IEDM

A > 64 Multiple States and > 210 TOPS/W High Efficient Computing by Monolithic Si/CAAC-IGZO + Super-Lattice Ferroelectric for Ultra-Low Power Edge AI Application

M.-C. Chen[1], S. Ohshita[2,*], S. Amano[2], Y. Kurokawa[2], S. Watanabe[2], Y. Imoto[2], Y. Ando[2], W.-H. Hsieh[1], C.-H. Chang[1], C.-C. Wu[1], S.-S. Chuang[1], H. Yoshida[1], M.-C. Lu[1], M.-H. Liao[3,*], S.-Z. Chang[1,*], and S. Yamazaki[2]

PSMC 力晶積成電子 SEL SEMICONDUCTOR ENERGY LABORATORY

[1] **Powerchip Semiconductor Manufacturing Corporation, Hsinchu, Taiwan.**

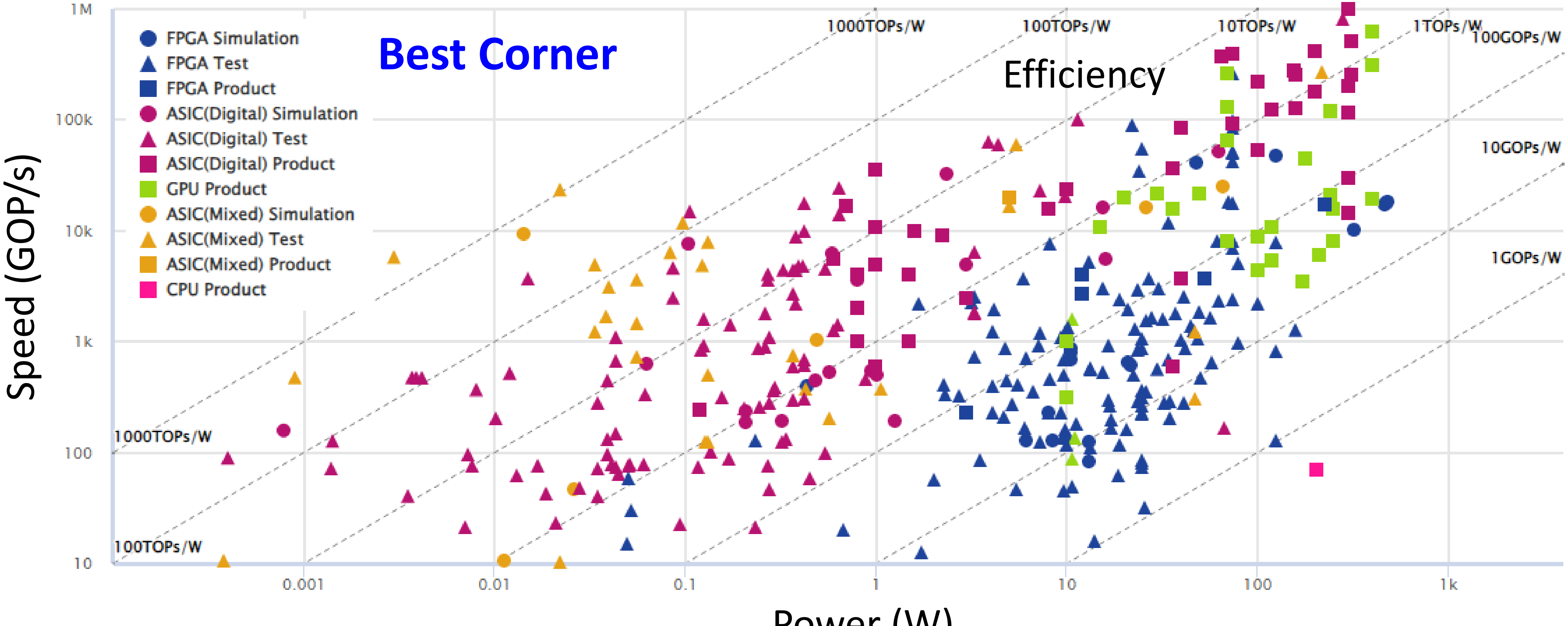
[2] **Semiconductor Energy Laboratory Co., Ltd., Kanagawa, Japan.**

[3] **National Taiwan University, Taipei, Taiwan. E-mail: mhliaoa@ntu.edu.tw**

Outline

- Introduction
- Experiments and Fabrications
- AiMC Chip Design and Operation
- Performance, Stability, and Reliability
- Benchmark & Conclusion

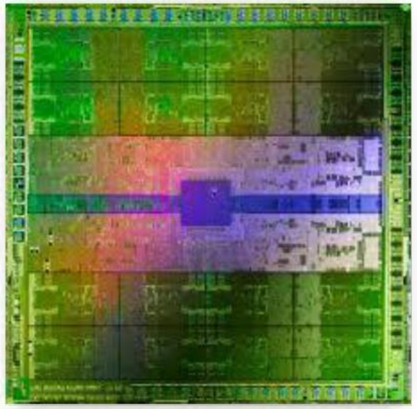
Accelerators for AI: Neural Network Accelerator



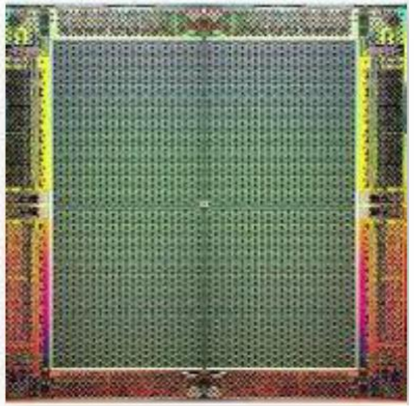
https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/

- High efficient AI accelerators (FPGA, ASIC, CPU, and GPU) are developed for different computing applications.

Accelerators for AI: GPU/FPGA/TPU/Compute-in-memory



GPU

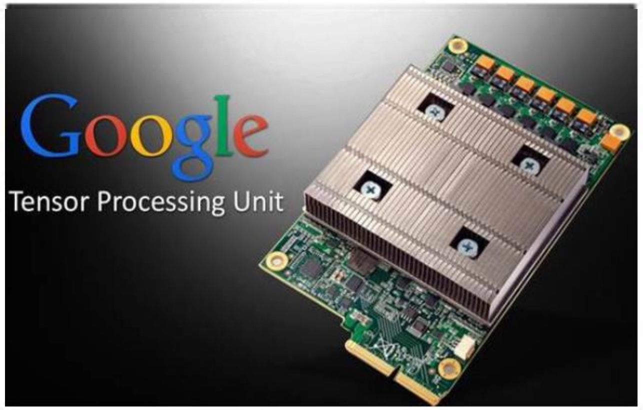


FPGA

Conventional computing platforms

~ 0.1 TOPS/W

Floating-point

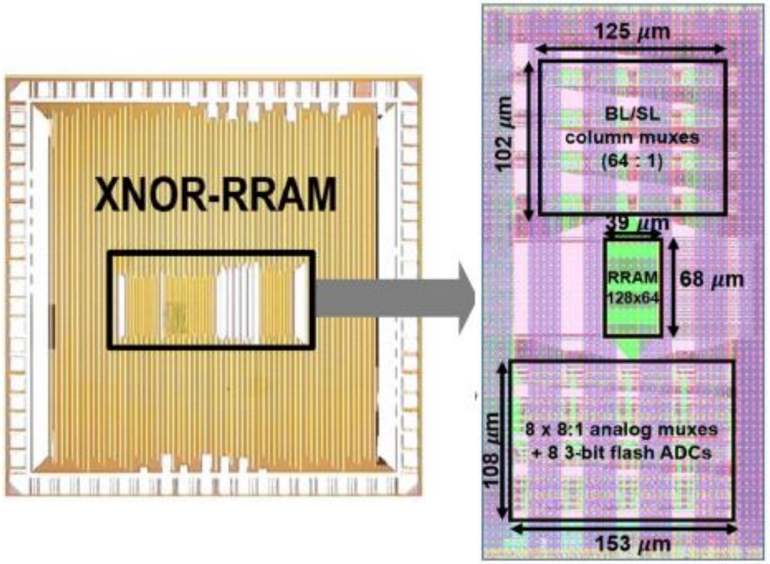


TPU

Digital CMOS ASICs

~ 1-10 TOPS/W

Fixed-point


XNOR-RRAM
125 μm
102 μm
BL/SL column muxes (64 : 1)
39 μm
RRAM 128x64
68 μm
108 μm
8 x 8:1 analog muxes + 8 3-bit flash ADCs
153 μm

Compute-in-memory (CIM)

Analog CMOS (or eNVMs)

~ 10-100 TOPS/W

Low-precision→ accuracy?

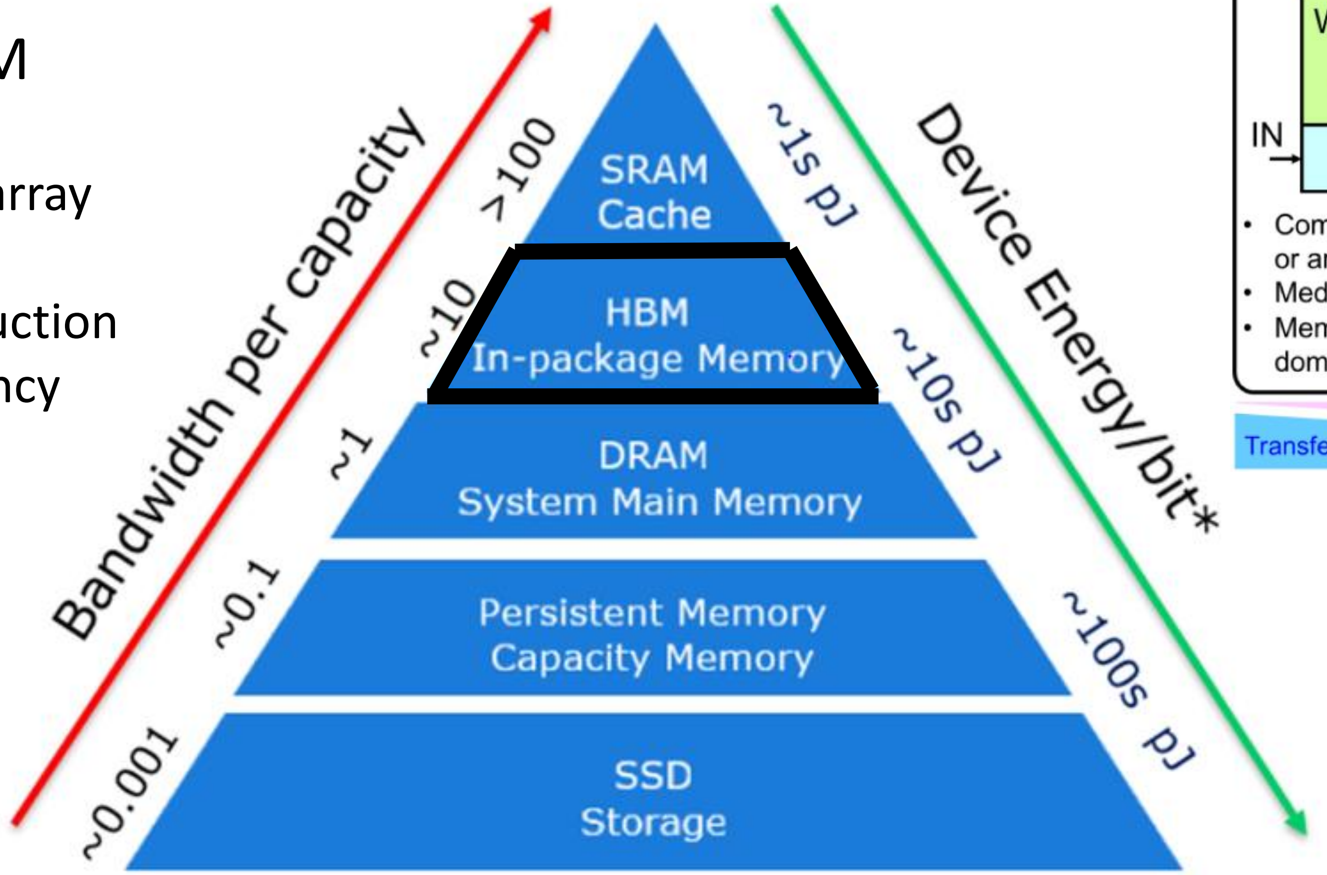
IEDM 2020 Short Course 2: Analog Memory Needs for AI (GIT)

- Compute-in-memory (CIM) chip is promising especially in the edge inference when model has pre-trained.
- It also supports incremental learning with new data input when deployed to the field.

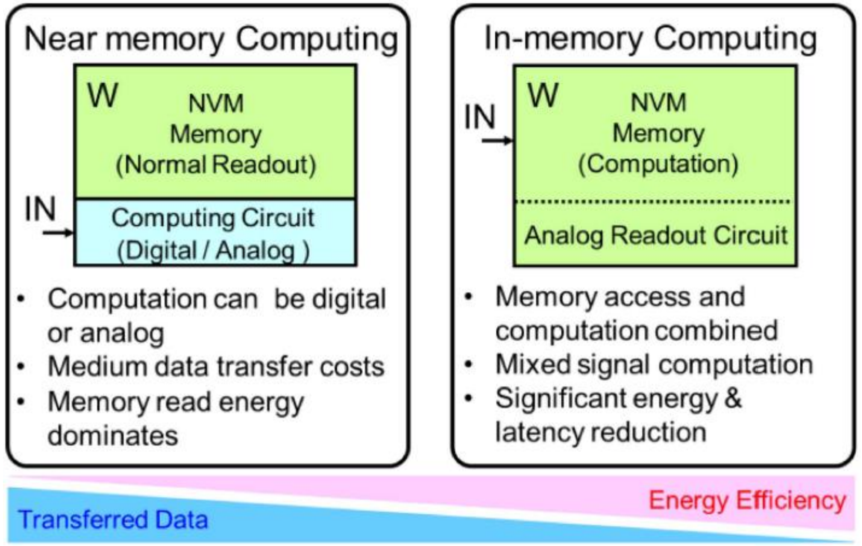
Memory Classification & High Band-Width Memory

Features of HBM

- 3D stack
- CMOS under array
- Faster
- Footprint reduction
- Energy efficiency



Intel & Shen Meng et al., Journal of Lightwave, 2019.



Near memory Computing:

- Computation can be digital or analog
- Medium data transfer costs
- Memory read energy dominates

In-memory Computing:

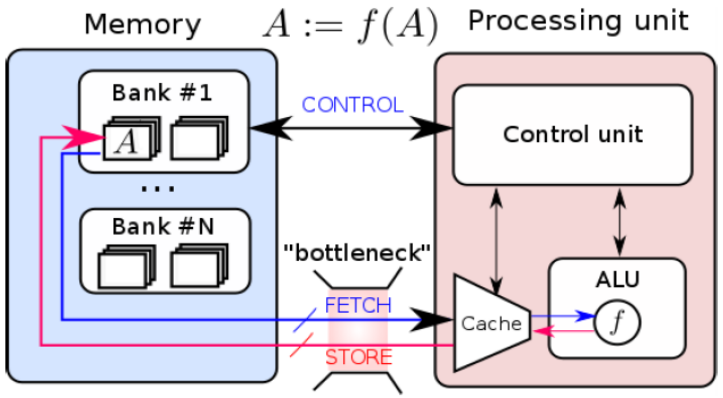
- Memory access and computation combined
- Mixed signal computation
- Significant energy & latency reduction

J.M. Hung et al., *Soild-State Circuit Society*, Vol. 1, pp. 171-183, 2021.

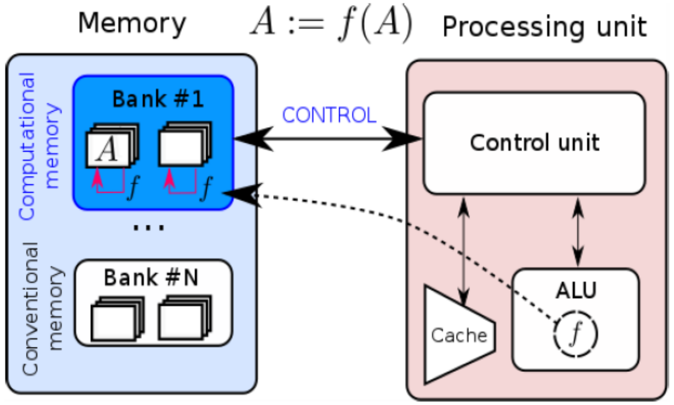
- Compute-in-memory (CIM) chip can also be used as L1-L3 memories, considering bandwidth and energy/bit.

Near Memory and In-Memory Computing

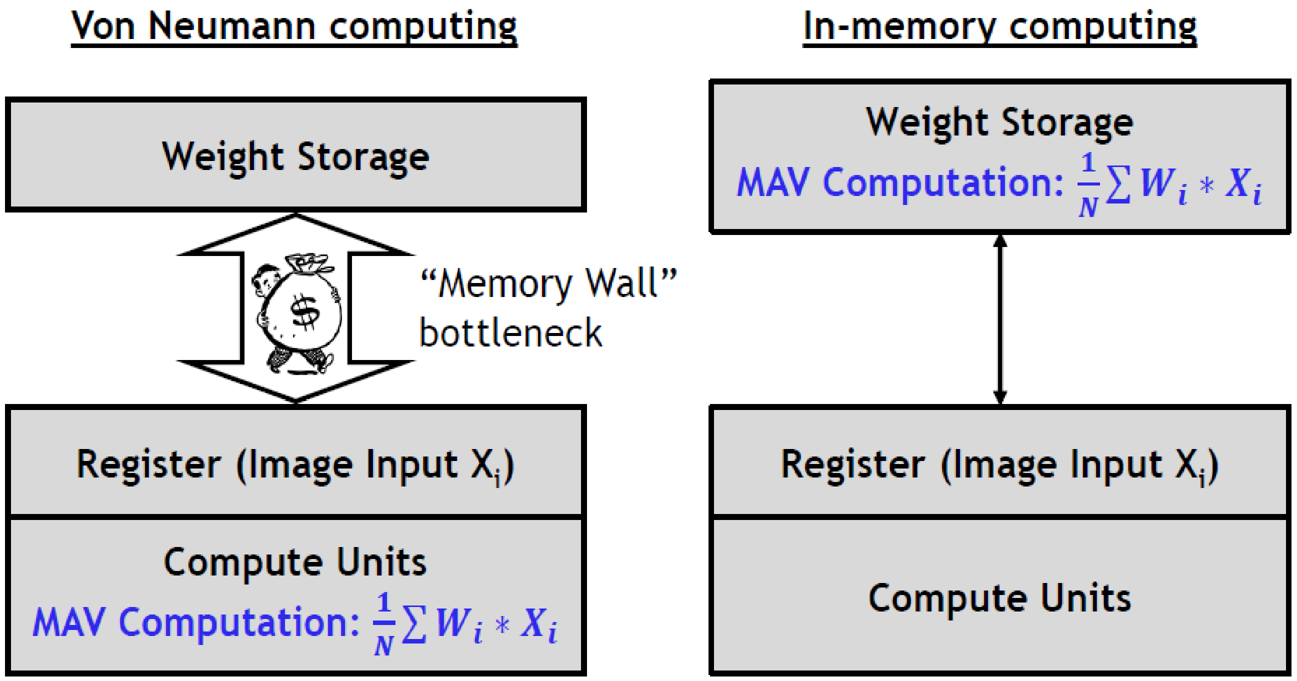
Processing unit & Conventional memory



Processing unit & Computational memory



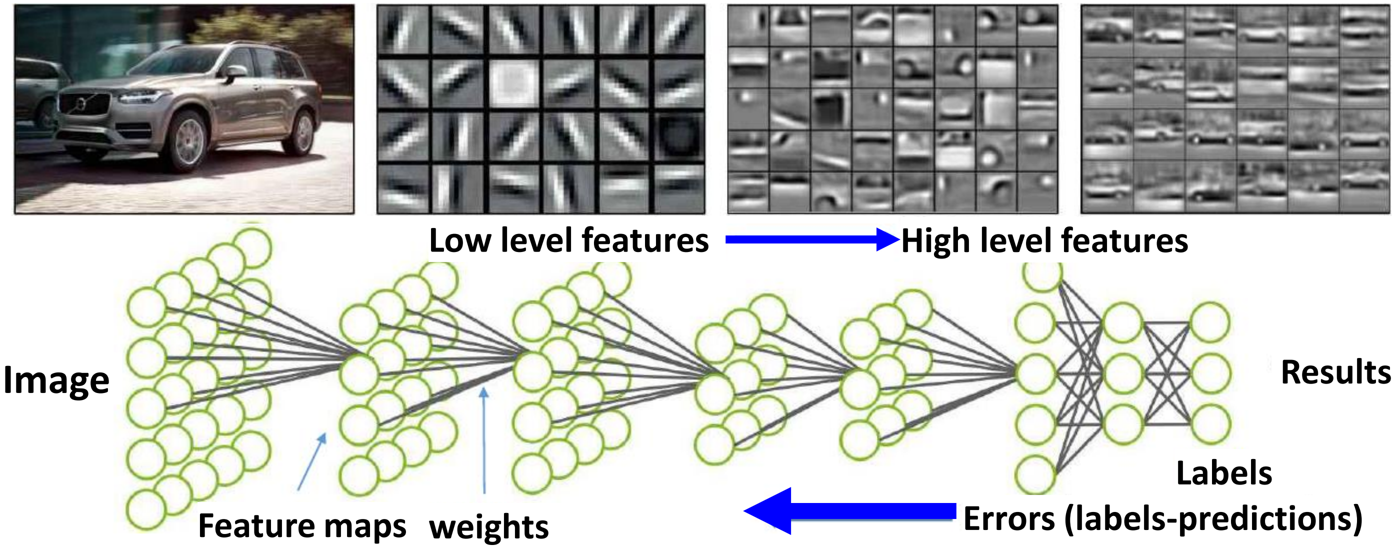
IEDM 2019 Tutorial 2: In Memory Computing for AI (IBM)


Von Neumann computing
Weight Storage
"Memory Wall" bottleneck
$
Register (Image Input X_i)
Compute Units
MAV Computation: $\frac{1}{N}\sum W_i * X_i$
In-memory computing
Weight Storage
MAV Computation: $\frac{1}{N}\sum W_i * X_i$
Register (Image Input X_i)
Compute Units

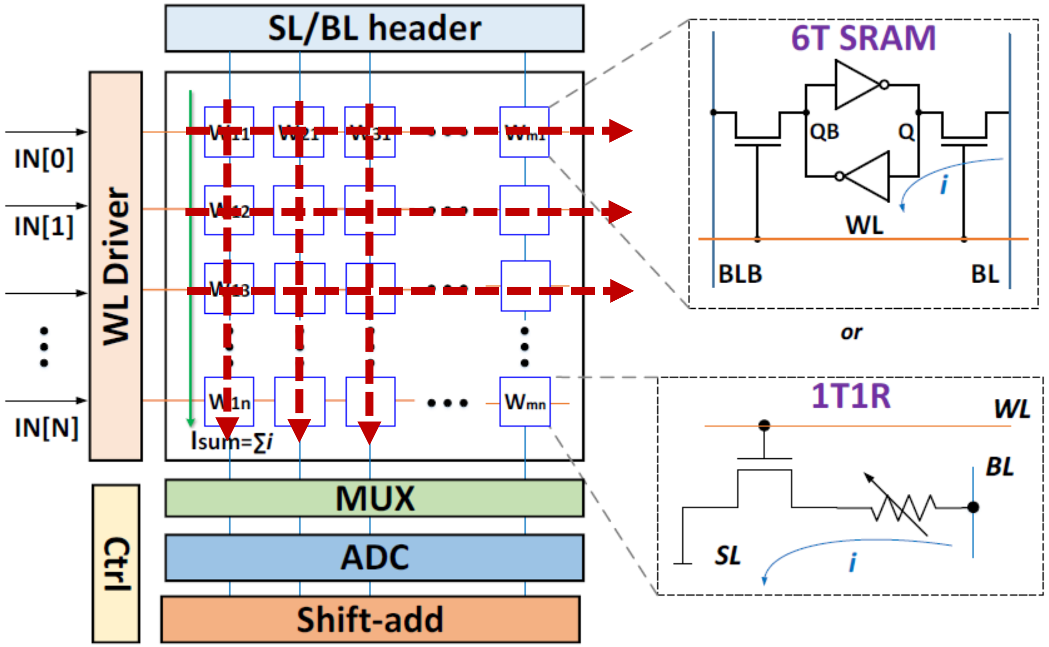
VLSI 2022 C12-4 (University of Texas)

- Compute-in-memory (CIM) chip reduces data movement between memory and processing units.
- It performs the computation within a memory array and release the “memory wall”.

AI Computing and Analog Computing Concept



IEDM 2020 Short Course 2: Analog Memory Needs for AI (GIT)



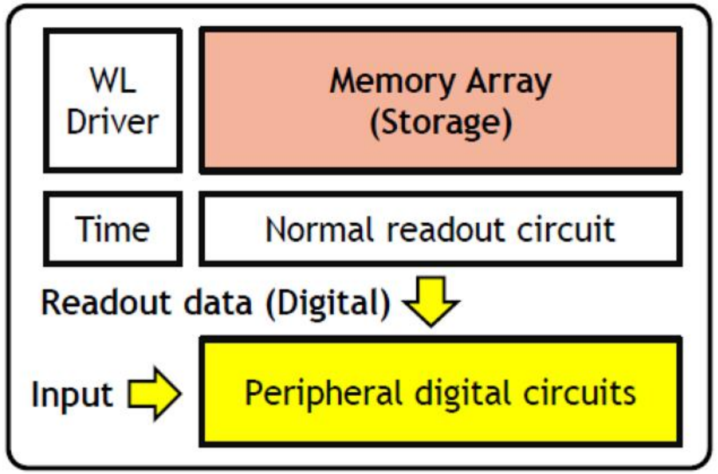
IEDM 2020 Short Course 2: Analog Memory Needs for AI (GIT)

- AI Computing has two basic functions:
  - ✓ Training: write intensive to synaptic weight memories.
  - ✓ Inference: read intensive to synaptic weight.
- Analog Computing: Intensive computation needs 『Vector-matrix-multiplication』 by 『analog/multi-states』 for parallel computing.

Digital *versus* Analog Computing in Memory

Digital Near-Memory

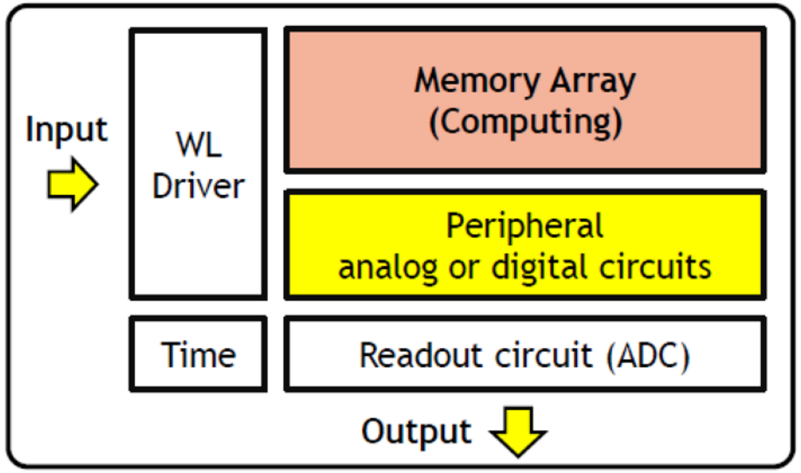
- Digital signal computation
- Multiply and accumulation in peripheral digital circuit
- No accuracy loss



[Y.-C. Chiu (NTHU), et al, ISSCC, 2022]

Analog In-Memory

- Mixed signal computation
- Multiply in cell array
- Accumulation in peripheral analog or digital circuit
- High energy efficiency



[S. D. Spetalnick (Georgia Tech), et al., ISSCC, 2022]
[W.-S. Khwa (TSMC), et al, ISSCC, 2022]
[J.-M. Hung (NTHU), et al, ISSCC, 2022]

Chang et. al., VLSI (2022)

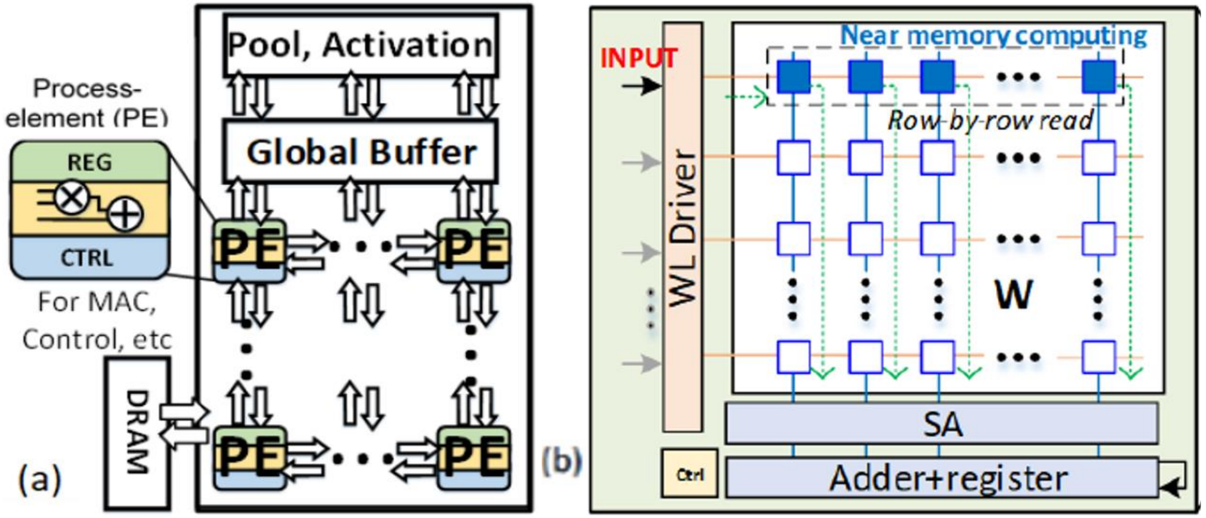
- Digital: Multiply and accumulation in peripheral digital logic → No accuracy loss.
- Analog: Multiply in Cell array, Accumulation in peripheral analog/digital circuit → High energy efficiency/parallel computing.

Digital *versus* Analog Computing in Memory

Digital Near-Memory

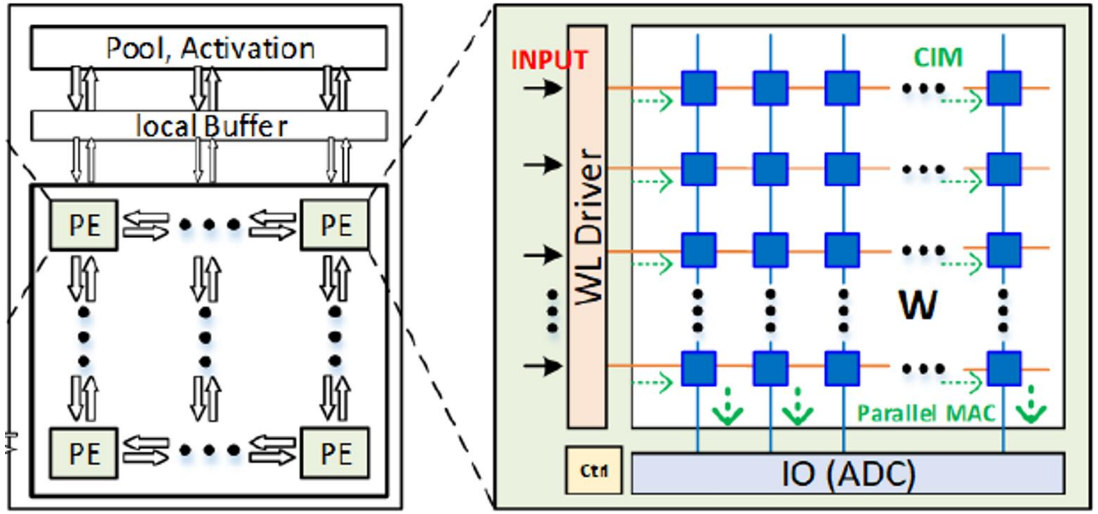
Digital accelerator (TPU-like)

Near-memory-compute accelerator (row-by-row)



Analog In-Memory

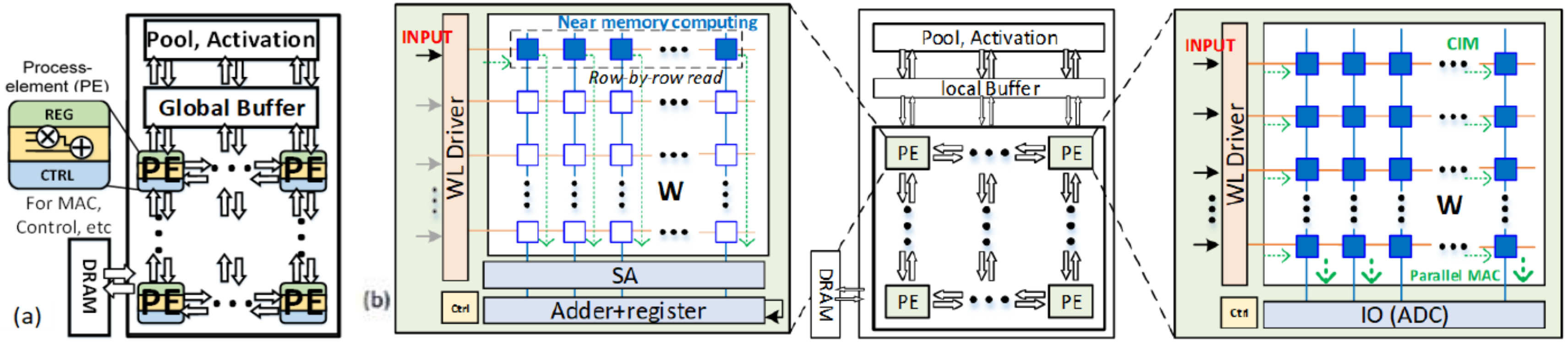
In-memory-compute accelerator (parallel)



IEDM 2020 Short Course 2: Analog Memory Needs for AI (GIT)

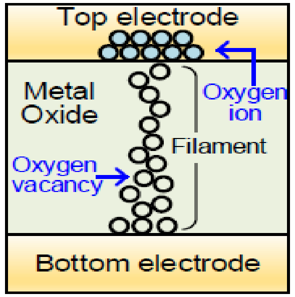
- Digital:
  - ✓ Single row access (TPU-like).
  - ✓ Row by Row with digital address at periphery.
- Analog: Parallel access and ADC for total sum quantization.

Analog Computing in Memory Candidates

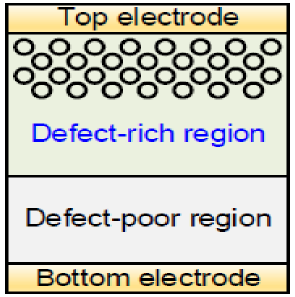

TPU-like digital accelerator

Near-memory-compute accelerator (row-by-row)

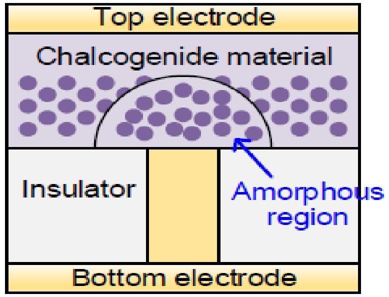
In-memory-compute accelerator (parallel)



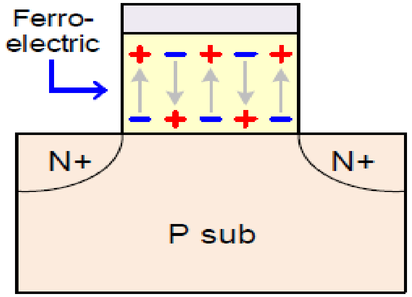
Filamentary RRAM



Non-filamentary RRAM



Phase Change Memory



Ferroelectric FET

CAAC-IGZO + MIM w/SL FE

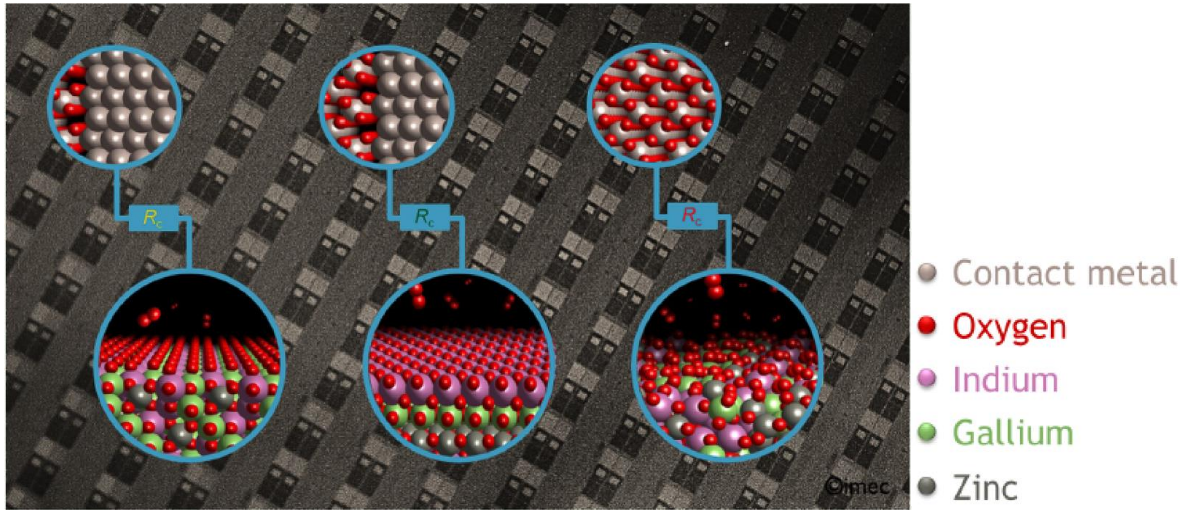
This Work

IEDM 2020 Short Course 2: Analog Memory Needs for AI (GIT)

- Analog: Parallel access and ADC for total sum quantization.
- RRAM, PCRAM, FE-RAM can be used for Analog computing.
- In this work, we propose new CAAC-IGZO + MIM w/SL-FE scheme.

IGZO: Material Properties

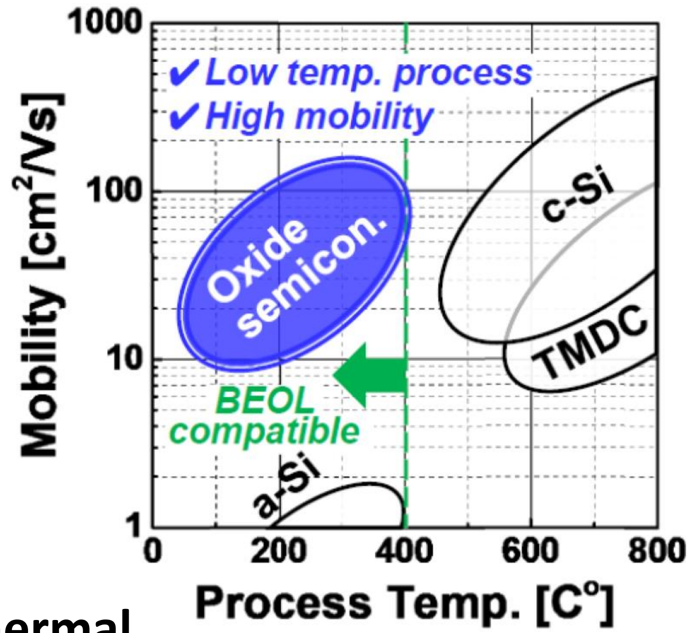
④ **Good uniformity**

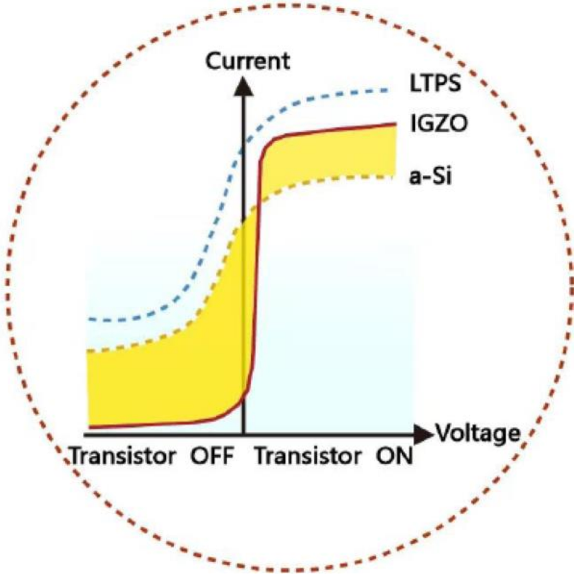


ACS Appl. Electron. Mater. Vol. 3(9), p. 4037, (2021)

② **Moderate mobility**

① **Ultra-low off-state current**



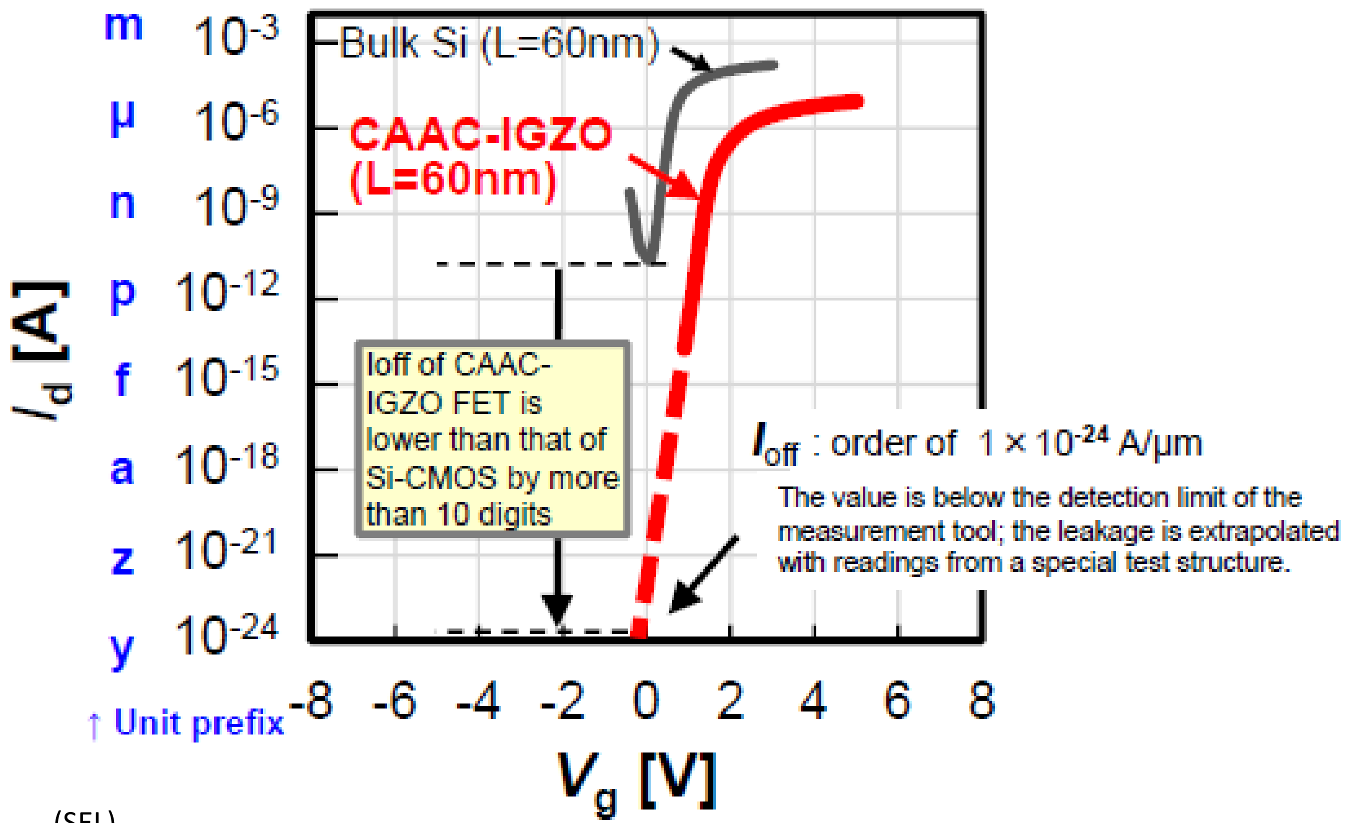


③ **Low thermal budget**

Wu et. al., VLSI (2020) and E. Fortunato et. al., Adv. Mater., (2012)

- IGZO is BEoL compatible, with acceptable mobility, low off-state current, low power consumption, high scalability, and excellent uniformity.

IGZO: Monolithic Properties

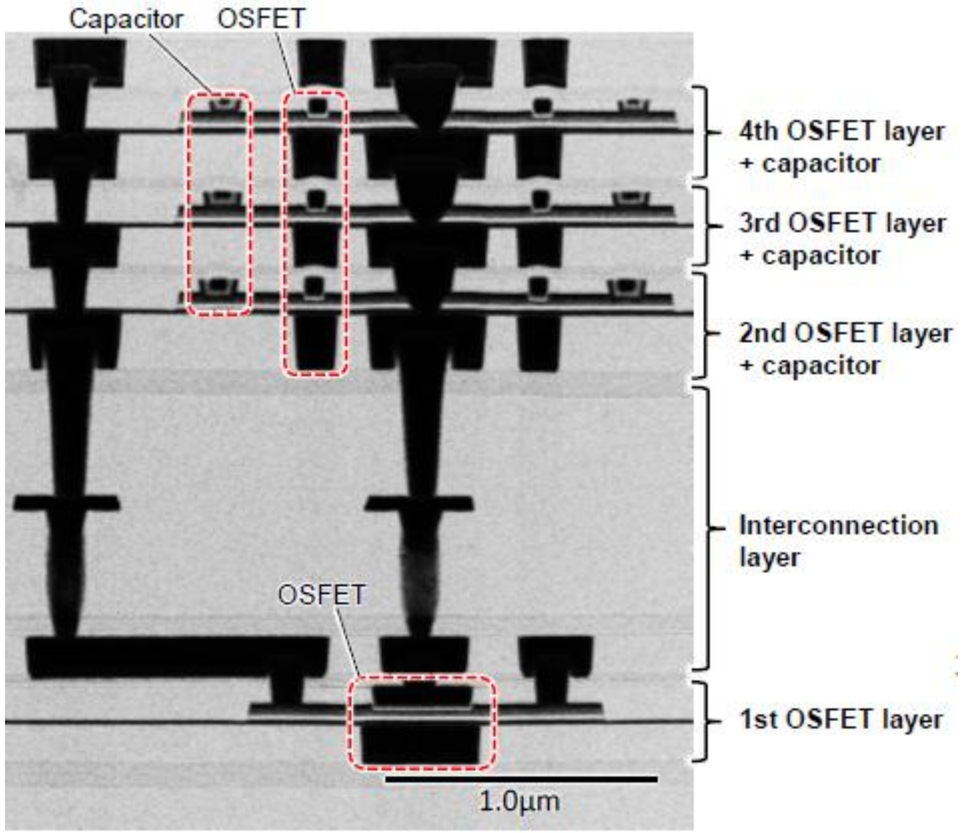


Ioff of CAAC-IGZO FET is lower than that of Si-CMOS by more than 10 digits

I_{off} : order of 1×10^{-24} A/μm

The value is below the detection limit of the measurement tool; the leakage is extrapolated with readings from a special test structure.

(SEL)



Capacitor
OSFET
4th OSFET layer + capacitor
3rd OSFET layer + capacitor
2nd OSFET layer + capacitor
Interconnection layer
OSFET
1st OSFET layer
1.0μm

CIMTEC Plenary talk 2022 (SEL)

- IGZO I_{off} ~ 10^{-24} A/μm, I_{on}/I_{off} ~ 10^{19}.
- Most other CIM chips (RRAM,..) suffer high read current with high IR power consumption.
- IGZO Low Temp. process with monolithic in BEoL is suitable for multi-stackings.

Highlights in this work

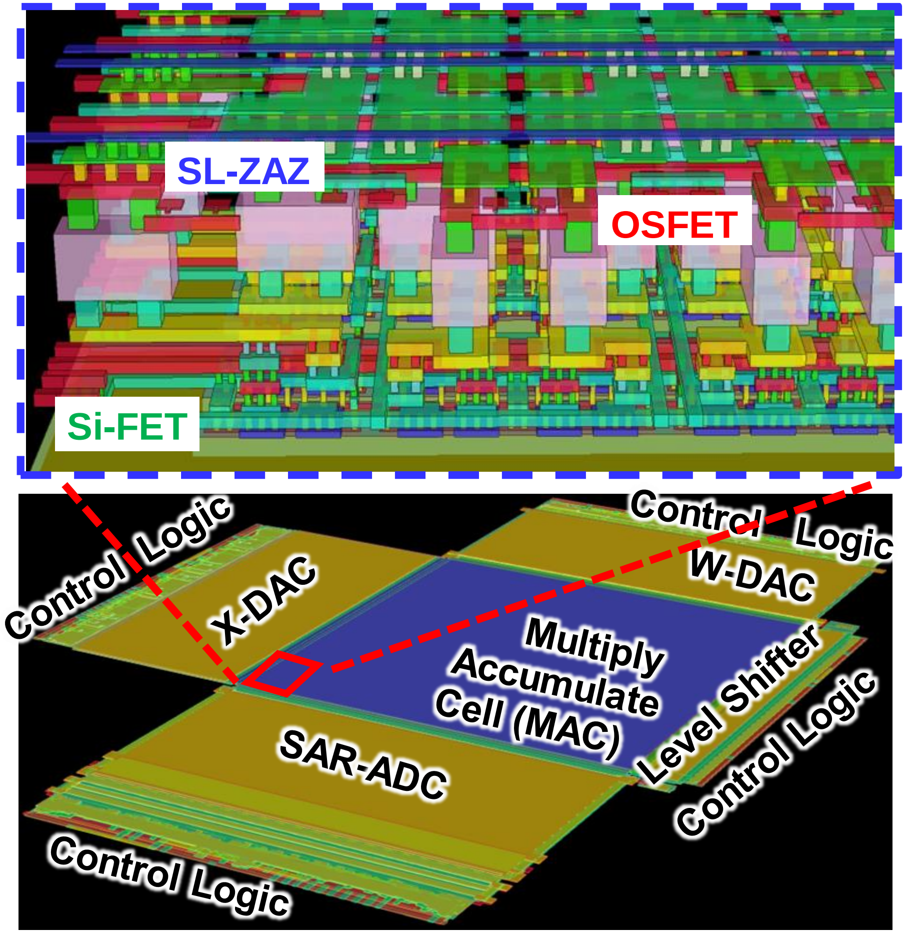
- Monolithic OS (IGZO) as BEoL device
- SL $ZrO_x/AlO_x/ZrO_x$ (SL-ZAZ) MIM capacitor
- Multi-states and High efficient Analog Memory Macro
- Efficiency, Temperature effect, Stability, and Reliability

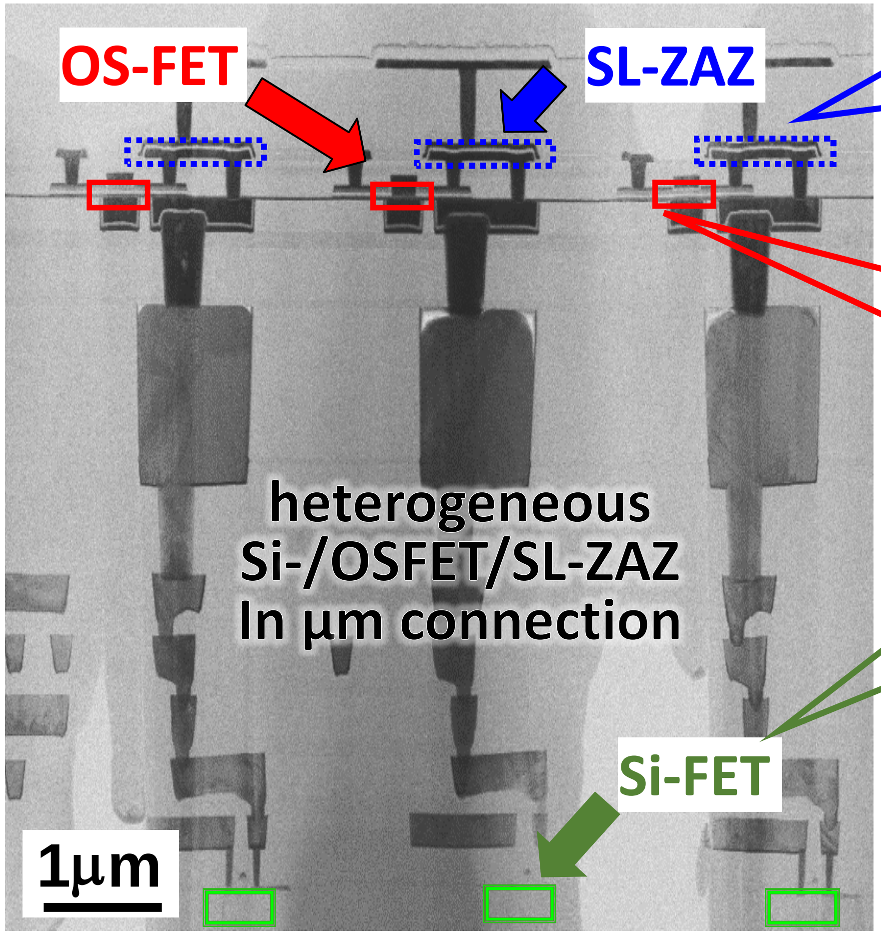
Outline

- Introduction
- Experiments and Fabrications
- AiMC Chip Design and Operation
- Performance, Stability, and Reliability
- Benchmark & Conclusion

Monolithic 3D Integration

- Integrated FEoL Si devices + BEoL OS devices and SL MIM to achieve Analog in-memory computing (AiMC) macro.





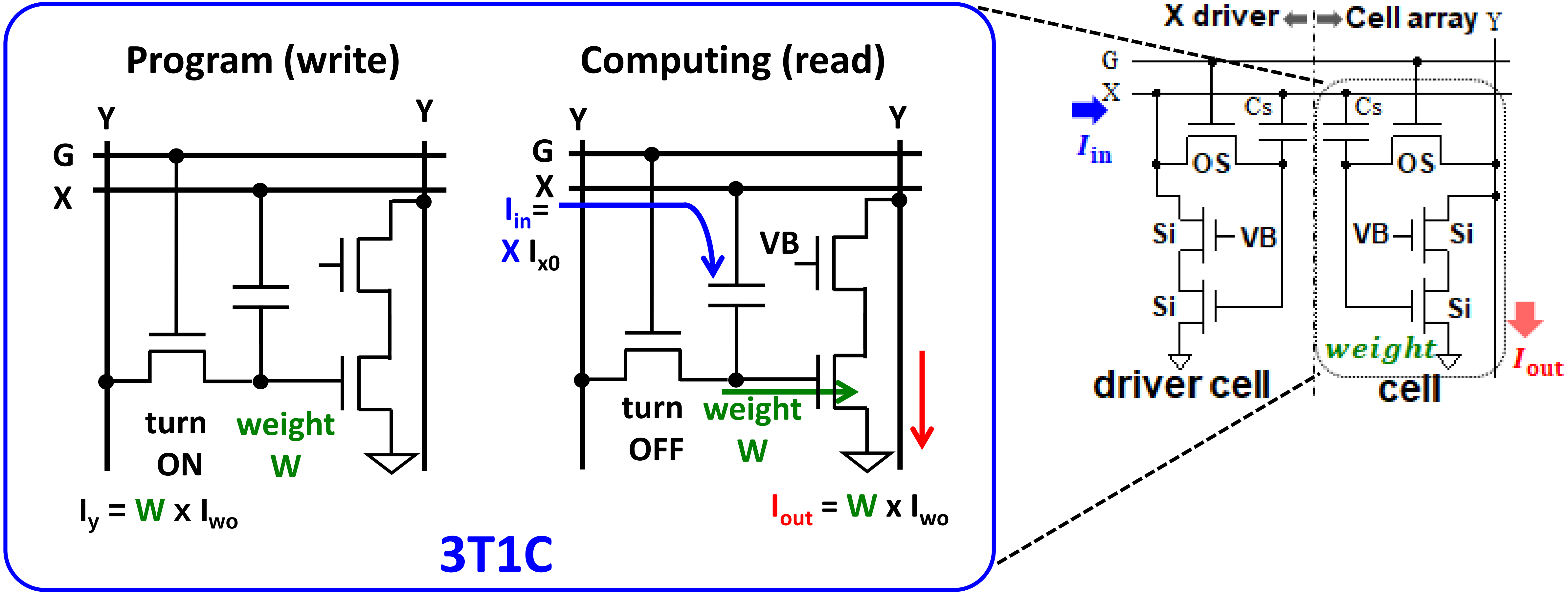
- MIM Capacitor
  - SL-ZAZ

- OS Devices
  - Low leakage

- Si Devices
  - CMOS

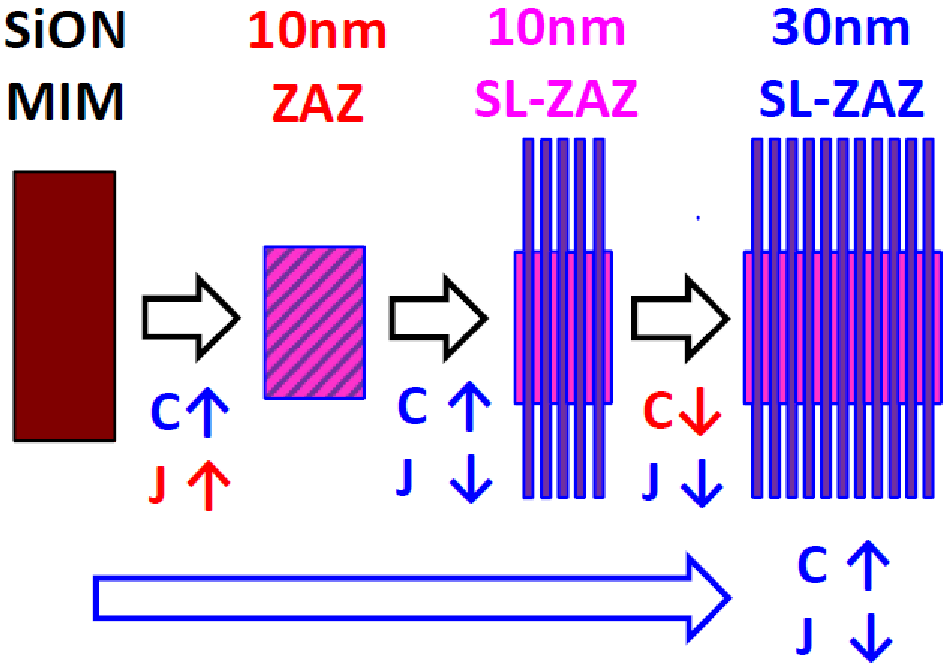
Cell Array Structure

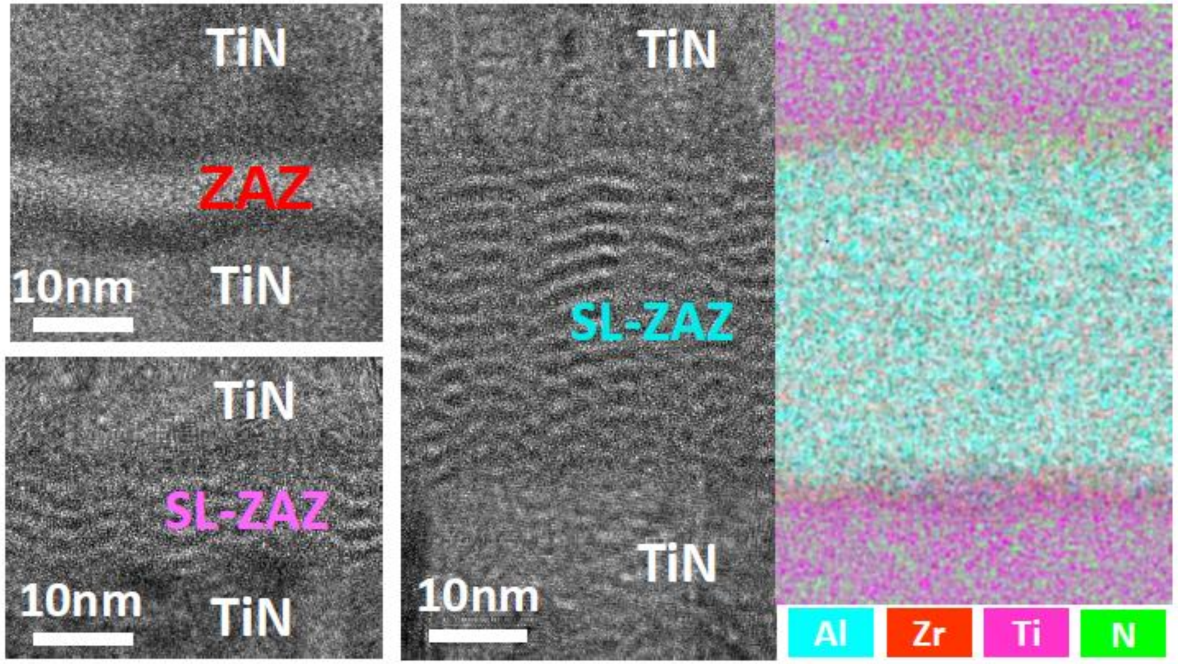
- Memory cell operates at sub-threshold with nA read current.
- Analog multiplication: I_{out} = Weight (W) × I_{in} (X).



Super-Lattice $ZrO_x/AlO_x/ZrO_x$ (SL-ZAZ) in MIM

- The laminated-ZAZ structure is processed by ALD.
- Optimize the ZAZ stacked structure for C & J performance.



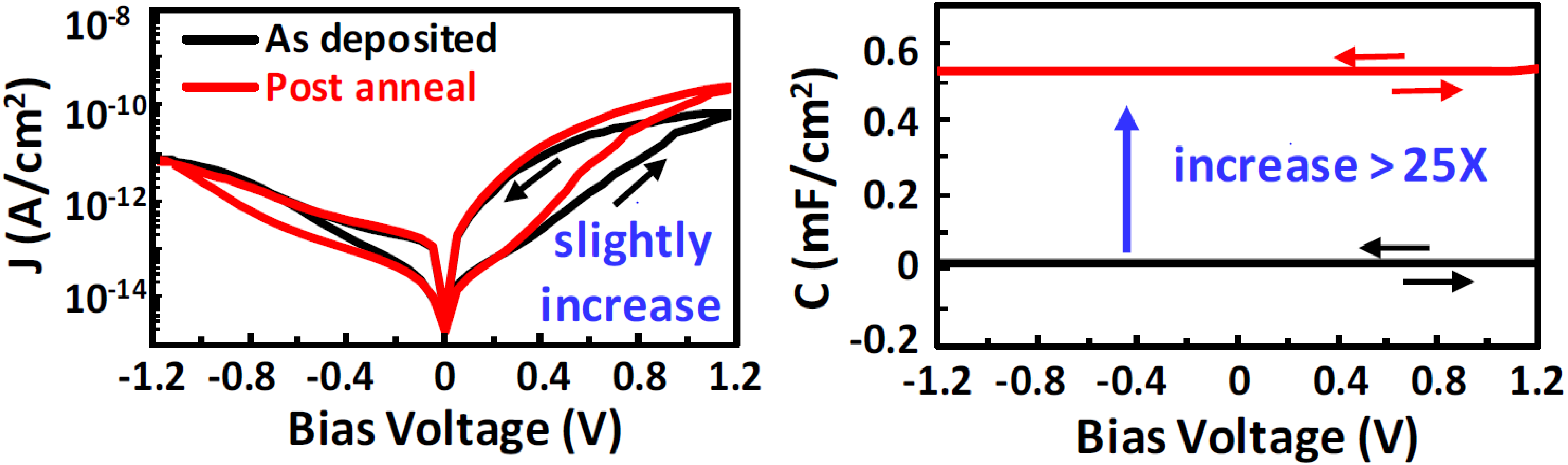


Sample from NTNU Prof. M.-H. Lee's group

ZAZ: K=17
SL-ZAZ: K=22

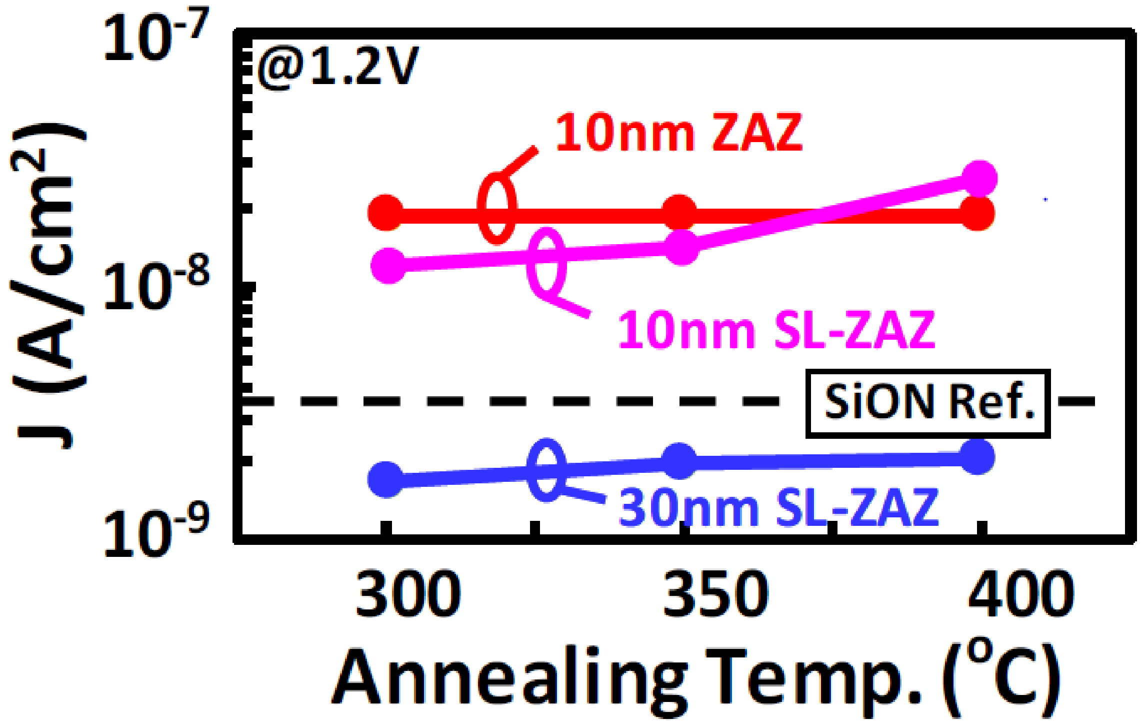
Super-Lattice $ZrO_x/AlO_x/ZrO_x$ (SL-ZAZ): RTA effect

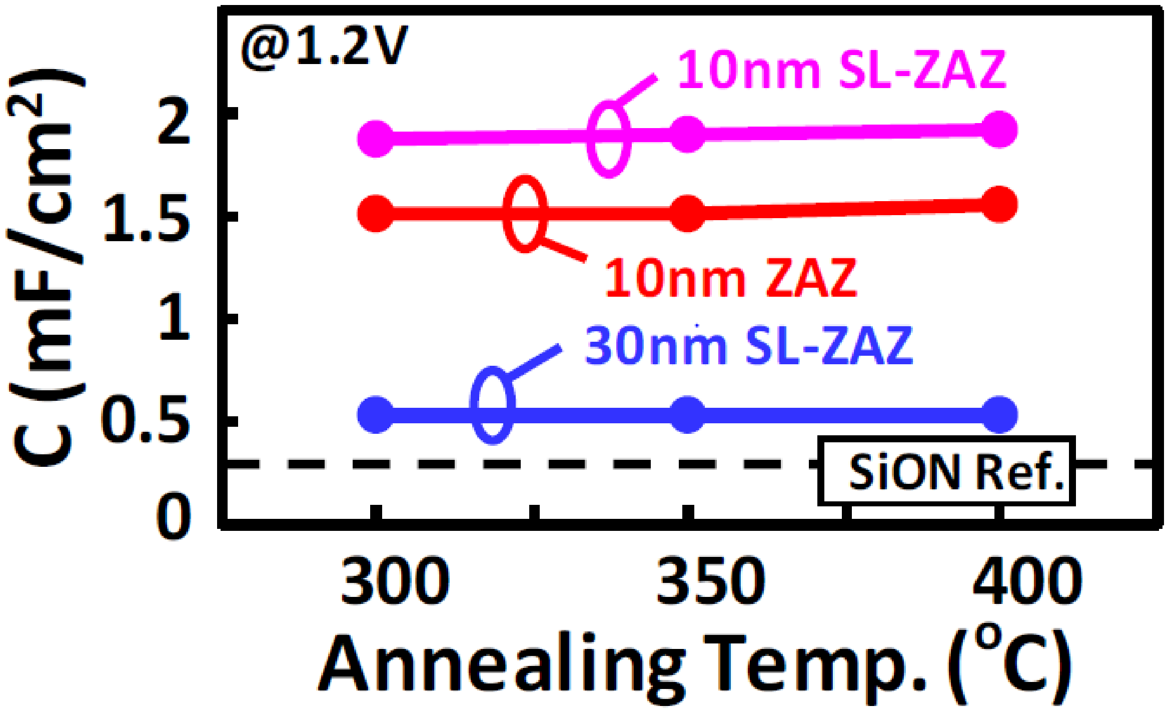
- RTA annealing for SL-ZAZ.
- **25X** Capacitance improvement.



Super-Lattice $ZrO_x/AlO_x/ZrO_x$ (SL-ZAZ)

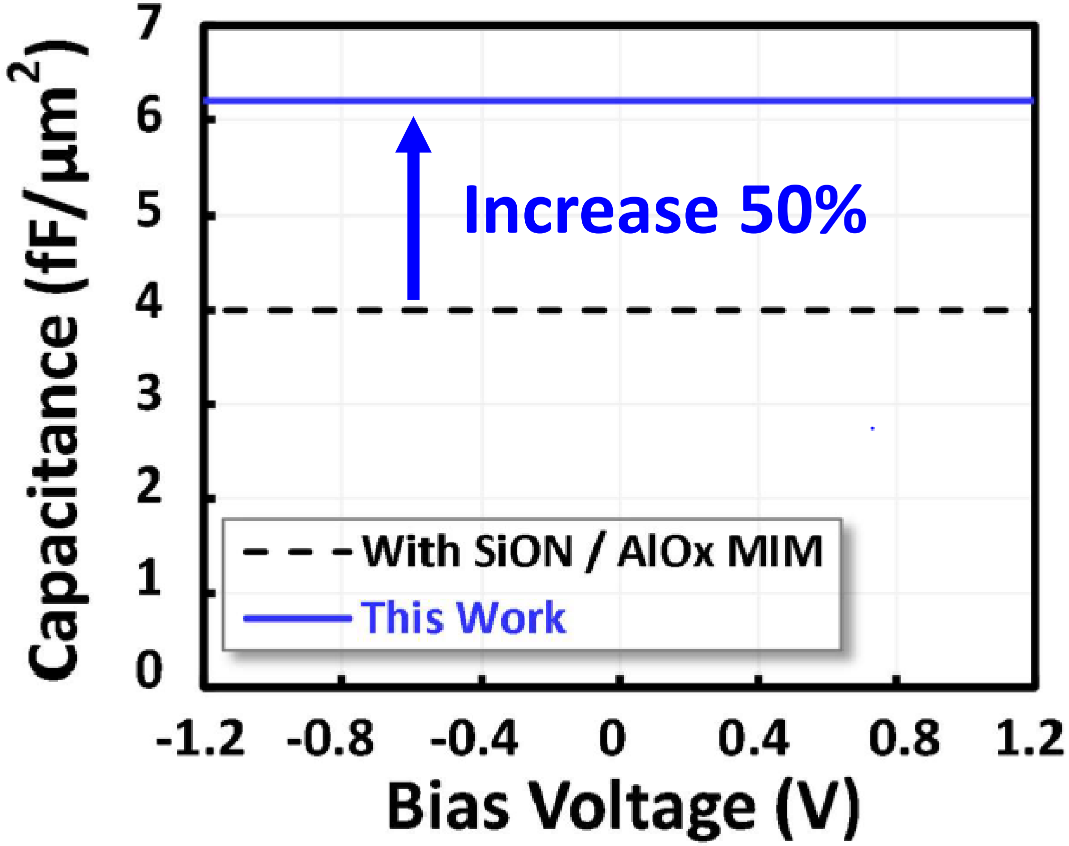
- 30 nm ZAZ & 400 °C anneal is used.

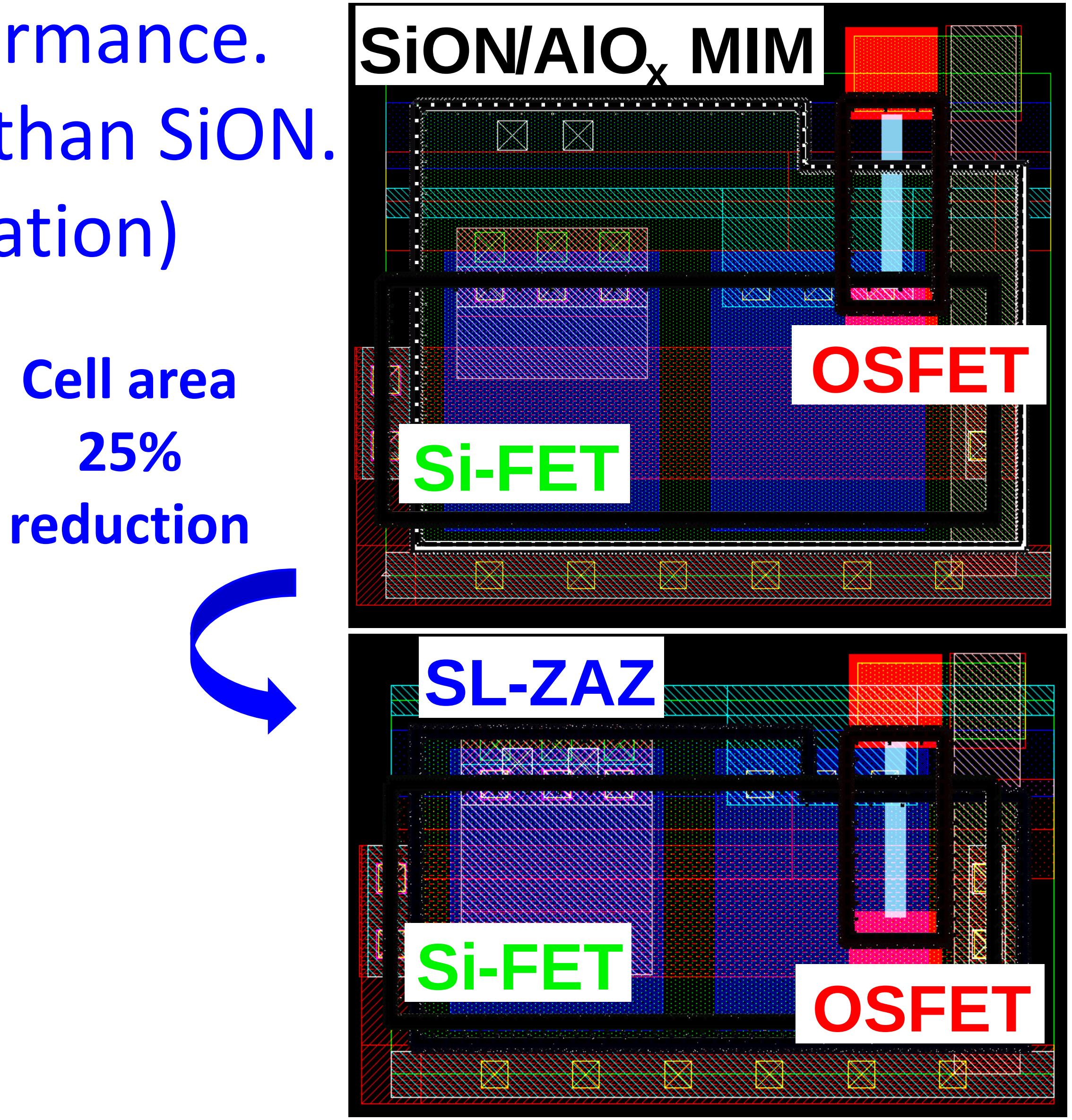




Super-Lattice $ZrO_x/AlO_x/ZrO_x$ (SL-ZAZ)

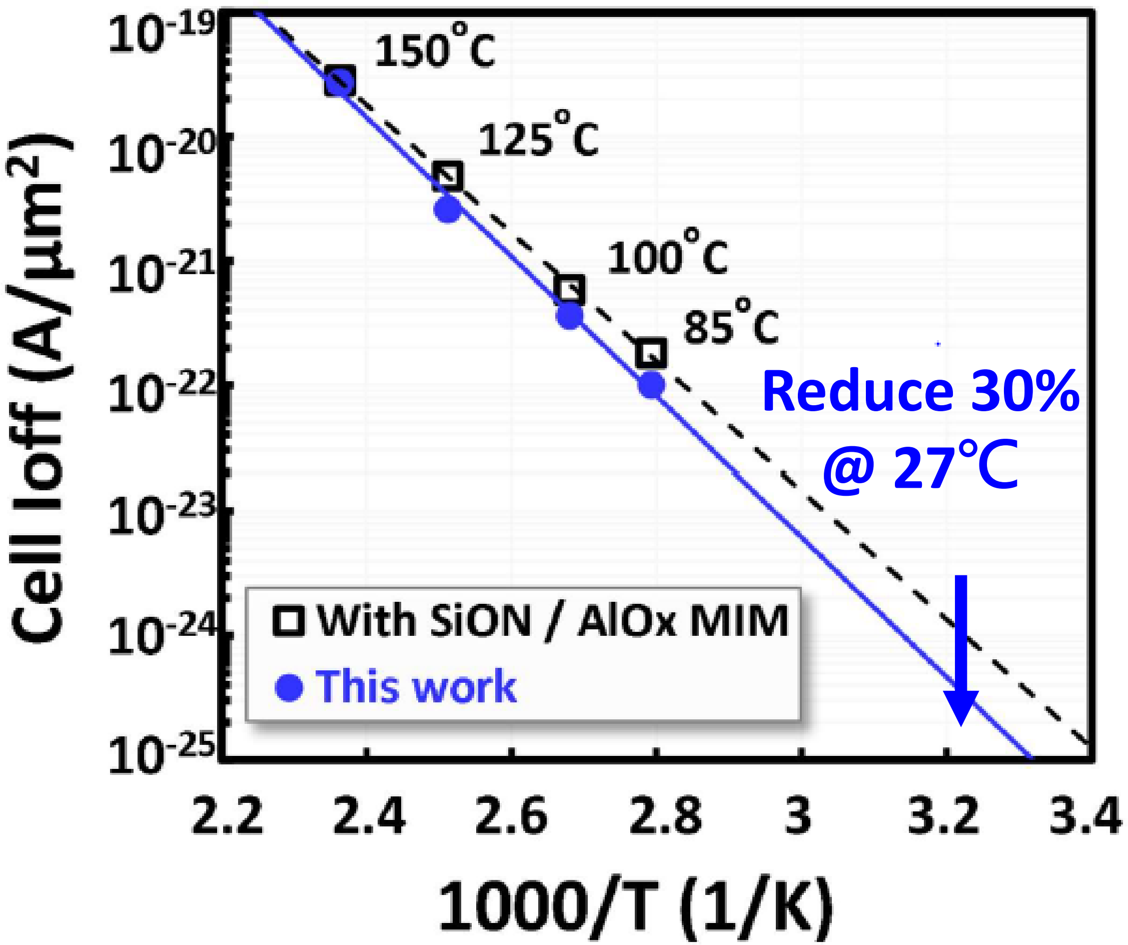
- SL-ZAZ has better capacitor performance.
  - **50%** capacitance improvement than SiON.
  - **25%** Cell area reduction. (calculation)





Super-Lattice $ZrO_x/AlO_x/ZrO_x$ (SL-ZAZ)

- SL-ZAZ has better capacitor performance than SiON MIM.
  - **30%** Cell I_{off} reduction.



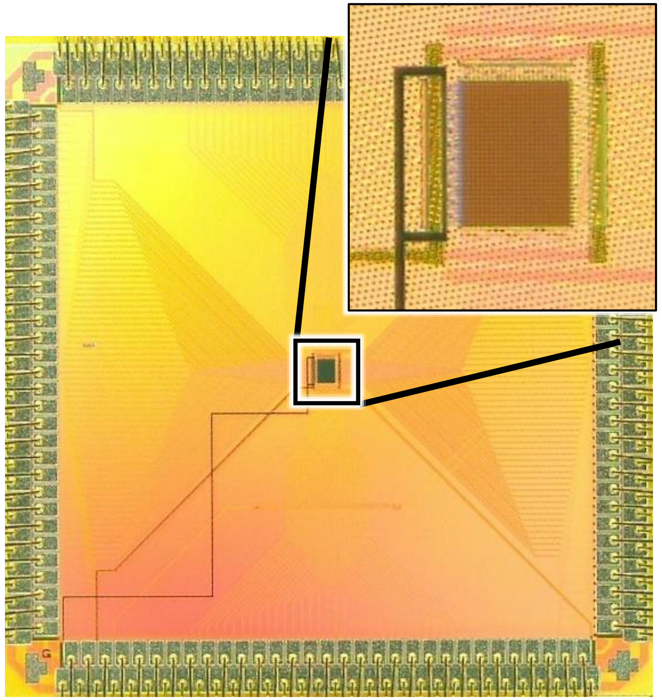
Outline

- Introduction
- Experiments and Fabrications
- AiMC Chip Design and Operation
- Performance, Stability, and Reliability
- Benchmark & Conclusion

Analog in Memory Computing (AiMC) Chip

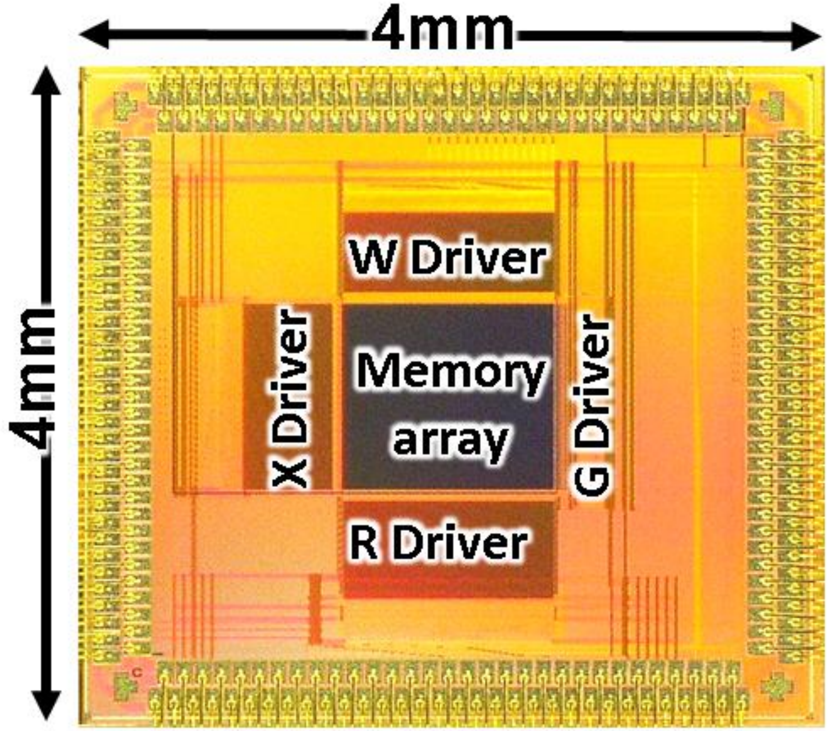
- The key chip performance and information.

IEDM 2021: 8 states
With SiON/AlOx MIM



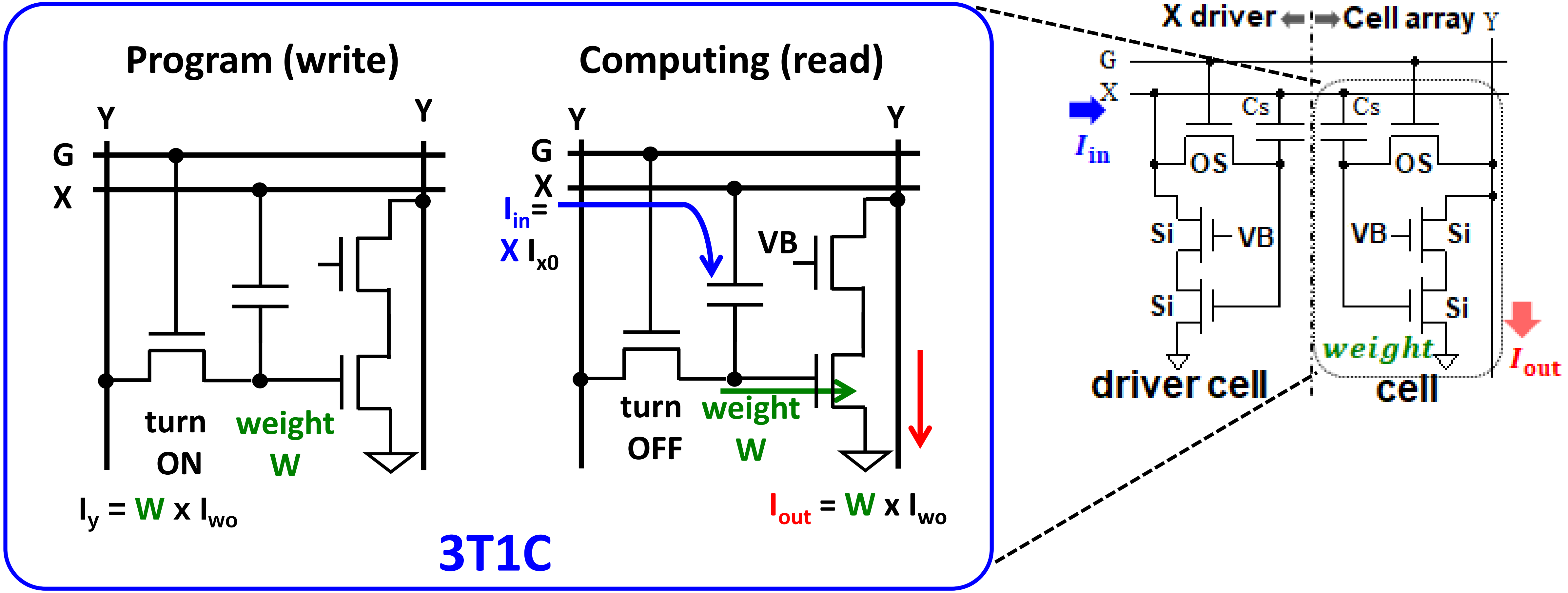
IEDM 2022: **64** states
With SL-ZAZ MIM

	Specifications
Die Area	4 mm x 4 mm
Precision	92.30%
Memory size	256 kb
Supply voltage	1.5 V
Frequency	Logic: 8MHz MAC: 20kHz
Performance	5.4 GOPS
Power Consumption	25.6 μW
Energy Efficiency	210 TOPS/W



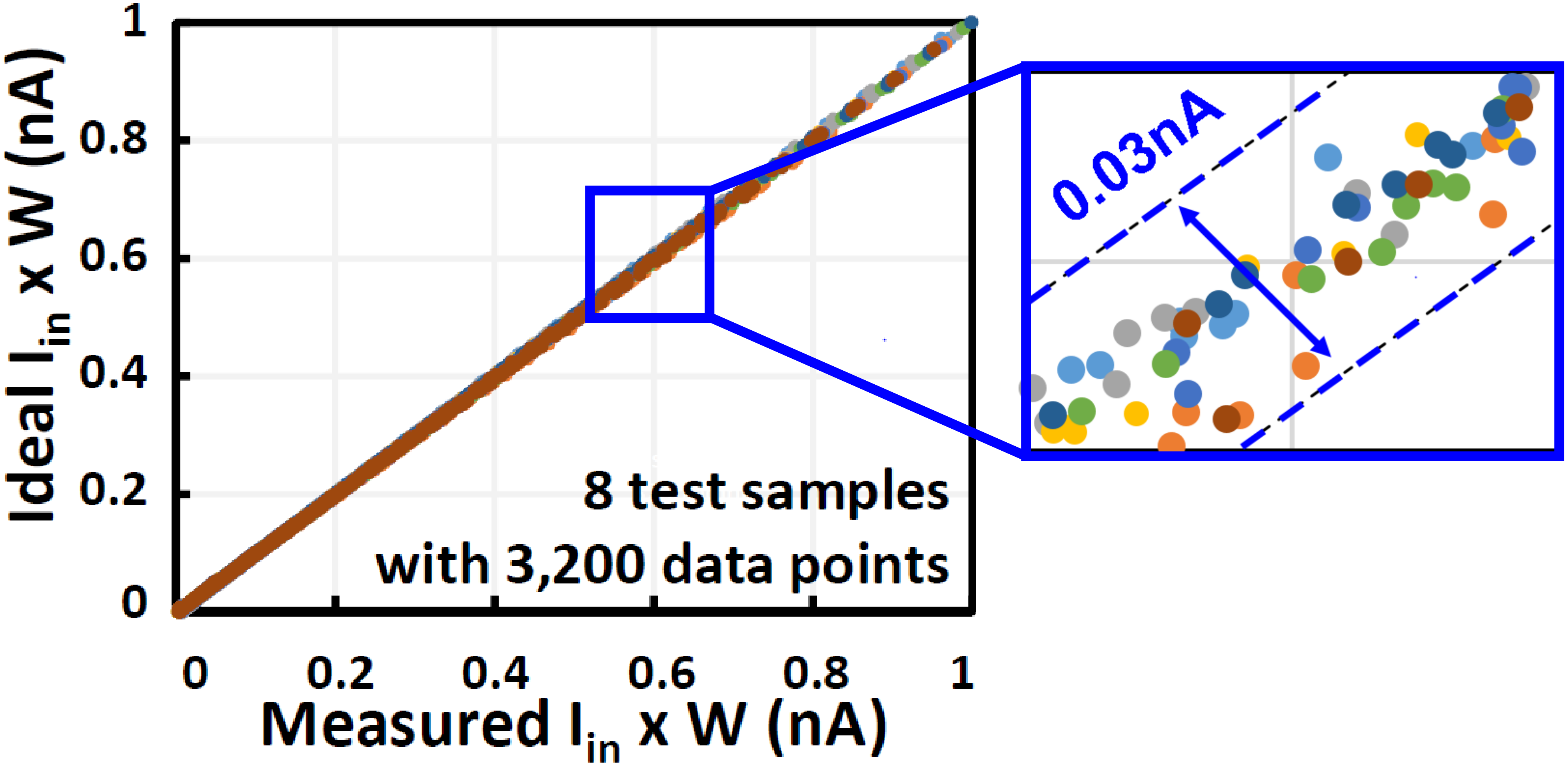
Cell Operation

- Our memory cell operate at sub-threshold with nA read current.
- Analog multiplication: I_{out} = Weight (W) × I_{in} (X).



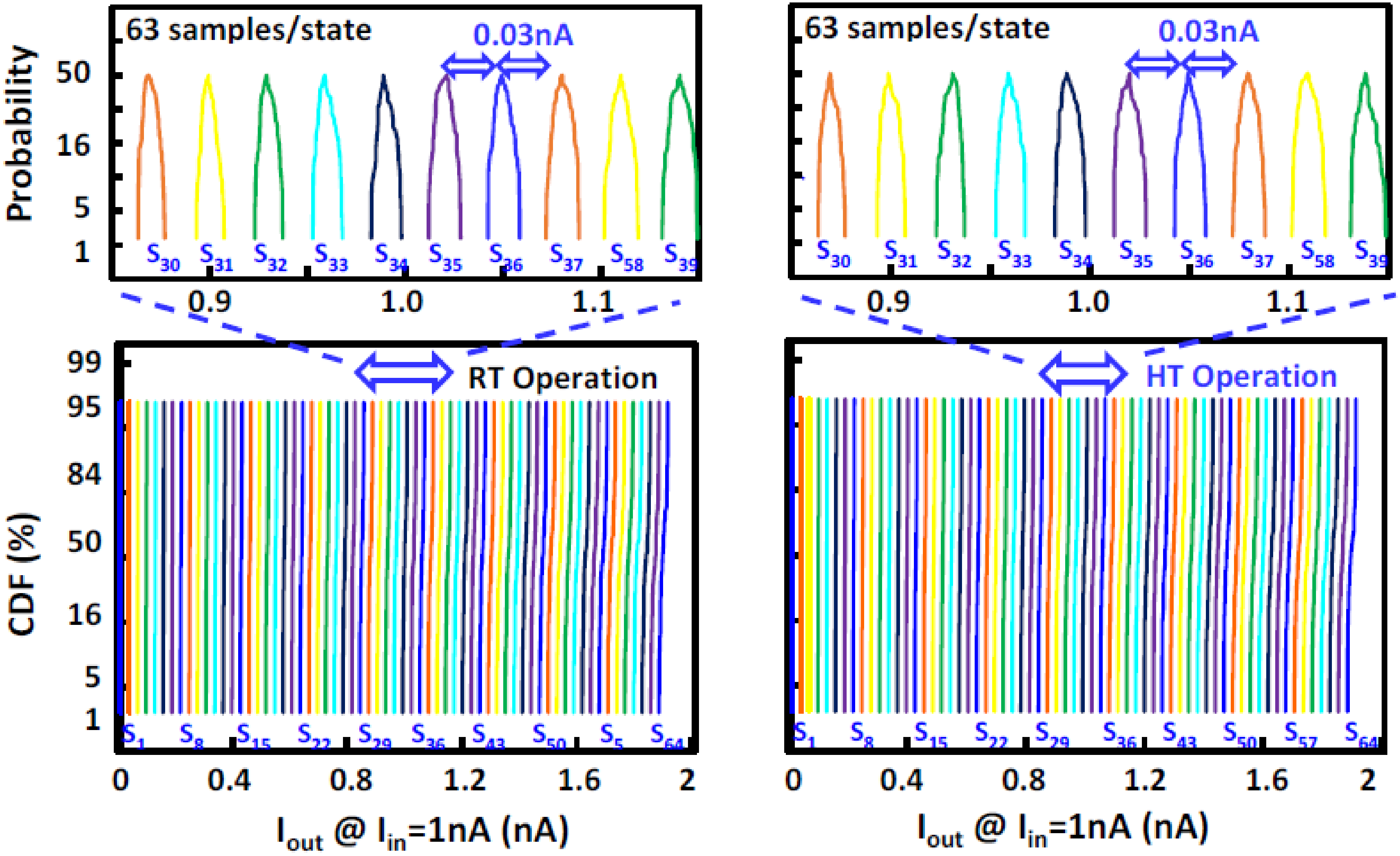
Analog Multi-states

- **0.03 nA** variation band width (w/ 3,200 data in 8 test samples).



Analog Multi-states

- **64** distinct weighting states at RT and HT (125°C) with highly cell-to-cell stability.

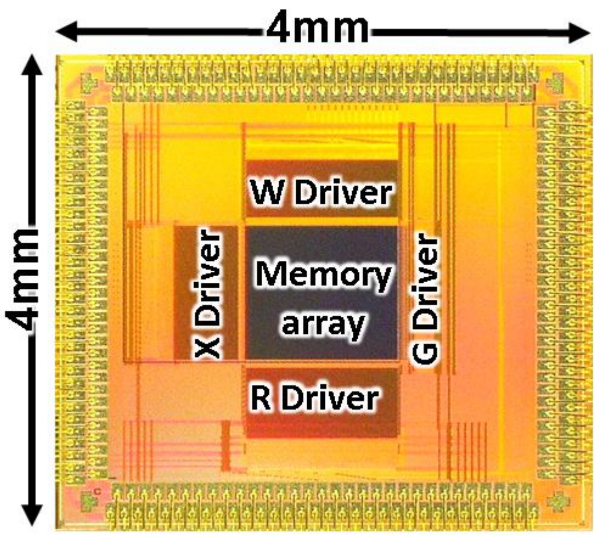


Outline

- Introduction
- Experiments and Fabrications
- AiMC Chip Design and Operation
- Performance, Stability, and Reliability
- Benchmark & Conclusion

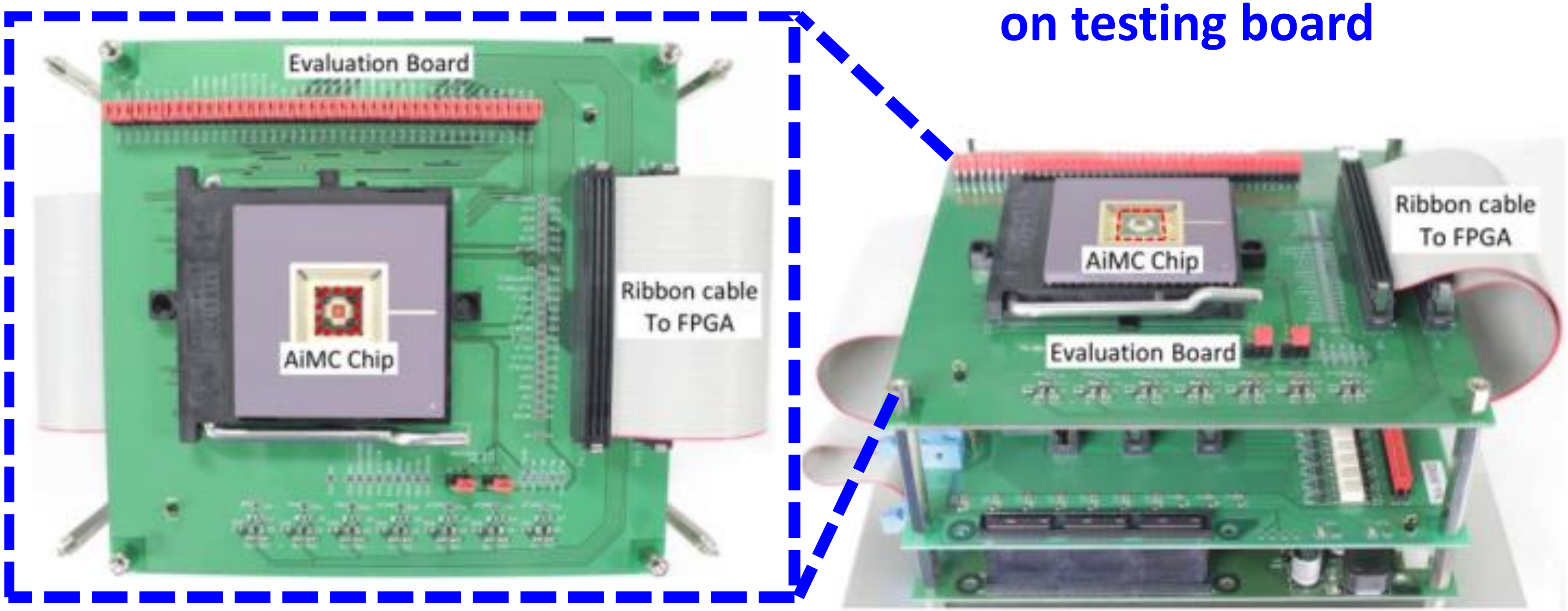
Test environment

- The measurement environment of AiMC.



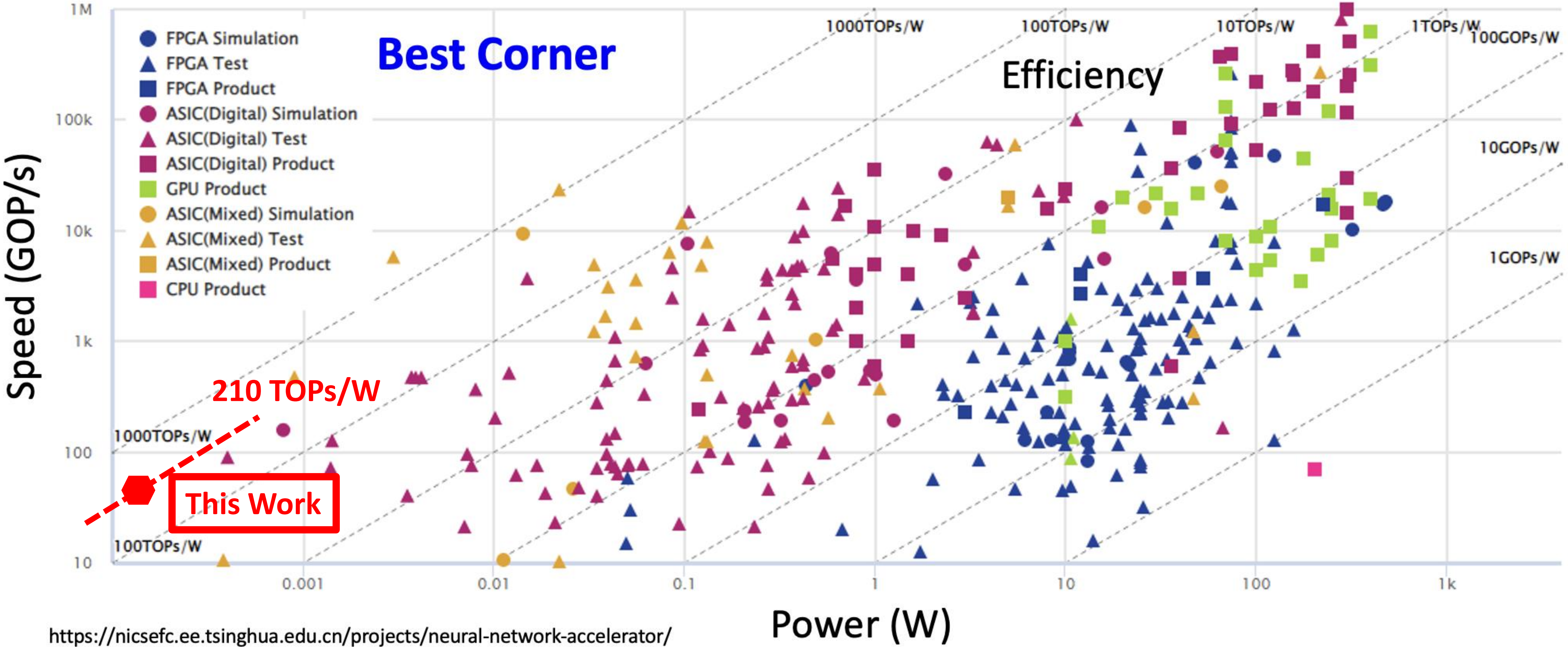
top view

Si/CAAC-IGZO AiMC chip on testing board



Performance

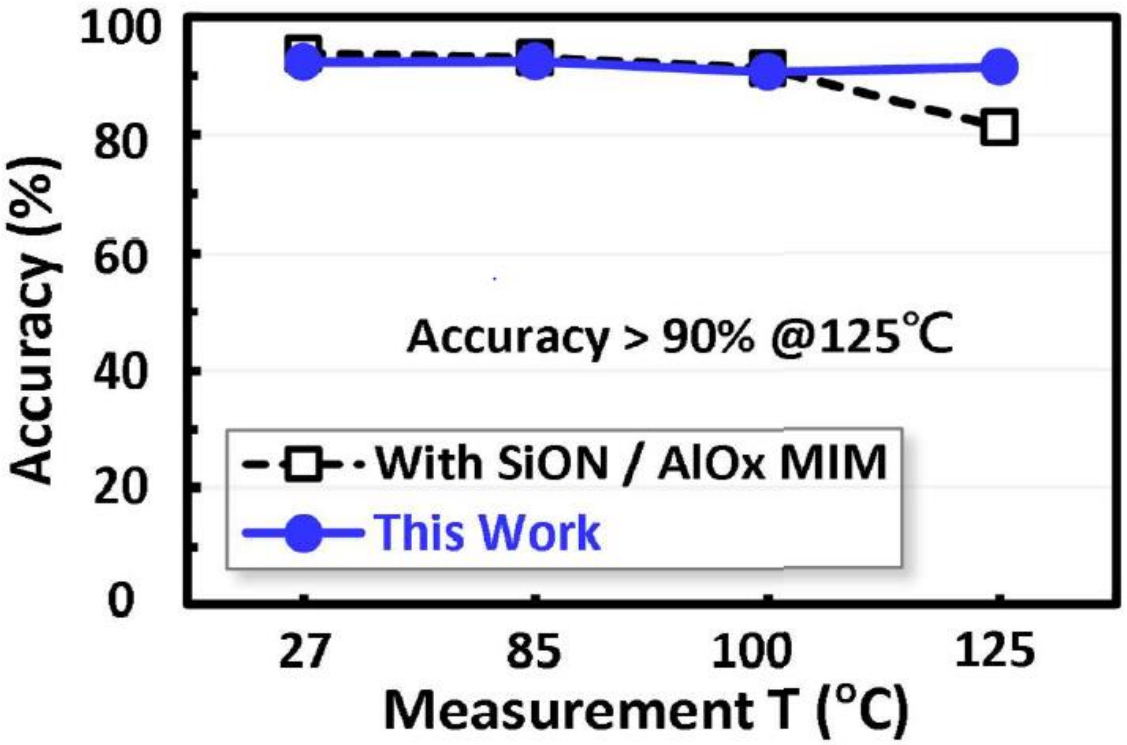
- **210 TOPS/W** and **25.6 µW** in our AiMC.



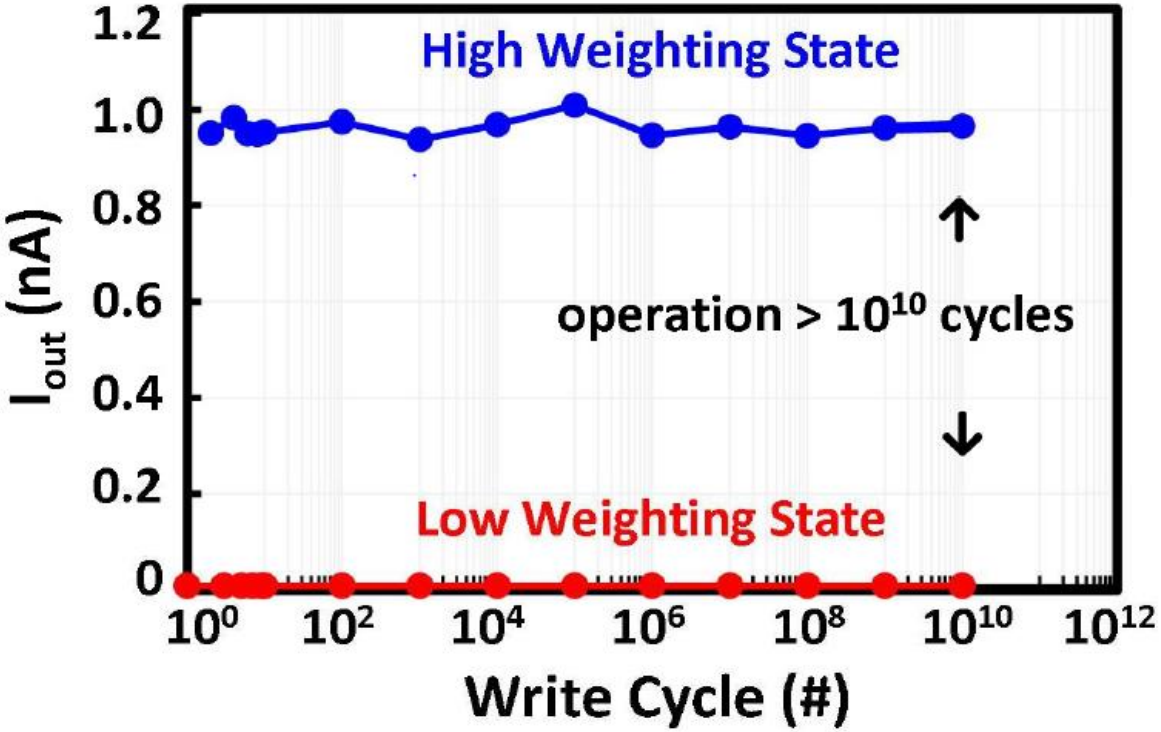
https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/

Stability and Temperature Effect

- **>90%** accuracy (MNIST) at both RT and HT.
- Stability is also promising.



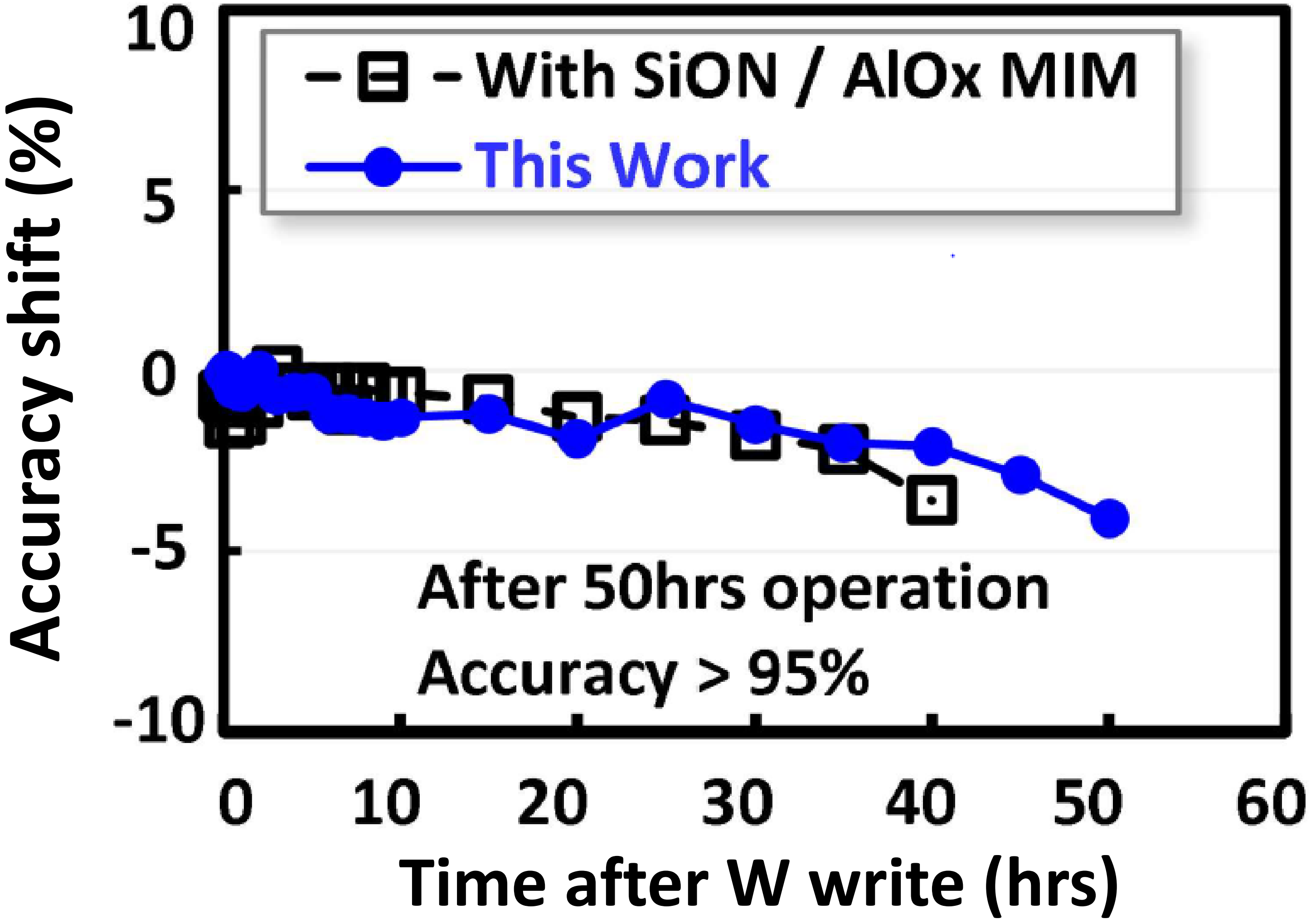
(Endurance)



Reliability Test

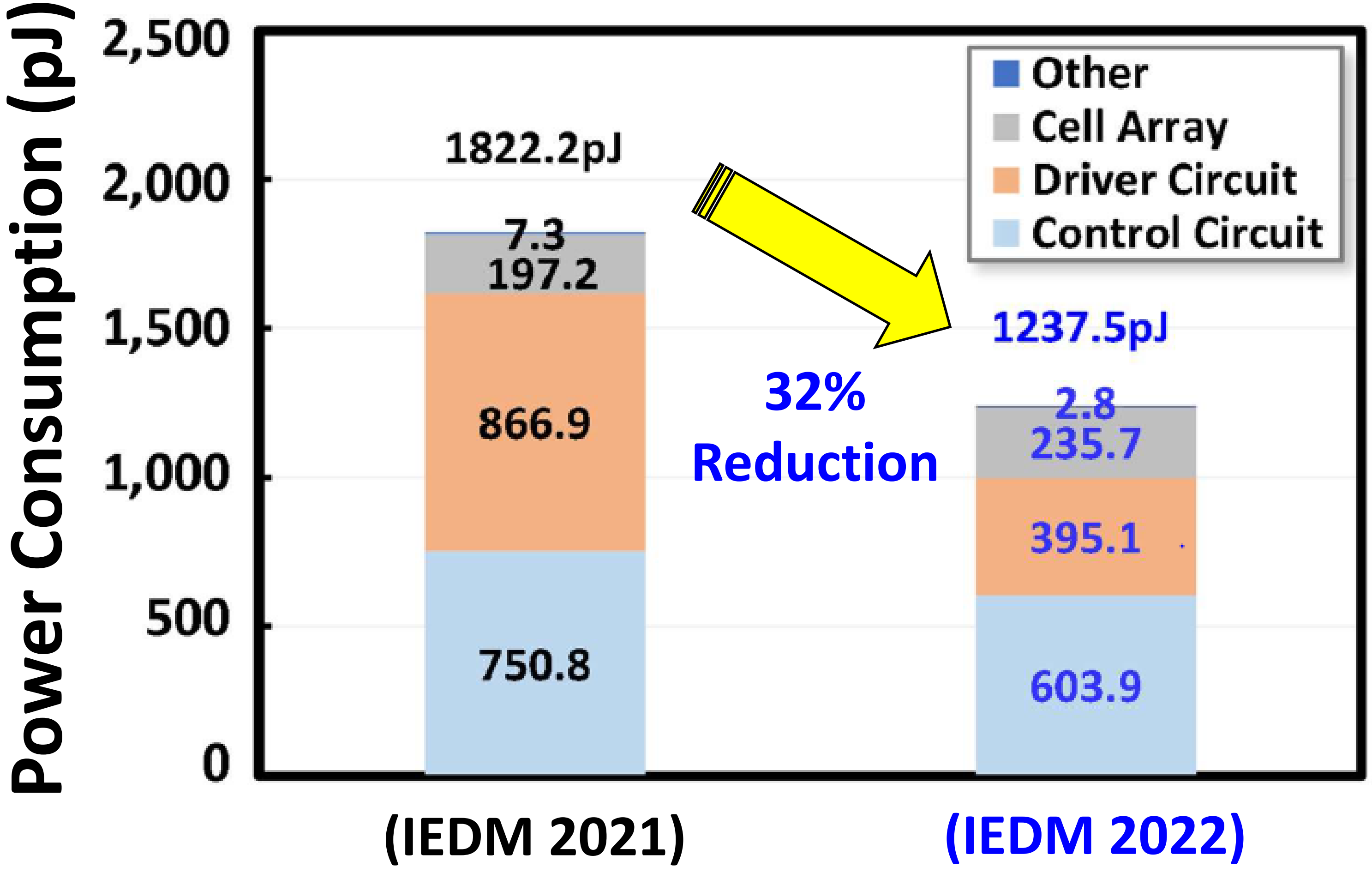
- **50** hrs operation for Reliability Test.

(Retention)



Power Consumption

- This AiMC is outperformance in terms of energy efficiency/consumption.
- **32%** power consumption reduction/improvement to IEDM 2021.

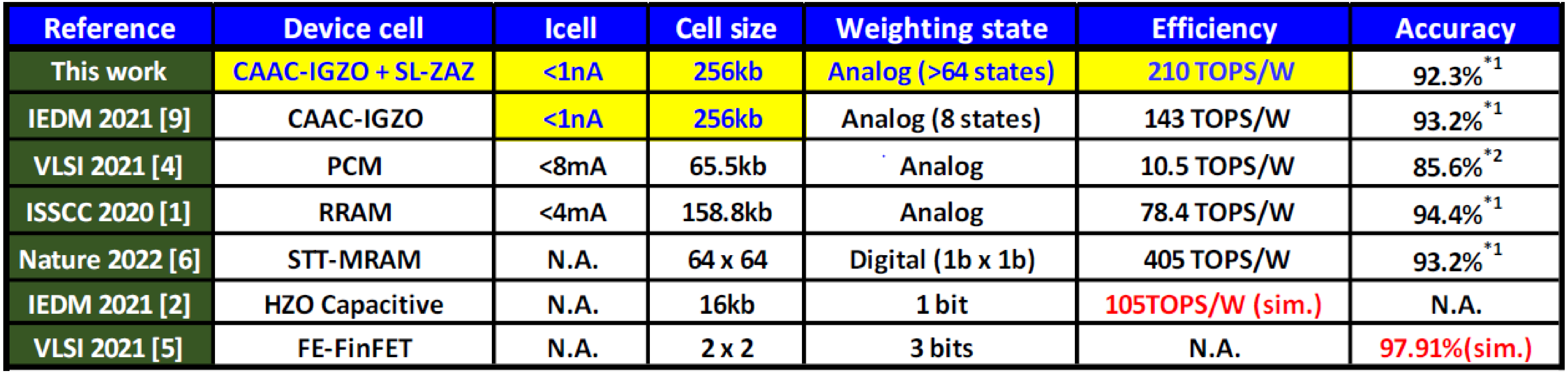


Outline

- Introduction
- Experiments and Fabrications
- AiMC Chip Design and Operation
- Performance, Stability, and Reliability
- Benchmark & Conclusion

Benchmark

- The CAAC-IGZO/Si hybrid AiMC is outperformance in terms of energy efficiency, multiple states for analog computing.
  - **210** TOPS/W
  - **64** states

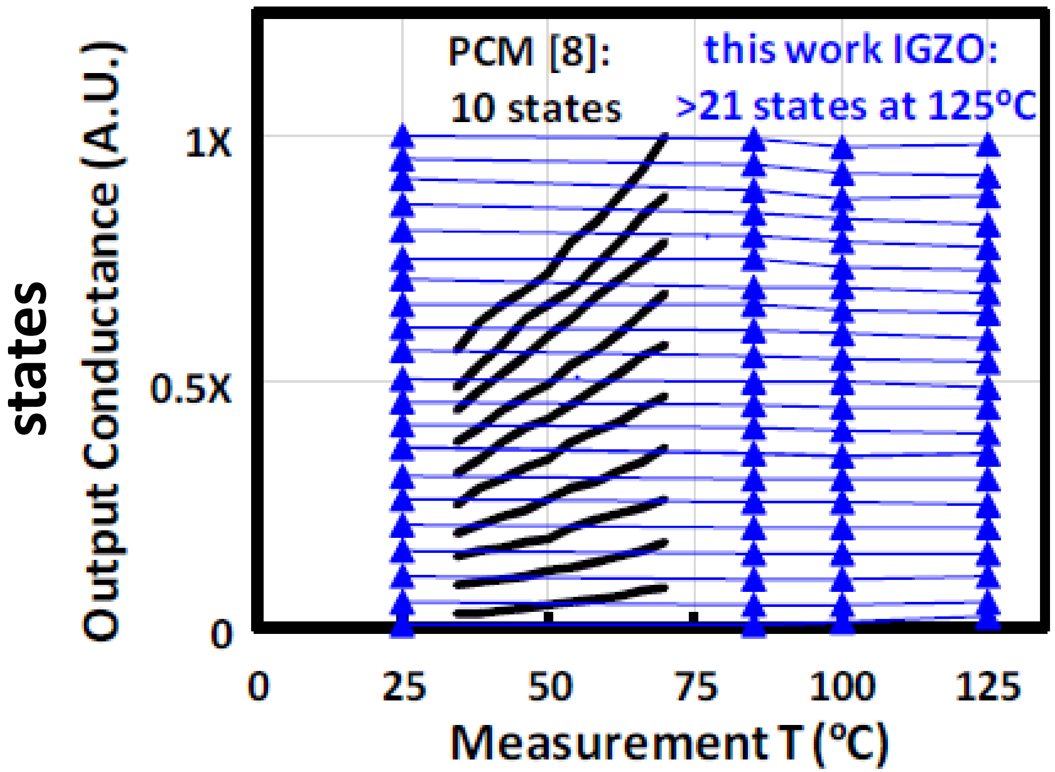


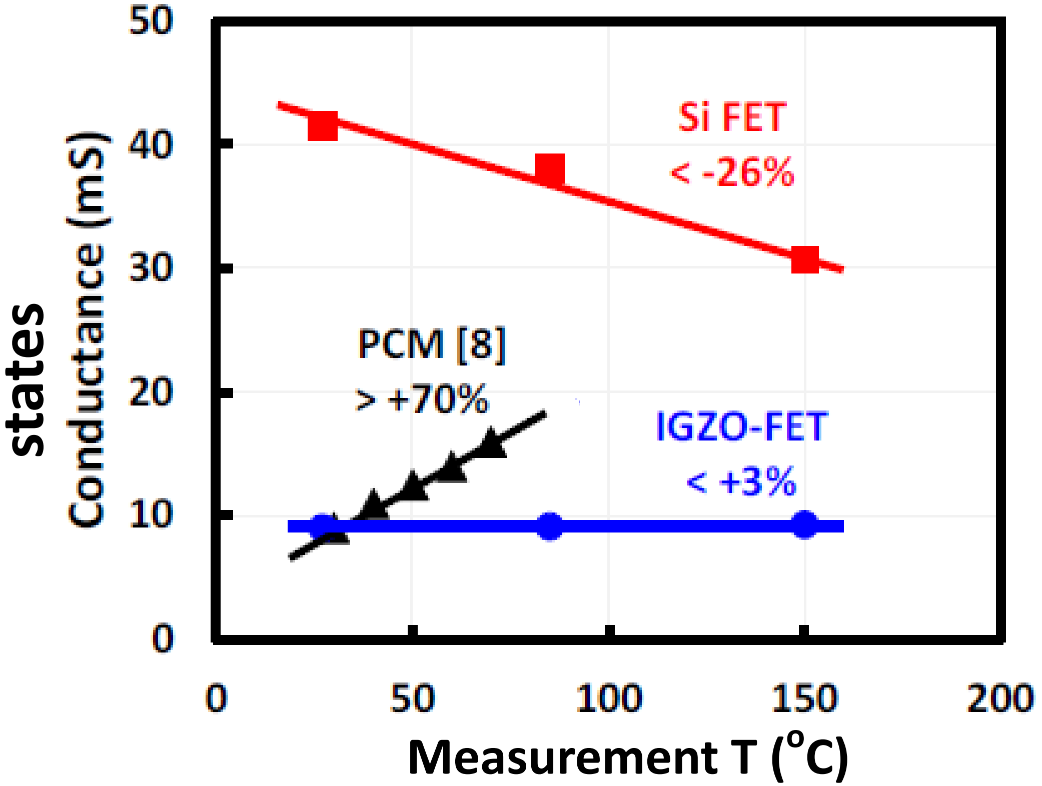
Reference	Device cell	Icell	Cell size	Weighting state	Efficiency	Accuracy
This work	CAAC-IGZO + SL-ZAZ	<1nA	256kb	Analog (>64 states)	210 TOPS/W	92.3%^{*1}
IEDM 2021 [9]	CAAC-IGZO	<1nA	256kb	Analog (8 states)	143 TOPS/W	93.2%^{*1}
VLSI 2021 [4]	PCM	<8mA	65.5kb	Analog	10.5 TOPS/W	85.6%^{*2}
ISSCC 2020 [1]	RRAM	<4mA	158.8kb	Analog	78.4 TOPS/W	94.4%^{*1}
Nature 2022 [6]	STT-MRAM	N.A.	64 x 64	Digital (1b x 1b)	405 TOPS/W	93.2%^{*1}
IEDM 2021 [2]	HZO Capacitive	N.A.	16kb	1 bit	105TOPS/W (sim.)	N.A.
VLSI 2021 [5]	FE-FinFET	N.A.	2 x 2	3 bits	N.A.	97.91%(sim.)

*1 MNIST; *2 CIFAR-10

Benchmark

- The CAAC-IGZO/Si hybrid AiMC is outperformance in terms of energy efficiency, multiple states for analog computing.
  - **125 °C** High Temp. operation capability

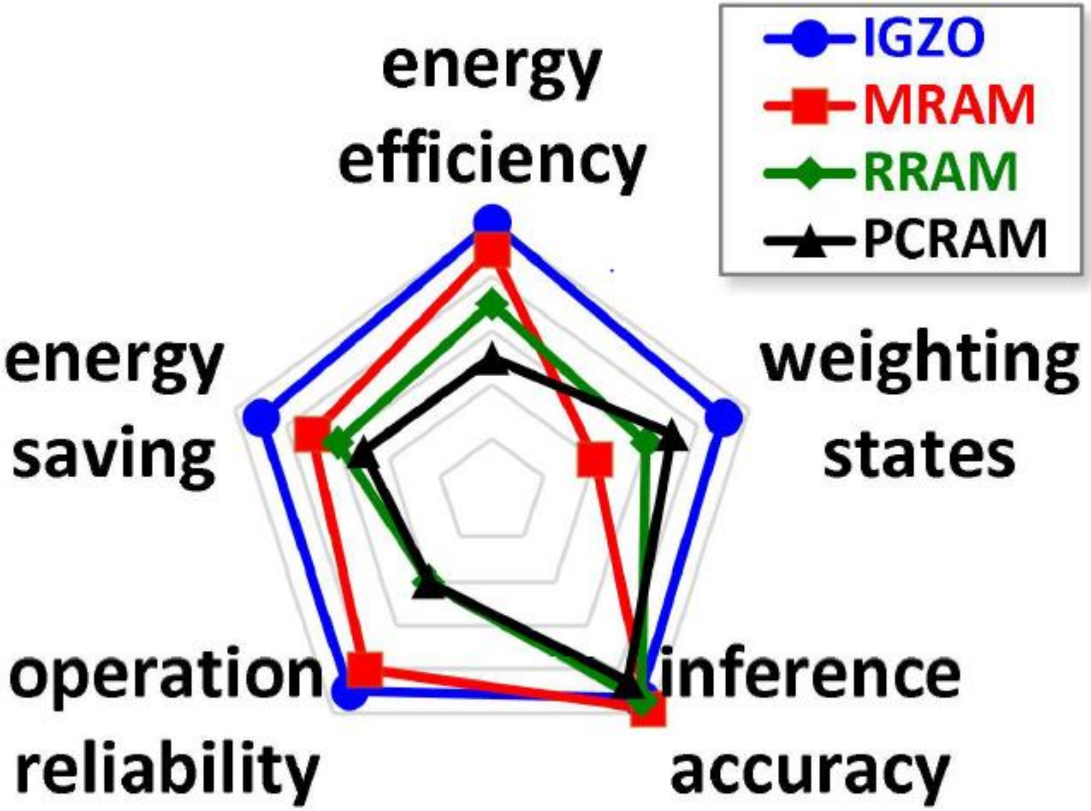




[8] I. Boybat, IEDM, 2021, pp.609.

Benchmark

- The CAAC-IGZO/Si hybrid AiMC is outperformance in terms of energy efficiency, multiple states for analog computing.



[Ref]
MRAM: S. Jung, Nature, 2022.
RRAM: H. Jia, JSSC, 2020.
PCRAM: R. Khaddam-Aljameh, VLSI, 2021.

Conclusions

- The monolithic CAAC-IGZO/Si technology is fully integrated with available Si CMOS process + SL MIM.
- The CAAC-IGZO/Si hybrid CMOS ring oscillator is capable of reducing 25% layout area.
- The AiMC with 210 TOPS/W energy efficiency, > 64 weighting states, 92.3% inference accuracy, and 125^{o}C operation have been demonstrated.

Thanks

Contact Information:

National Taiwan University
Ming-Han (Miller) Liao
mhliaoa@ntu.edu.tw

National
Taiwan
University
國立臺灣大學

Back-Up