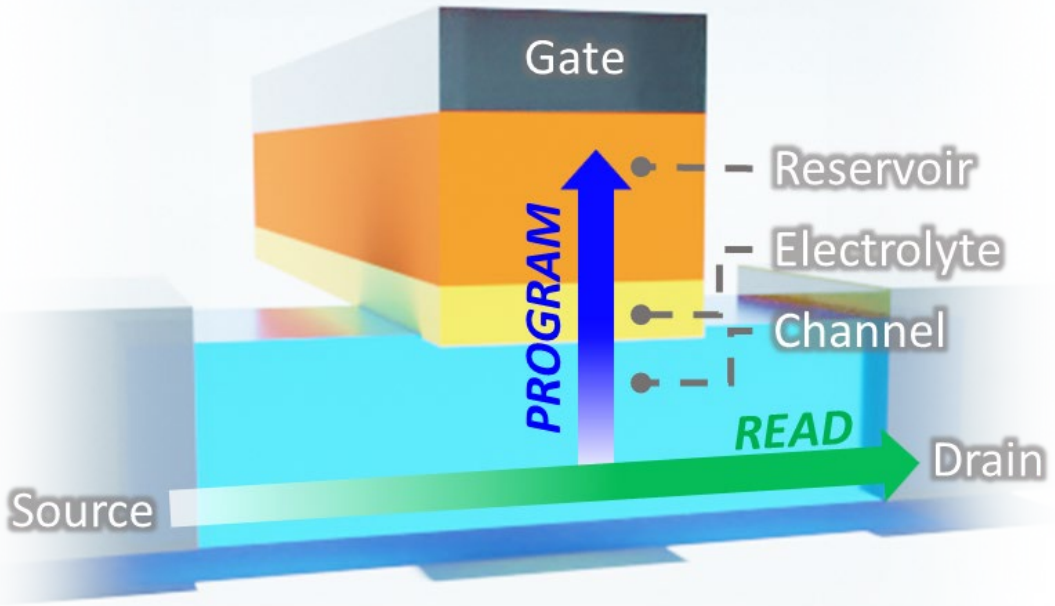
To realize analog acceleration at system level, resistive cross-point array are organized into hierarchical structures from device, to array, to full AI chip

[1] Gokmen, T. et al., Frontiers in Neuroscience 2016 [2] Kim, S., et al., MWSCAS 2017

Ⅱ. Electrochemical RAM

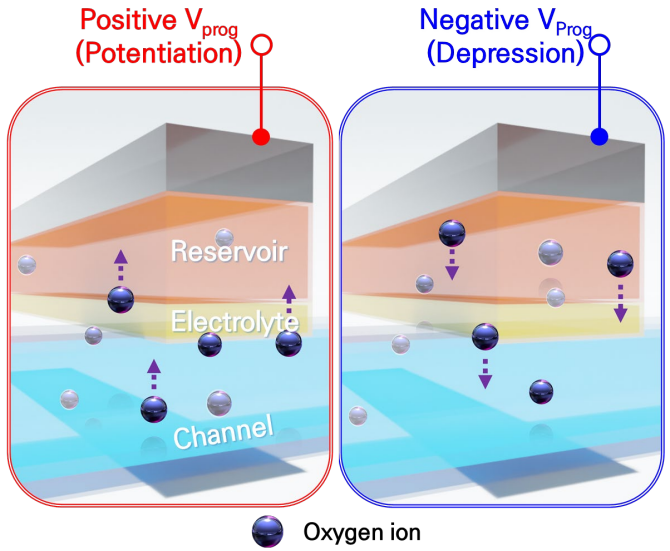
❖Electrochemical Random-Access Memory for High-Performance Accelerators

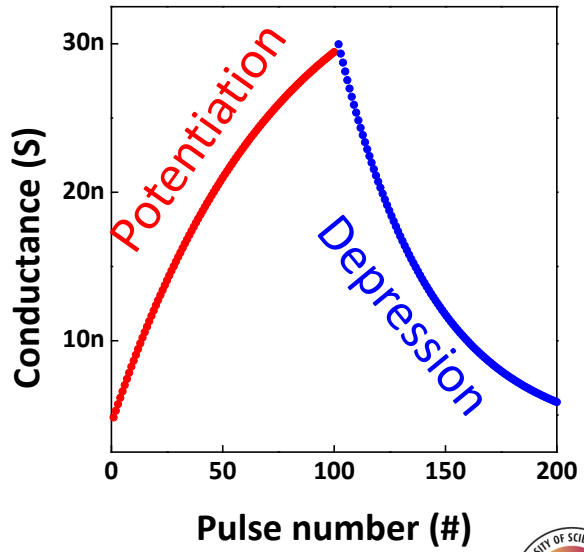
ECRAM structure & operation



K. Noh and H. Kwak et al., Sci. Adv. (2024)

- **Separation of read and write operation**
  - **Program (Write)** : Control the conductivity of the channel by injecting/removing ions into the channel by applying a current/electric field to the gate
  - **Read** : Conductivity is measured by applying a voltage between drain and source and reading the current



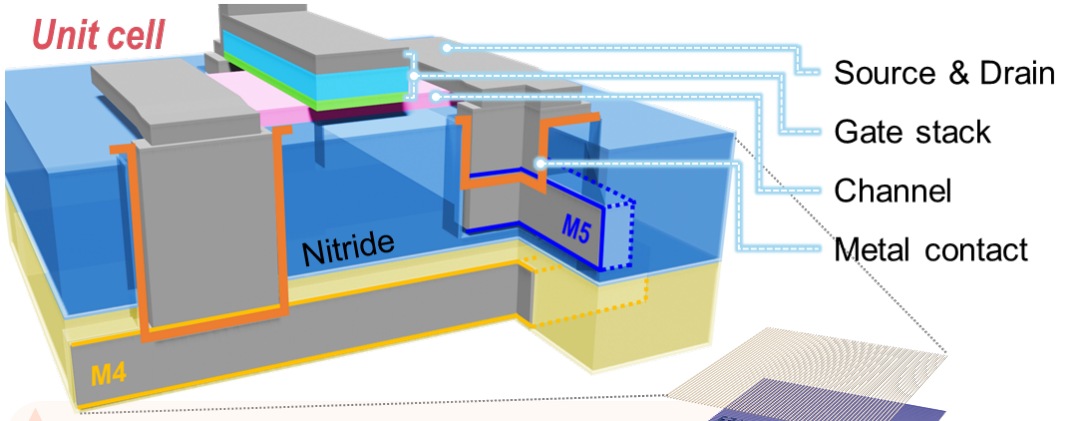


Ⅲ-1. Monolithic 4K ECRAM-Based Analog AI Chip

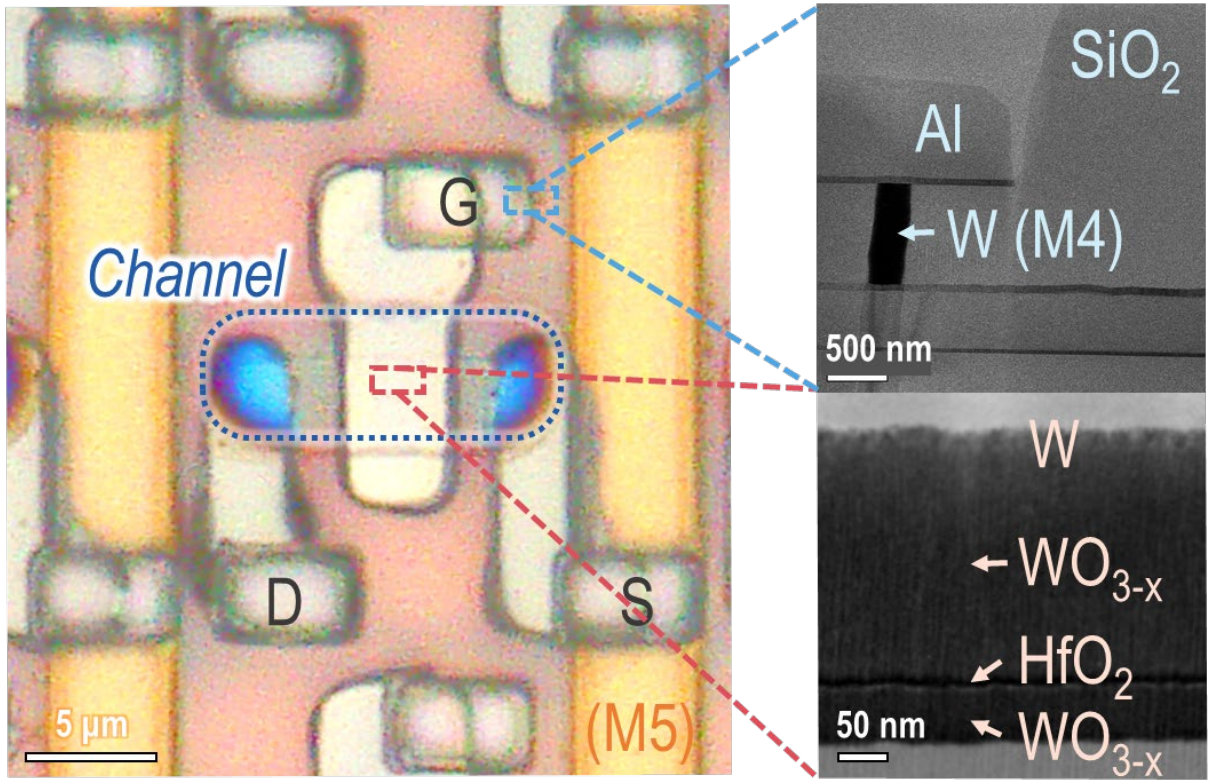
❖Fabrication of Monolithic 4K ECRAM Analog AI Chip

H. Kwak et al., (Manuscript in preparation)

Device, chip, and BEOL-compatible fabrication flow integrating ECRAMs on the 250nm-tech peripheral circuit chip



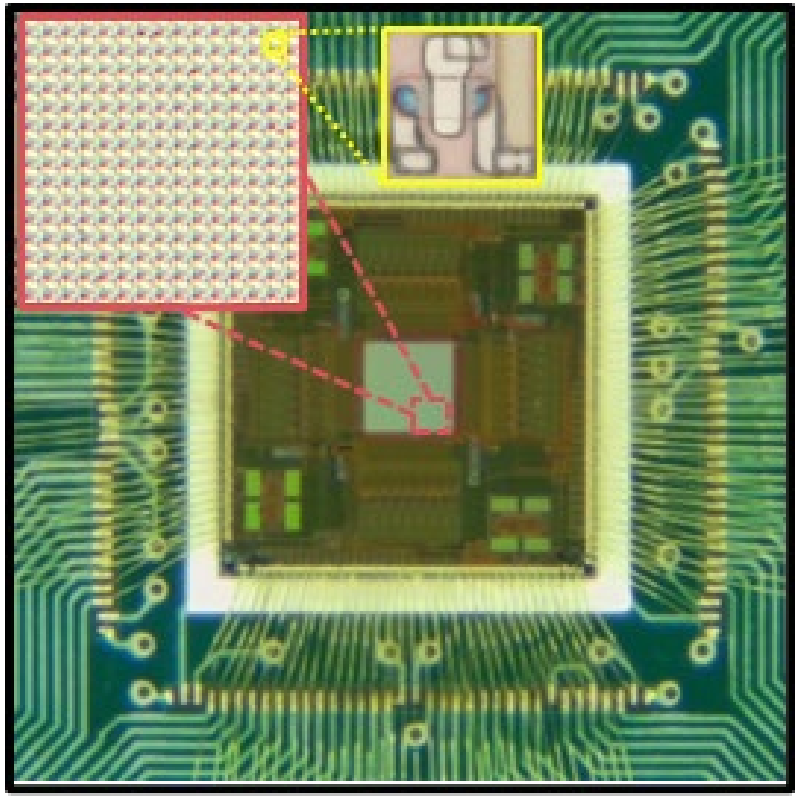
- Metal deposition to connect ECRAM devices and metal lines
- Gate stacks (WO_{3-x}/W) deposition
- Electrolyte (HfO_2) deposition & Etching of the connected area
- Channel deposition (WO_{3-x})
- Via filling to connect metal lines and ECRAM arrays
- 250-nm CMOS peripheral circuits

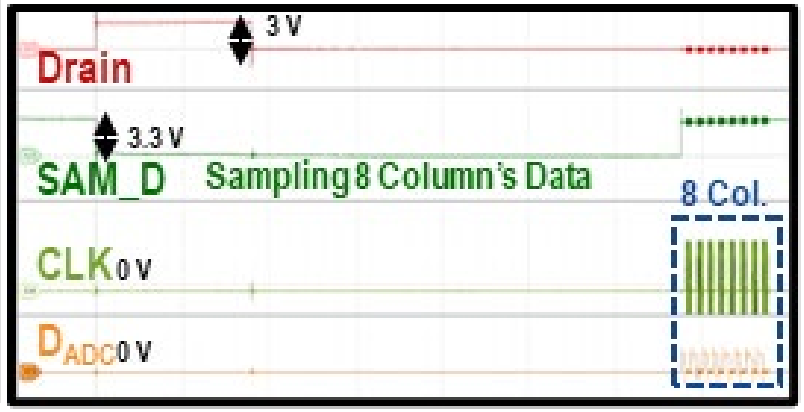


Ⅲ-1. Monolithic 4K ECRAM-Based Analog AI Chip

❖Monolithic 4K ECRAM Analog AI Chip

H. Kwak et al., (Manuscript in preparation)





- ECRAM array size: 64x64
- Chip size: 4.65 x 4.65 mm^2
- Unit cell area: 20x20 um^2
- Peripheral with SAR ADC 10 bits, 16 channels analog front-end, FPGA signal buffer

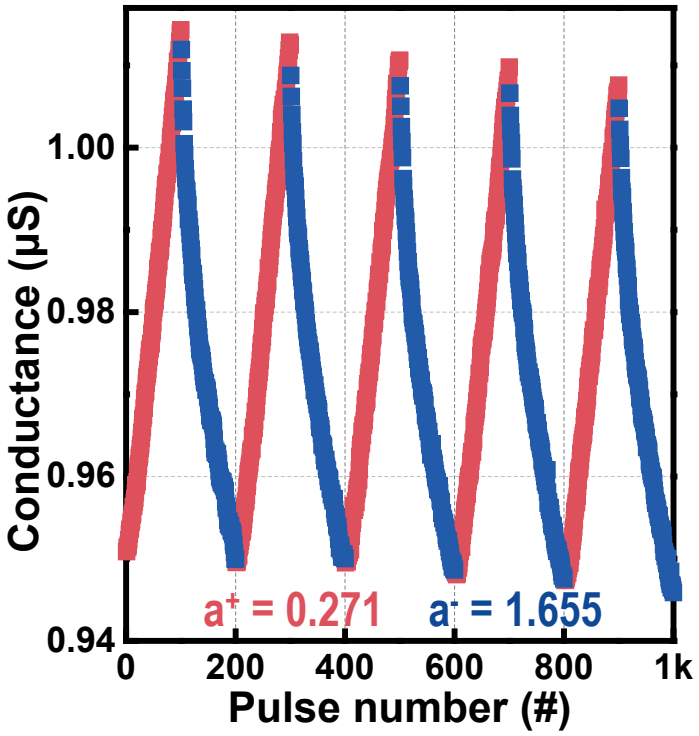
System-level demonstration by integrating ECRAM devices on foundry-built accelerator chip

Ⅲ-1. Monolithic 4K ECRAM-Based Analog AI Chip

❖Switching Performance of ECRAM Devices

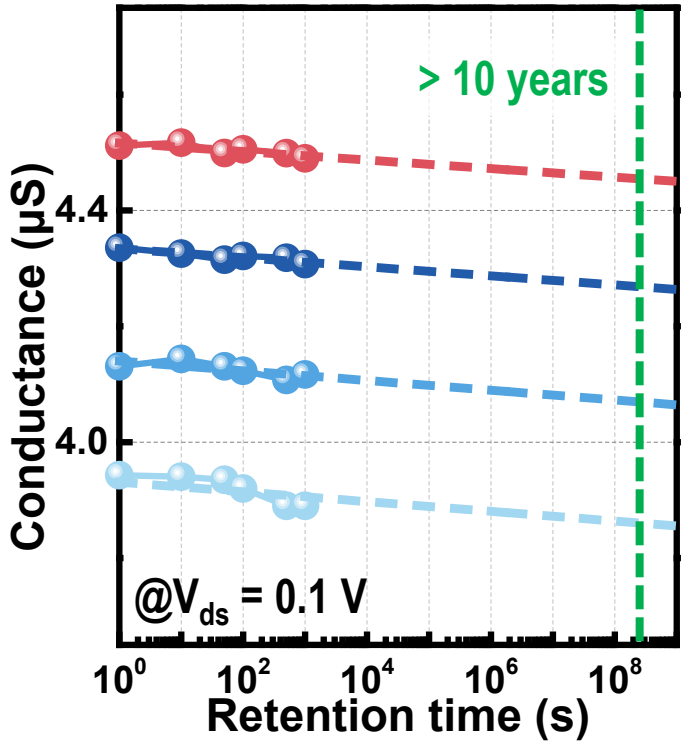
H. Kwak et al., (Manuscript in preparation)

Update Behavior



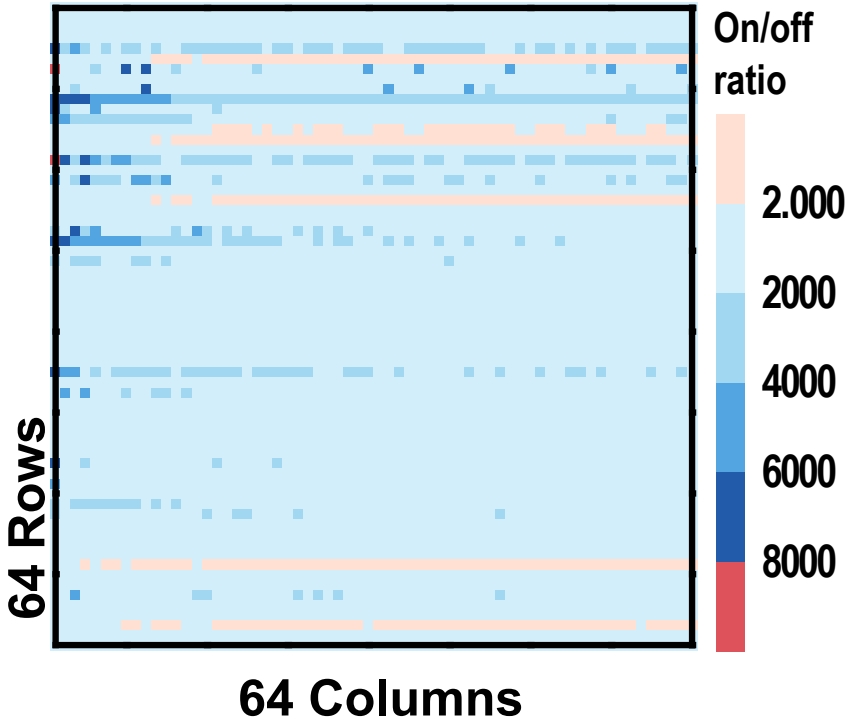
- 100 up (+3.0 V) /100 down (-2.15 V) pulses of 2.5 ms width

Retention



- Retention of various programmed states with extrapolated results of >10 years

On/off Ratio Heatmap



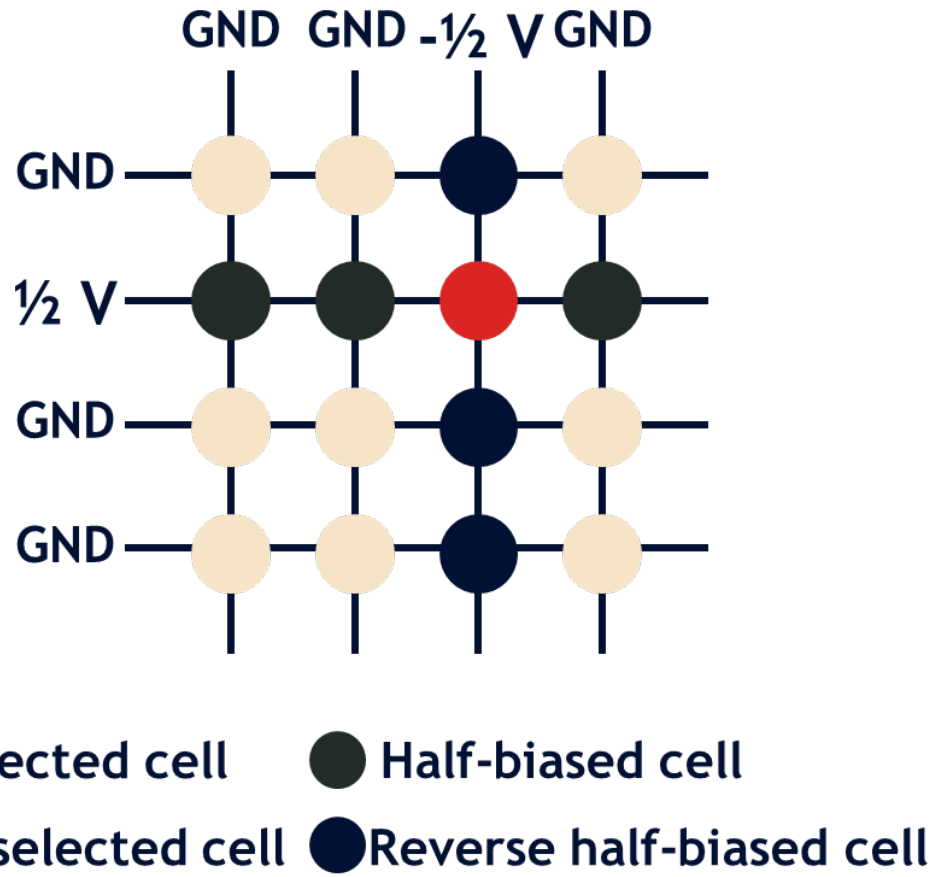
- Yield of switching devices is 84.4 % out of 4,096

Ⅲ-2. *in-situ* Training Demonstration

❖Hardware in-situ Training Demonstration in a 21×21 Array

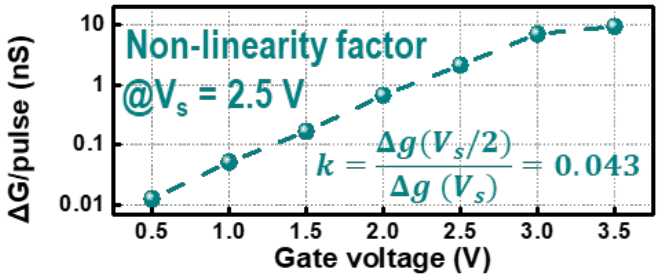
H. Kwak et al., (Manuscript in preparation)

Half-Bias Scheme

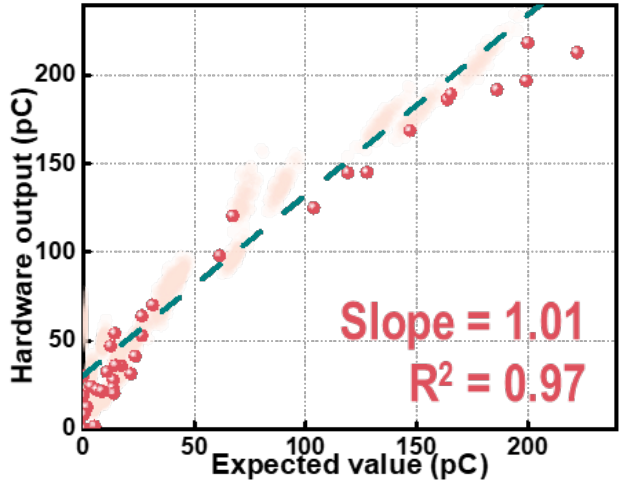


Pivotal Aspects for Array Operation

Half-bias selectivity



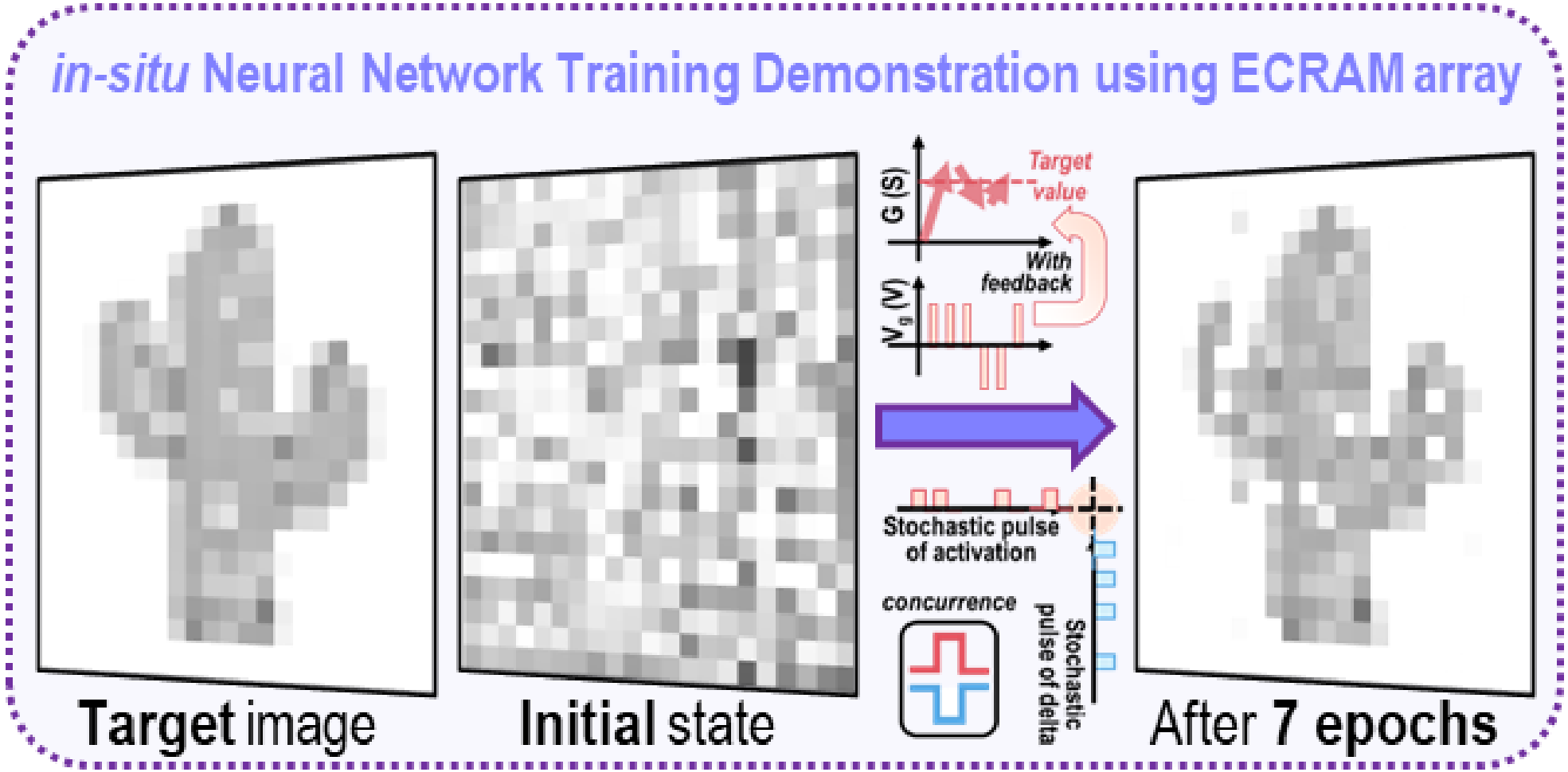
Vector-matrix multiplication



Ⅲ-2. *in-situ* Training Demonstration

❖Hardware in-situ Training Demonstration in a 21×21 Array

H. Kwak et al., (Manuscript in preparation)

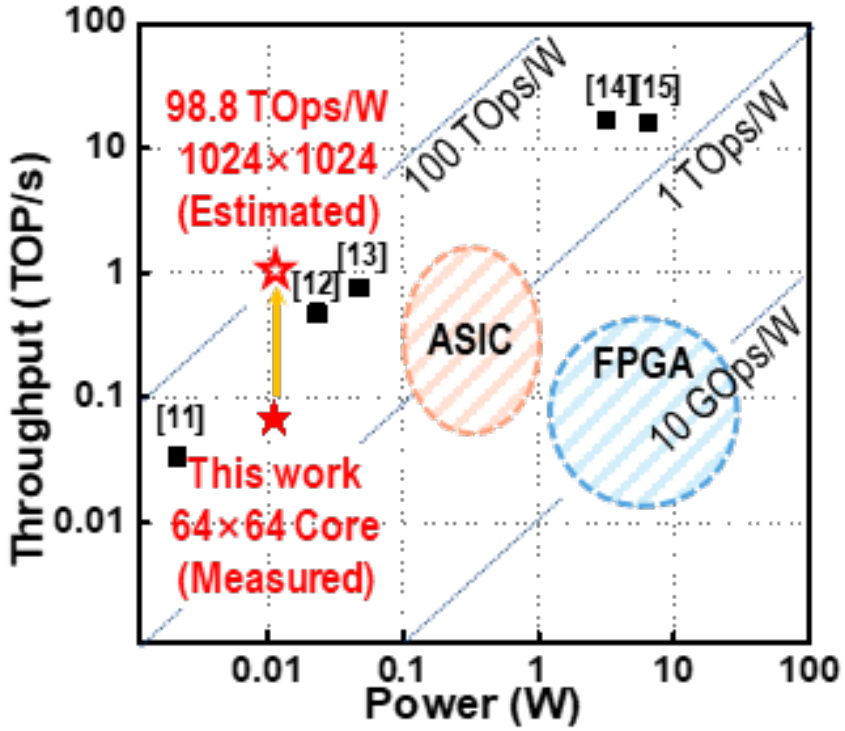


The largest-scale hardware demonstration of in-situ learning, involving 441 selector-free devices, which is a state-of-the-art result

Ⅲ-2. *in-situ* Training Demonstration

❖Expected Energy Efficiency of Our ECRAM Analog AI Chip

H. Kwak et al., (Manuscript in preparation)



	Nature '22 [13]	ISSCC '22 [14]	Nat. Elec. '23 [15]	This work
NVM	RRAM	nor-Flash	PCM	**ECRAM**
Chip area (mm^2)	159	190	144	**30.25**
CMOS Technology	130 nm	40 nm	14 nm	**250 nm**
Output precision	6b	8b	8b	**10b**
TOps/W	16	5.2	2.48	**6.17**
Core size	256×256	1024×1024	256×256	**64×64**
TOps/Core (/K)	0.25	0.005	0.039	**1.543**
Power (mW)	47.125	3190	6490	**11.08**

[11] J. Hung, Nat. Electron. (2021). [12] W. Khwa, ISSCC (2022). [13] W. Wan, ISSCC (2022). [14] L. Fick, ISSCC (2022). [15] M. Gallo, Nat. Electron. (2023).

Acknowledgement

AI & Neuromorphic Device Lab

Dr. Oki Gunawan & Chaeyoun Kim

- **Prof. Hyunsang Hwang** at POSTECH
- **Prof. Donghwa Lee** at POSTECH
- **Prof. Jiyong Woo** at KNU
- **Prof. Byungha Shin** at KAIST
- **Dr. Oki Gunawan** at IBM
- **Prof. Hyung-Min Lee** at Korea Univ.

Thank you for your attention