ASIA SOUTH PACIFIC DESIGN AUTOMATION CONFERENCE

北京航空航天大學
BEIHANG UNIVERSITY

A 24.65 TOPS/W@INT8 Hybrid Analog-Digital Multi-core SRAM CIM Macro with Optimal Weight Dividing and Resource Allocation Strategies

Yitong Zhou, Wente Yi, Sifan Sun, Wenjia Wang,
Jinyu Bai, **He Zhang***, **Wang Kang***
School of Integrated Circuit Science and Engineering,
Beihang University

Outline

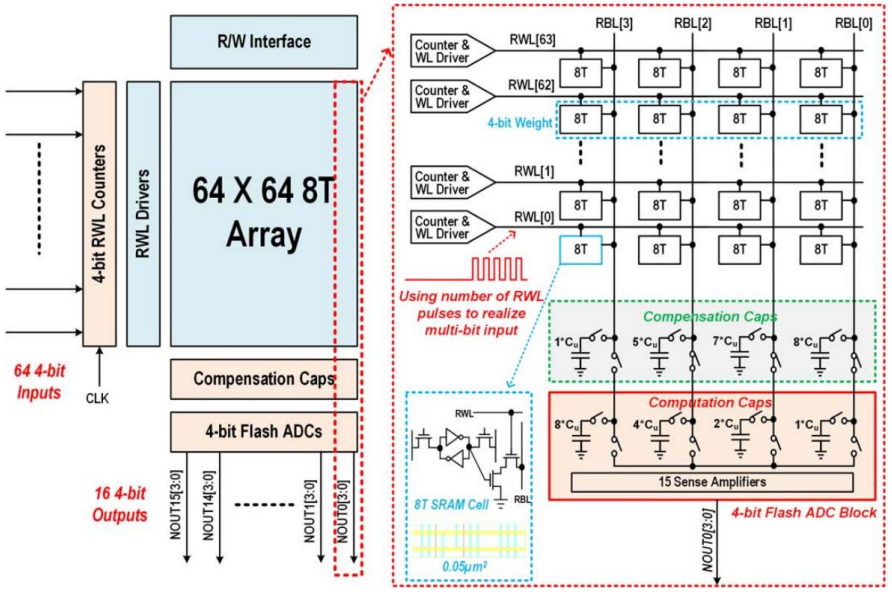
- Background
- Proposed Multi-core hybrid CIM architecture
  - Hybrid Weighting Scheme
  - Weight Divide Strategy & Computing Resource Allocation
- Experiment & Results
- Conclusion

Outline

- **Background**
- Proposed Multi-core hybrid CIM architecture
  - Hybrid Weighting Scheme
  - Weight Divide Strategy & Computing Resource Allocation
- Experiment & Results
- Conclusion

Background – Analog CIM

- **Analog CIM**: Use **physical quantities** to represent data and completes the **analog calculations** based on certain **physical laws**
  - **Current-domain** computing paradigm
  - **Charge-domain** computing paradigm
  - **Time-domain** computing paradigm

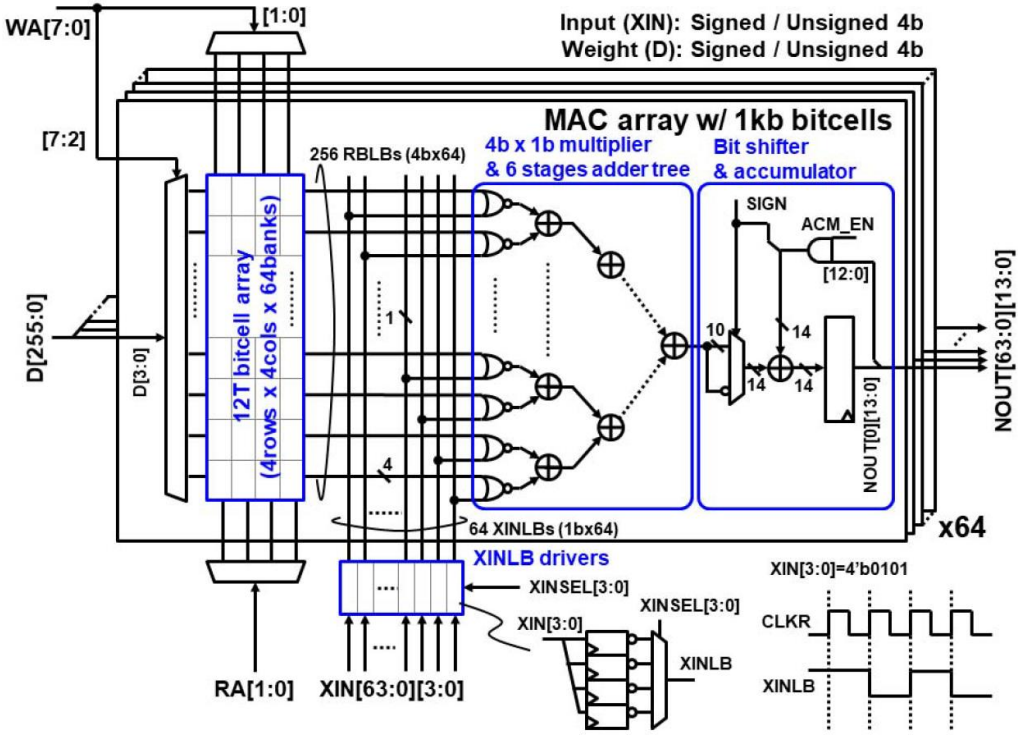


- Extremely high **energy efficiency** at **medium to low precision**
- Computational **errors**
- Substantial **area and power** overhead of peripherals at **medium to high precision**

A 351 TOPS/W and 372.4 GOPS Compute-in-Memory SRAM Macro in 7nm FinFET CMOS for Machine-Learning Applications (ISSCC 2020)

Background – Digital CIM

- **Digital CIM**: Integrate **logic gates** into memory cells to perform operations in **digital domain**



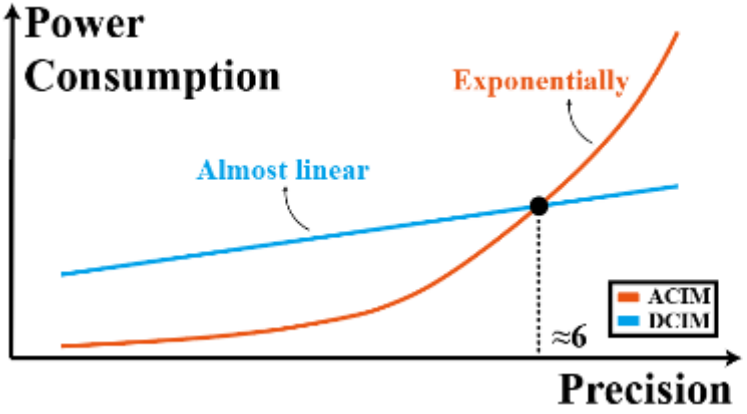
- **No** calculation loss
- Significant **area** overhead of peripherals at **medium to low precision**

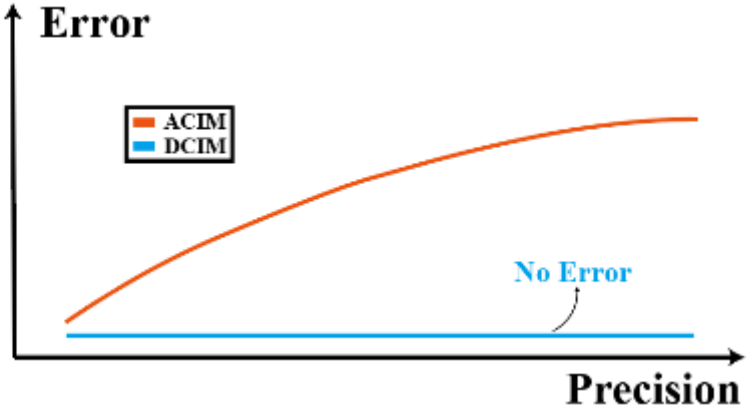
A 5-nm 254-TOPS/W 221-TOPS/mm Fully-Digital Computing-in-Memory Macro Supporting Wide-Range Dynamic-Voltage-Frequency Scaling and Simultaneous MAC and Write Operations (2022 ISSCC)

Background – Hybrid CIM

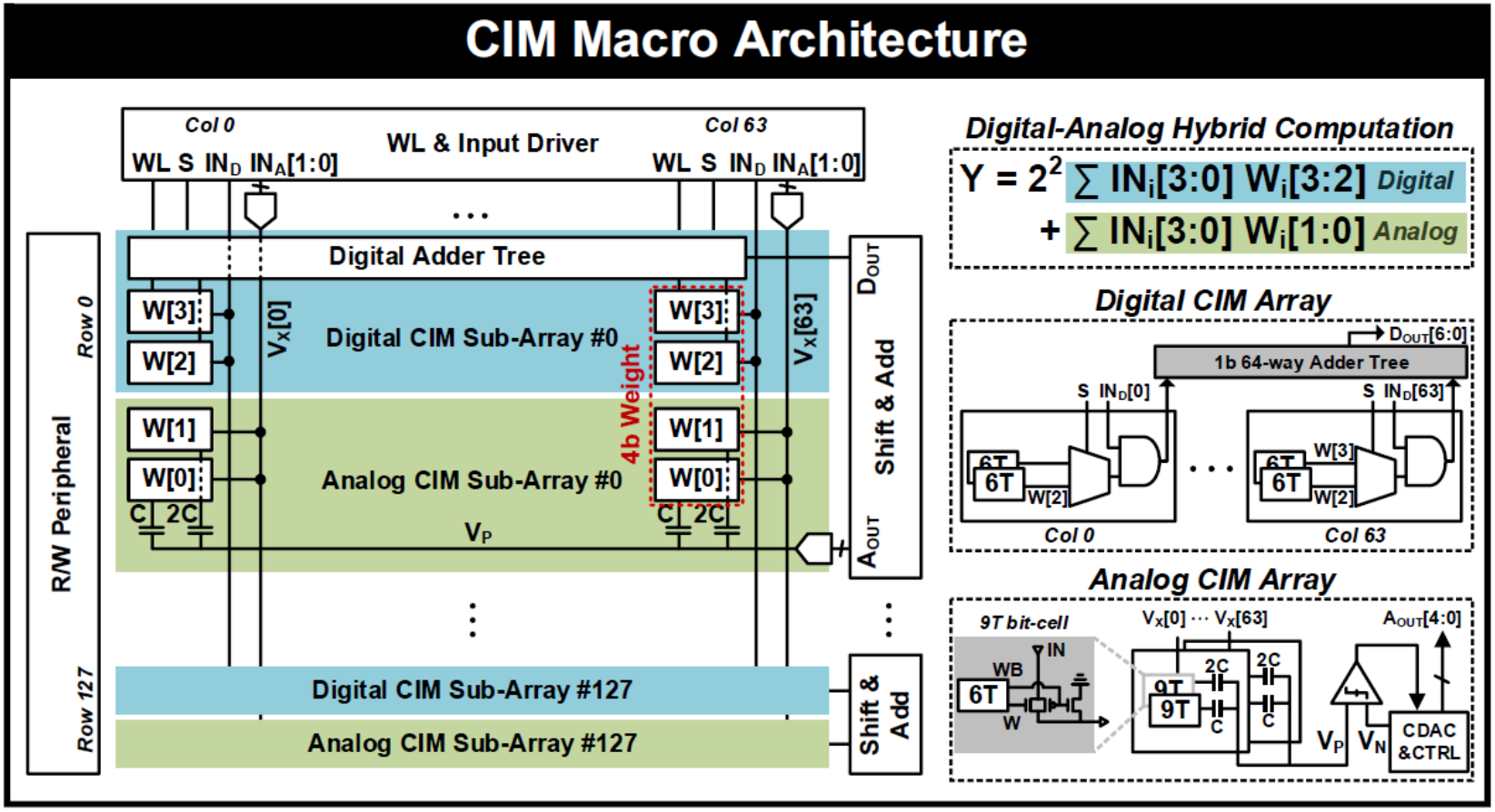
- **Performance comparison of ACIM and DCIM with different precision**

Category / Index	L2M Precision		M2H Precision	
	ACIM	DCIM	ACIM	DCIM
Energy Efficiency	✓	✗	✗	✓
Area Efficiency	✓	✗	✗	✓
Calculation Accuracy	✗	✓	✗	✓





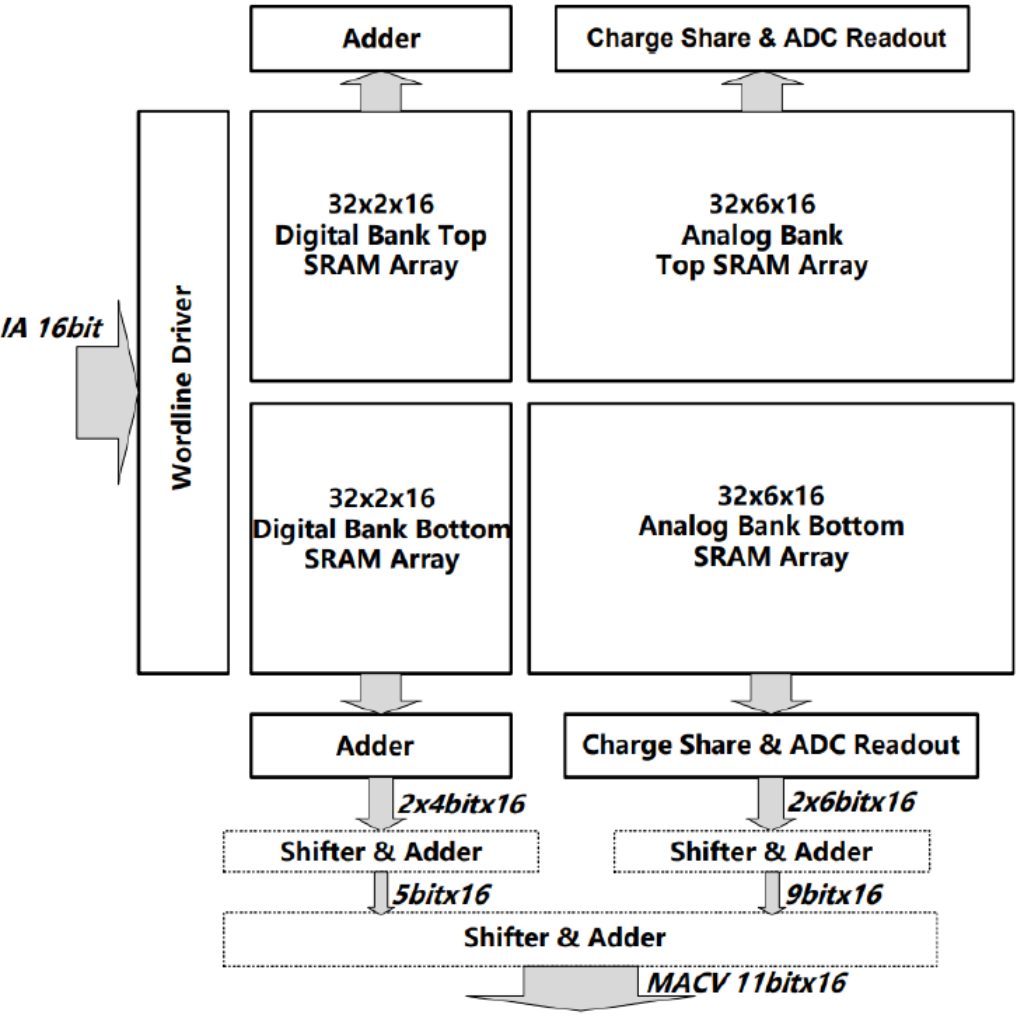
Background – Related Work



A 28nm 157TOPS/W 446.9Kb/mm2 Compute-In-Memory SRAM Macro with Analog-Digital Hybrid Computing for Deep Neural Network Inference

The optimal divide strategy for energy efficiency?

Background – Related Work



$$X * W = \sum_{i=0}^{7} 2^i x_i * \sum_{j=0}^{7} 2^j w_j$$

$$= \sum_{i=0}^{7} 2^i [2^6 * (2w_7 x_i + w_6 x_i) + 2^3 * (4w_5 x_i + 2w_4 x_i + w_3 x_i) + (4w_2 x_i + 2w_1 x_i + w_0 x_i)]$$

The optimal divide strategy for energy efficiency?

A Charge-Digital Hybrid Compute-In-Memory Macro with full precision 8-bit Multiply-Accumulation for Edge Computing Devices

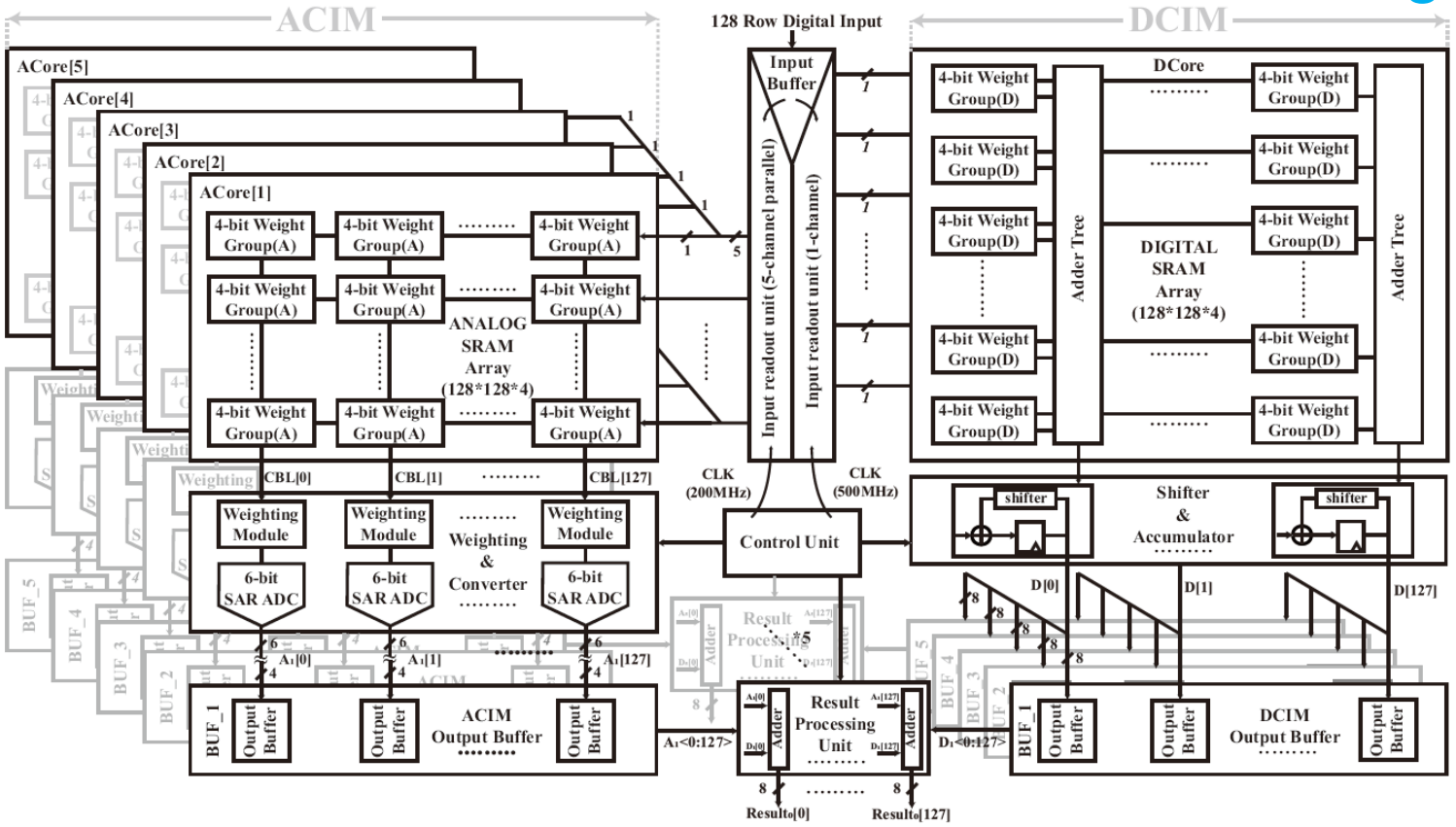
Outline

- Background
- **Proposed Multi-core hybrid CIM architecture**
  - **Hybrid Weighting Scheme**
  - Weight Divide Strategy & Computing Resource Allocation
- Experiment & Results
- Conclusion

Hybrid Weighting Scheme

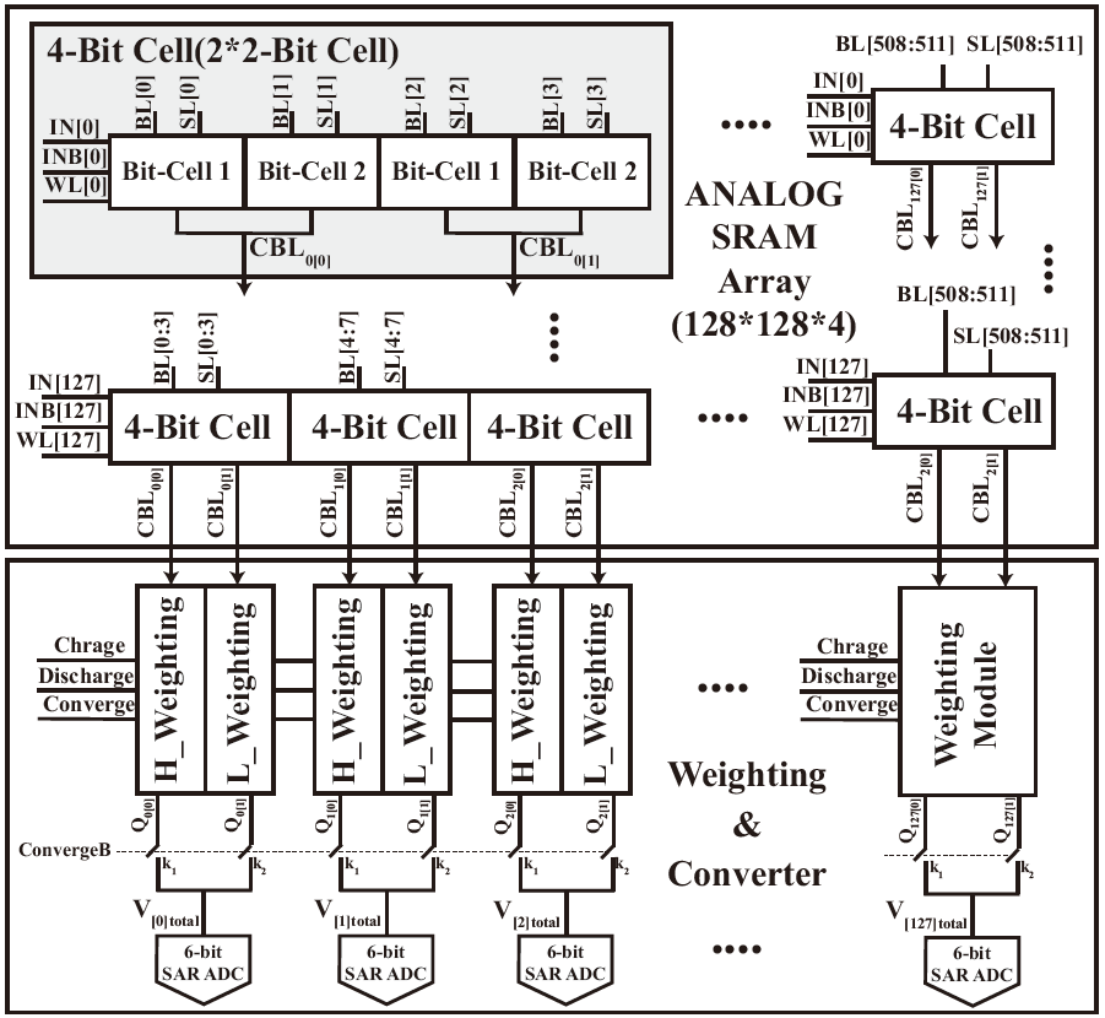
- **Overall Architecture**:
  - AC*5
  - DCore*1
  - Peripheral circuits

- Storage
- Input Handling
- MAC Operation
- Result Processing



Hybrid Weighting Scheme

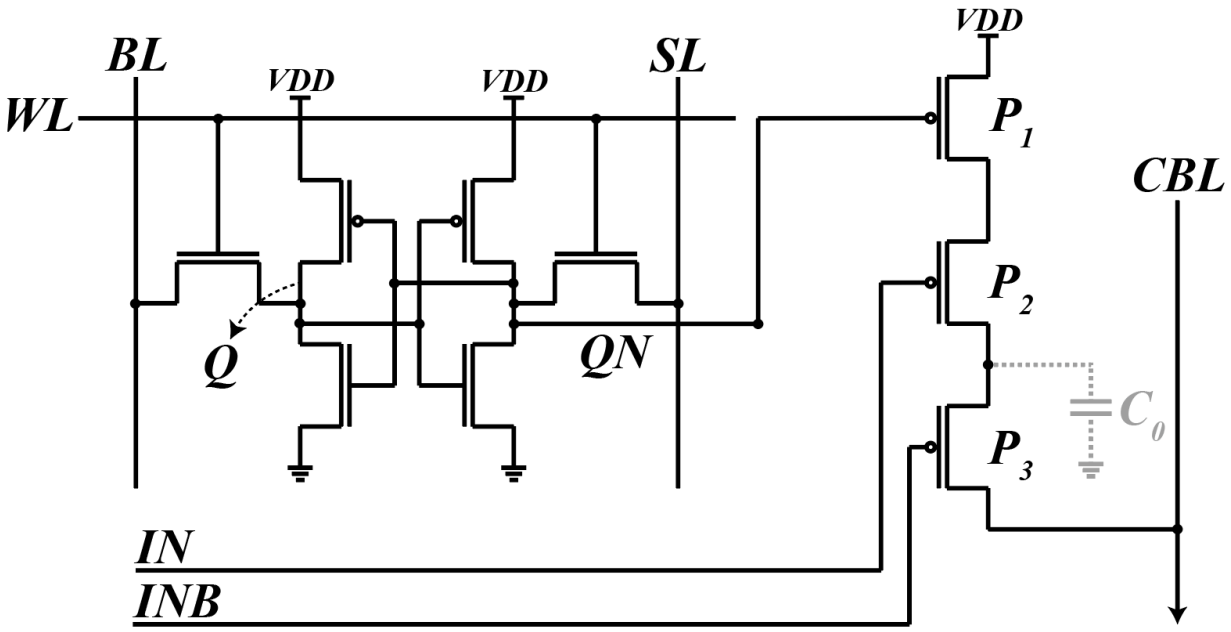
- **ACore Architecture**: 8-Bit input * 4 Bit weight



- **High precision** weighting scheme
- **Minimize power** consumption and **area** overhead

Hybrid Weighting Scheme

- **Bit-Cell 1 Structure**: integrates a 6T SRAM, and control transistors ($P1$, $P2$, $P3$)
  - **Charge domain** calculation reduces power consumption



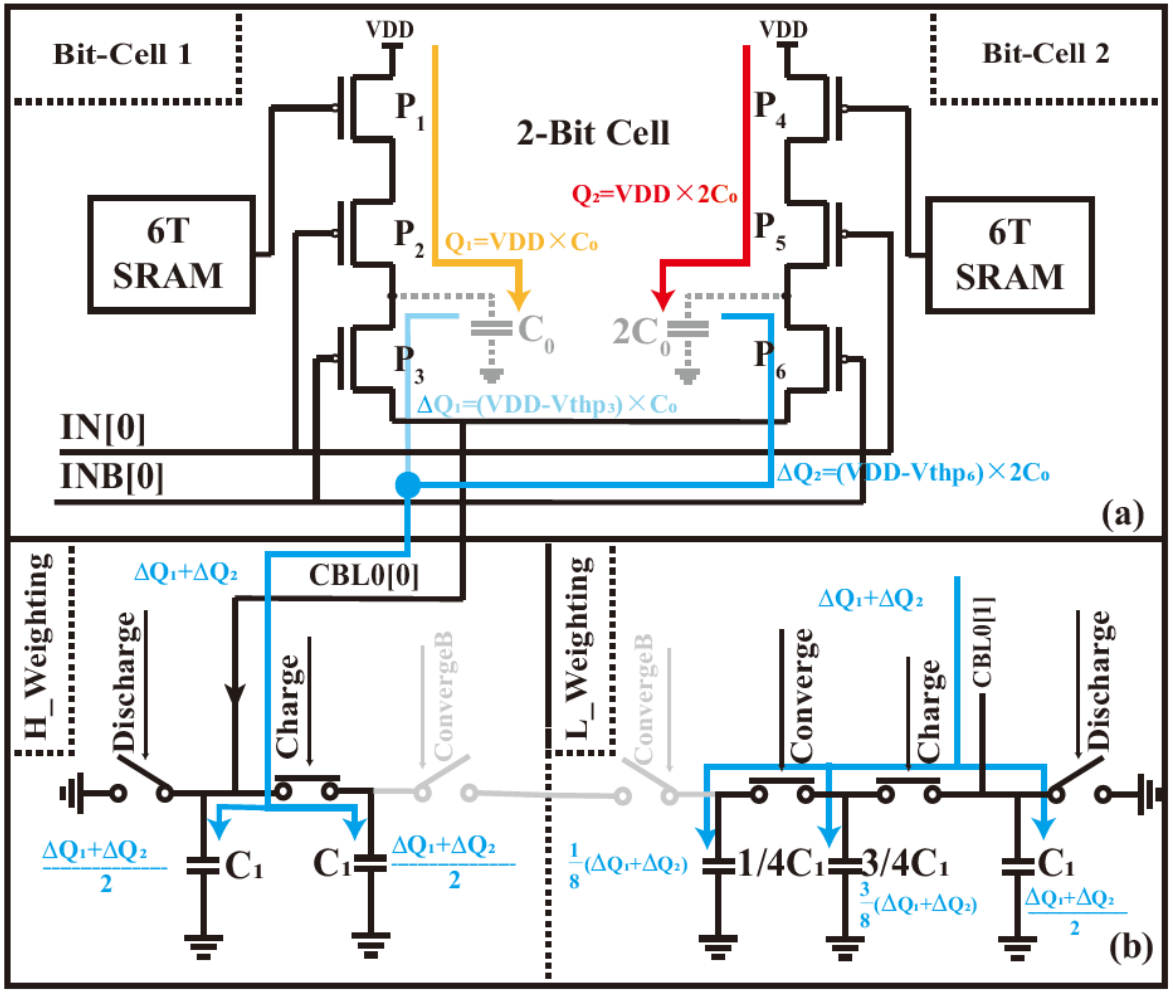
C0: **Parasitic** capacitance

IN & INB: **complementary** input signals

$$\Delta Q = (VDD - V_{THP_3}) \times C_0$$

Hybrid Weighting Scheme

- **2-Bit Cell structure & Weighting Module structure**

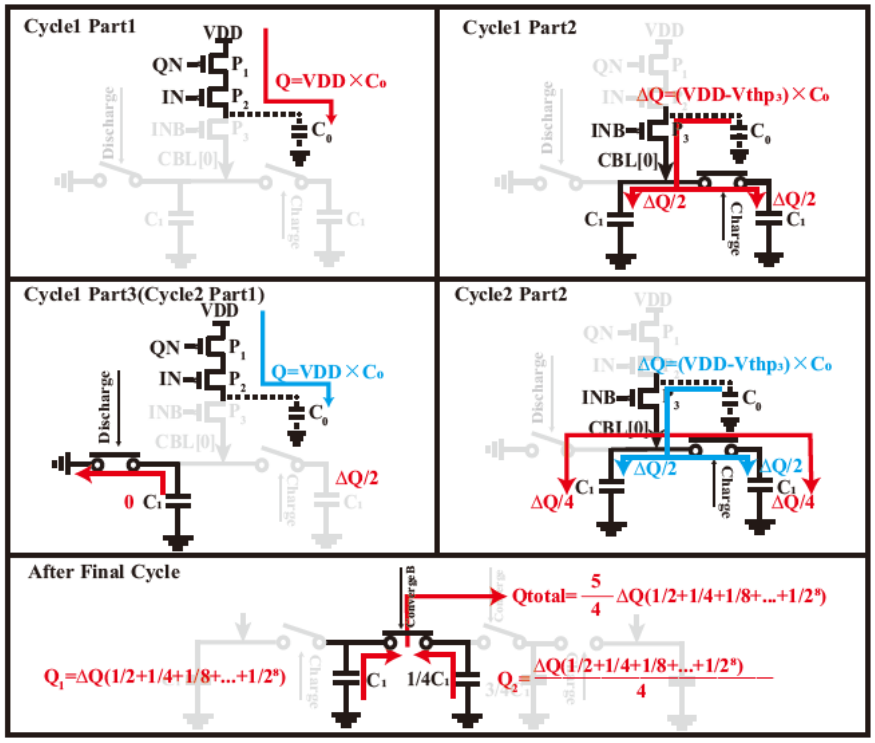


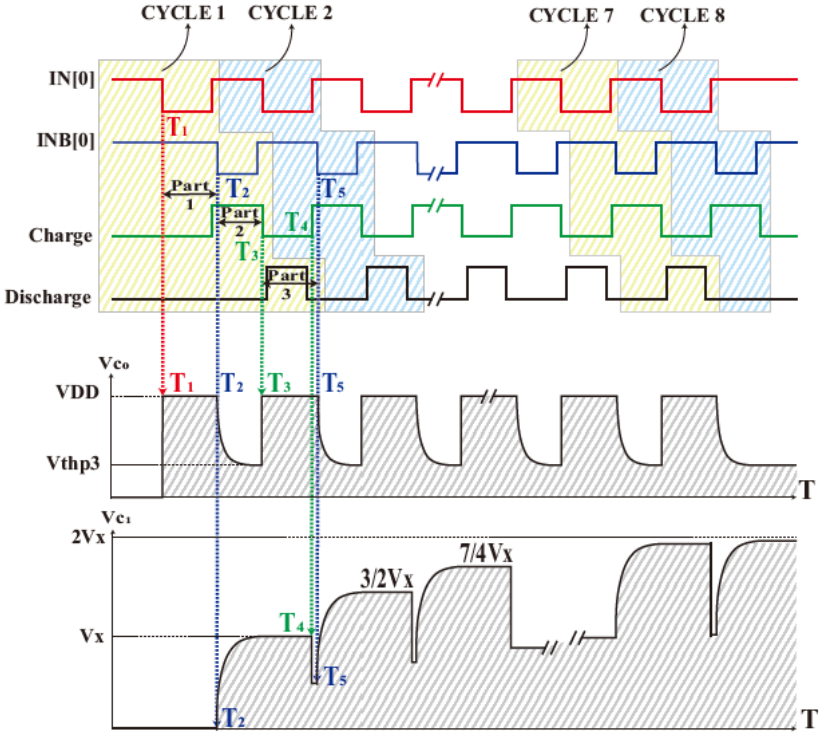
- **<u>Multi-cycle</u>** capacitor weighting module to **reduce area overhead**

Hybrid Weighting Scheme

- **Calculation process**: using the example of input '11111111' and weight '1111'

$$Q_{total} = \Delta Q \cdot \sum_{n=1}^{8} 1/2^n + \frac{\Delta Q \cdot \sum_{n=1}^{8} 1/2^n}{4} = \frac{5}{4}\Delta Q \cdot \sum_{n=1}^{8} 1/2^n$$





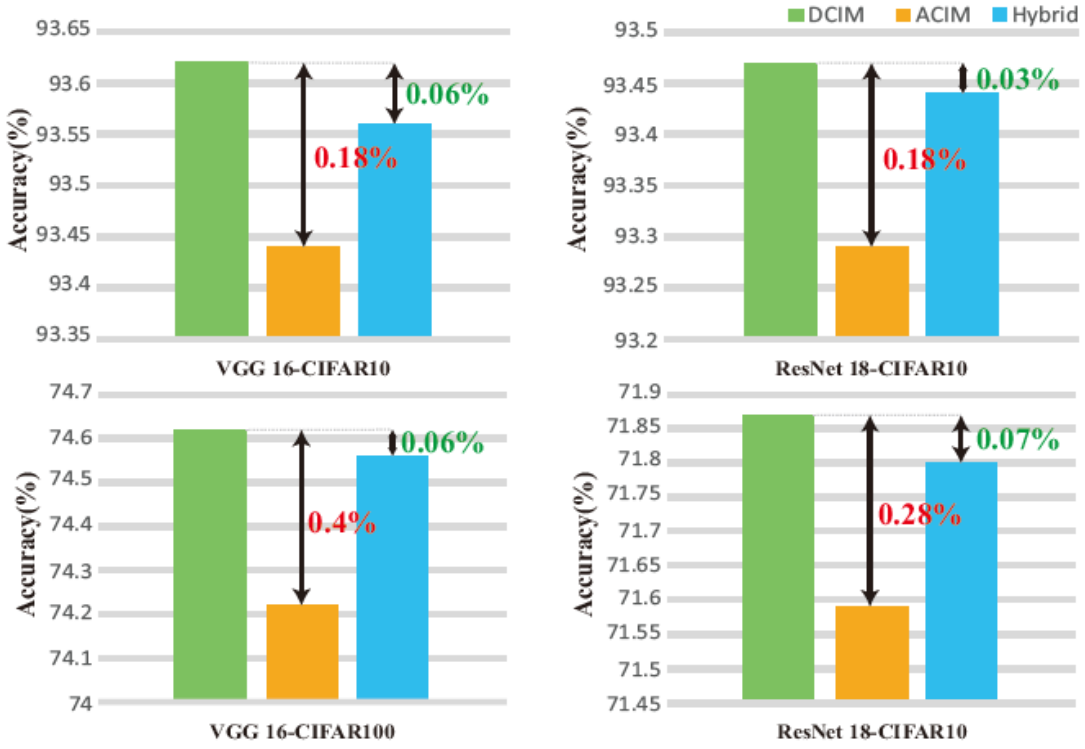
Outline

- Background
- **Proposed Multi-core hybrid CIM architecture**
  - Hybrid Weighting Scheme
  - **Weight Divide Strategy & Computing Resource Allocation**
- Experiment & Results
- Conclusion

Weight Divide Strategy & Computing Resource Allocation

- **Performance comparison**

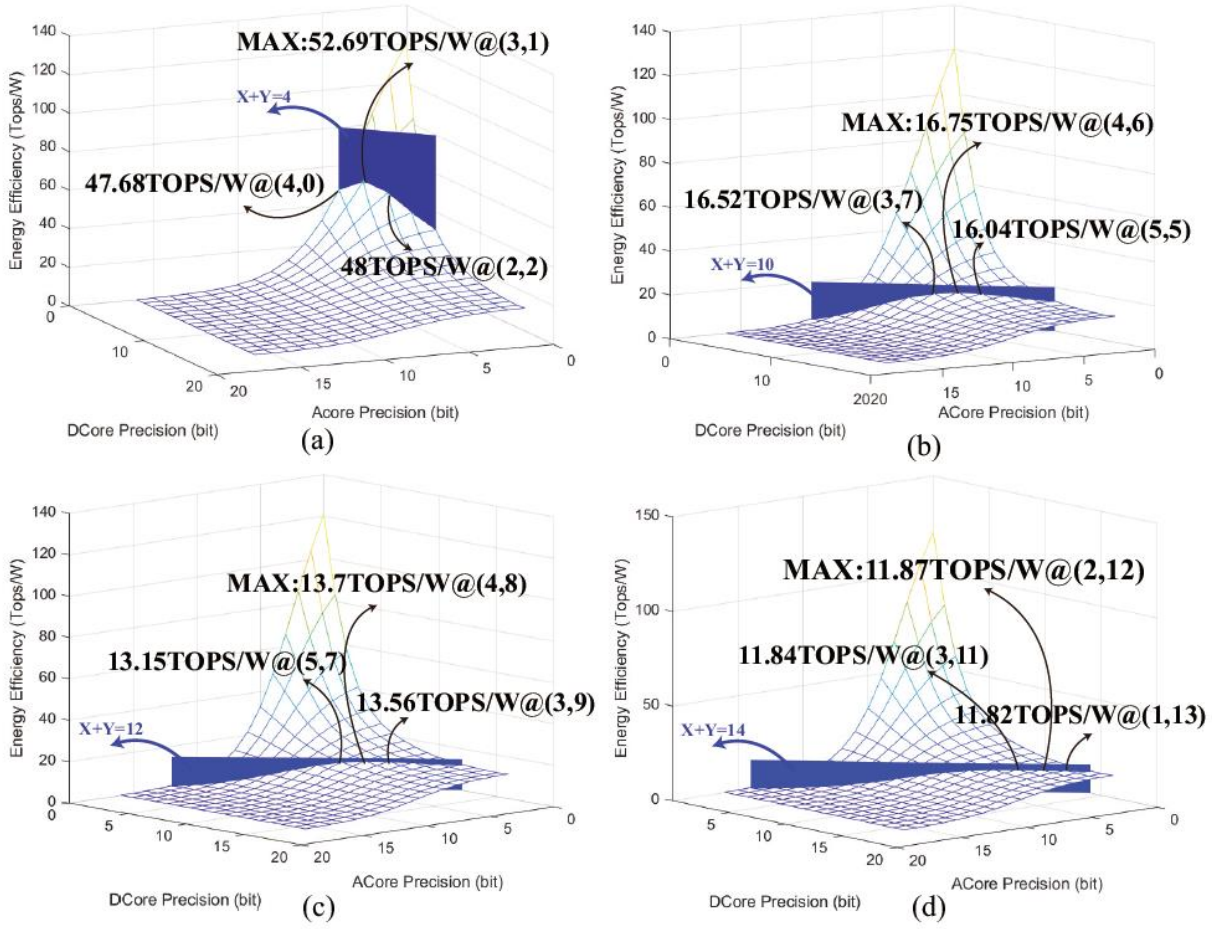
Divide Method	Energy Efficiency (TOPS/W)	Error
4+4	24.65	0.4%
2+6	19.73	0%



- The **4+4** divide strategy has better **energy efficiency** and **tolerable error**
- Hybrid CIM achieves **higher accuracy** than ACIM and is **comparable** to DCIM

Weight Divide Strategy & Computing Resource Allocation

- **Optimal divide strategies under different precision**



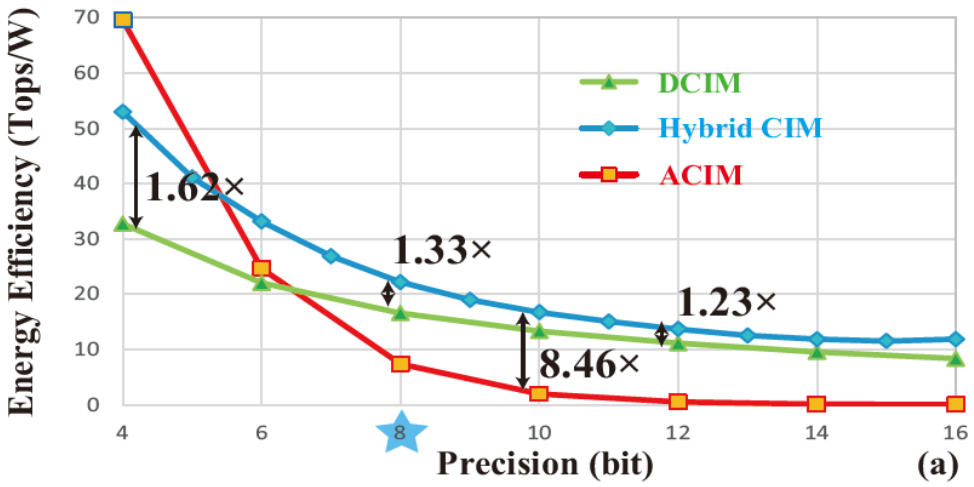
- Applicable to **various levels** of computational **precision**
- **Each precision** corresponds to an **optimal** energy efficient divide strategy

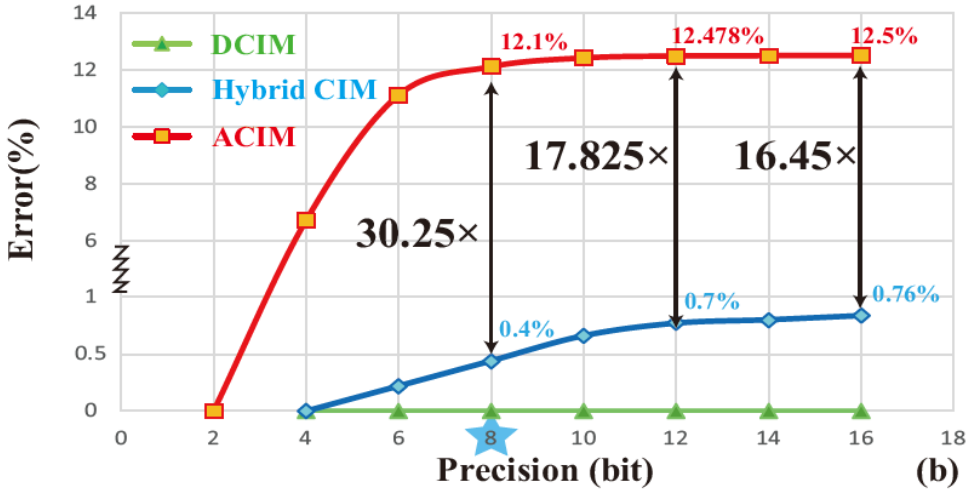
Outline

- Background
- Proposed Multi-core hybrid CIM architecture
  - Hybrid Weighting Scheme
  - Weight Divide Strategy & Computing Resource Allocation
- **Experiment & Results**
- Conclusion

Experiment & Results

- **Energy Efficiency & Error comparison of hybrid CIM with ACIM and DCIM**





Energy Efficiency

	4-Bit	8-Bit	12-Bit
DCIM	1	1	1
Hybrid CIM	1.62	1.33	1.23

Error

	8-Bit	12-Bit	16-Bit
ACIM	30.25	17.825	16.45
Hybrid CIM	1	1	1

Outline

- Background
- Proposed Multi-core hybrid CIM architecture
  - Hybrid Weighting Scheme
  - Hybrid Divide Strategy & Computing Resource Allocation
- Experiment & Results
- **Conclusion**

Conclusion

- **Propose a multi-core analog-digital hybrid CIM macro.**
- **Achieves 24.65 TOPS/W at 8-bit precision**
- **Compared to DCIM: 1.33× higher energy efficiency**
- **Compared to ACIM: 30.25× lower error**

Thanks !