DCiROM: A Fully Digital Compute-in-ROM Design Approach to High Energy Efficiency of DNN Inference at Task Level

Tianyi Yu, Tianyu Liao, Mufeng Zhou, Xiaotian Chu, Guodong Yin, Mingyen Lee, Yongpan Liu, Huazhong Yang, and **Xueqing Li**[1†]

[1]Tsinghua University, LFET/BNRist

†Email: xueqingli@tsinghua.edu.cn

Outline

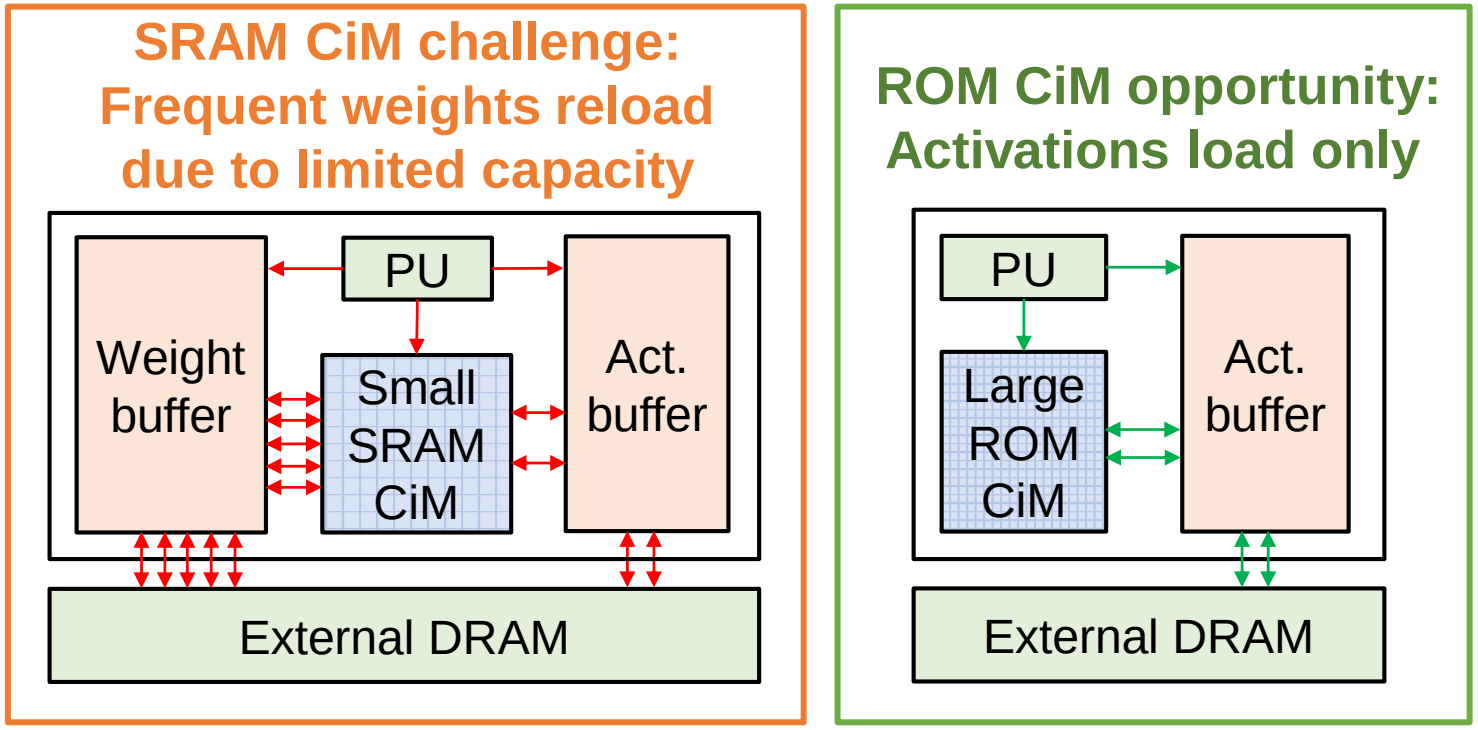
- Background
- Motivation
- Proposed Design
- Measurement
- Conclusion

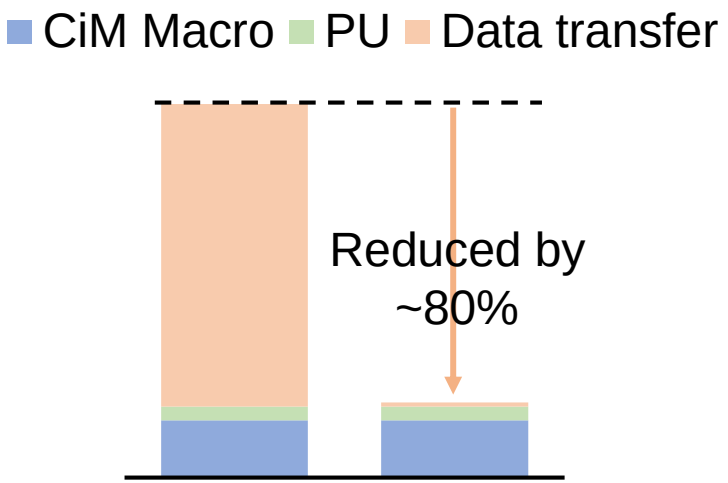
Outline

- **Background**
- Motivation
- Proposed Design
- Measurement
- Conclusion

Background

- SRAM CiM enhances performance in data-intensive AI tasks, but due to limited capacity, it suffers from frequent weights reload
- Recently, a high-density ROM CiM (G. Yin et al., 2023) has been proposed to address the limited capacity challenges of SRAM CiM

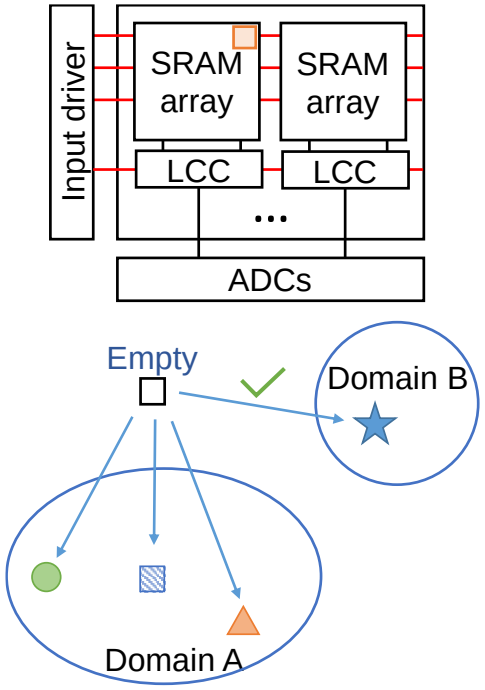




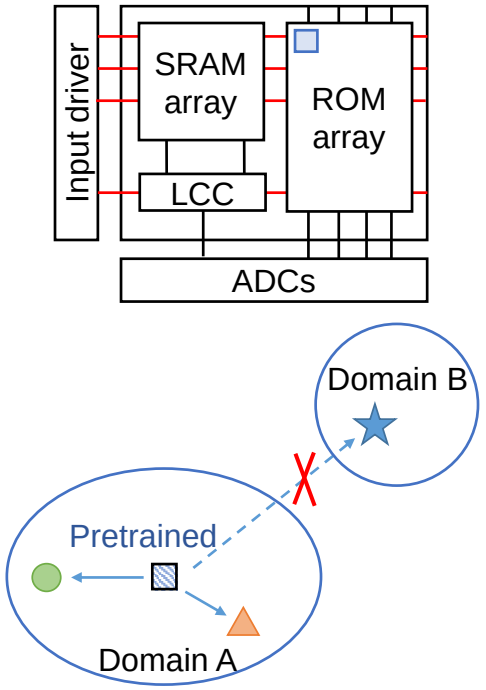
Task-level energy reduction of ROM-based CiM

Background

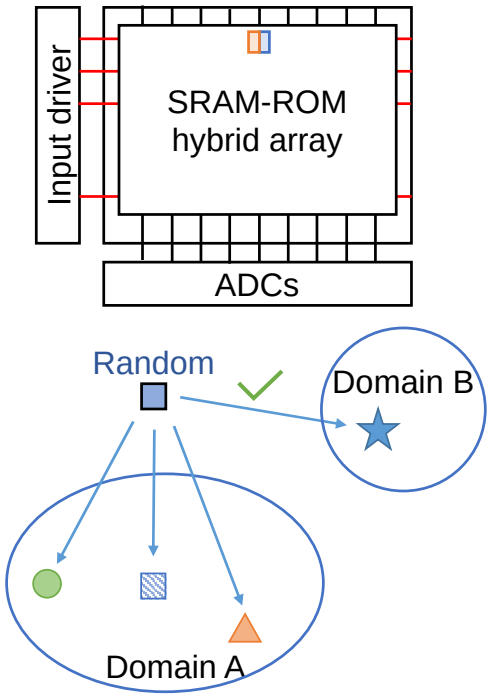
- By introducing SRAM CiM as finetuning weights, YOLoC and Hidden-ROM (Y. Chen et al., 2022) are proposed to release the bottleneck of flexibility issue



SRAM-CiM



YOLoC



Hidden-ROM

Source: Y. Chen *et al.*, *ICCAD'22*.

Outline

- Background
- **Motivation**
- Proposed Design
- Measurement
- Conclusion

Motivation

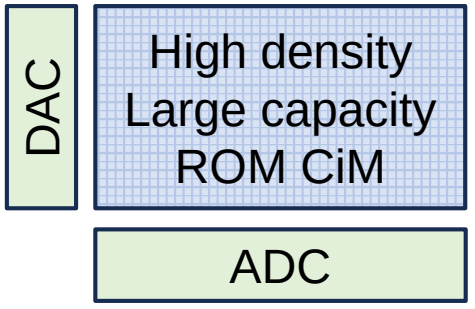
- Computing density of analog ROM CiM is limited by ADC
- Memory density of digital SRAM CiM is limited by adder tree

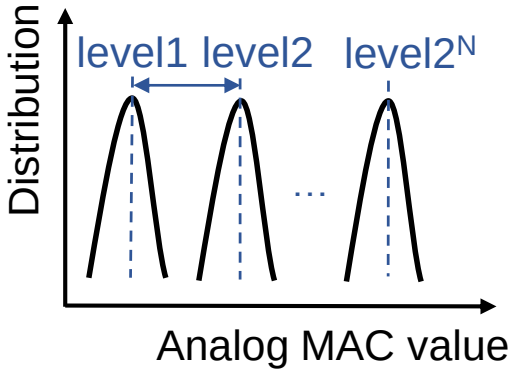
Existing high-memory-density ROM CiM

Memory density

Computing density

The number of activated rows (N) is limited by low SNR

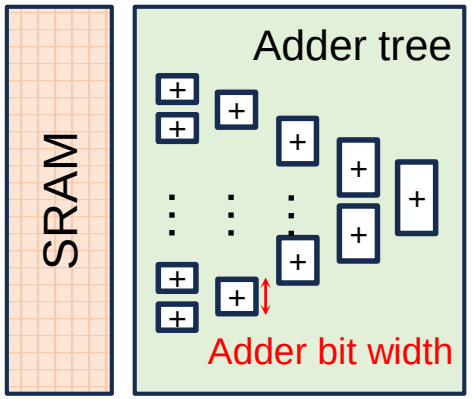


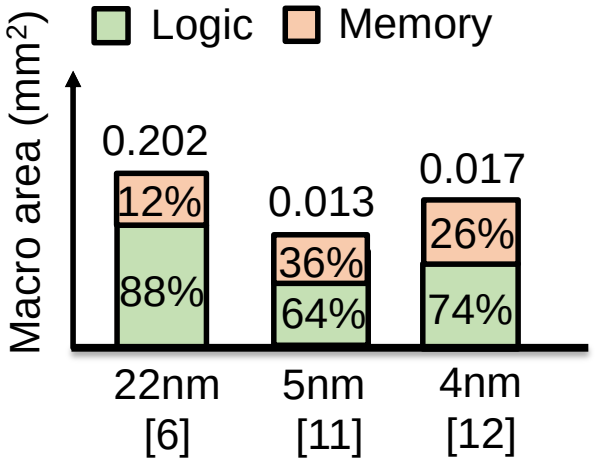


Existing high-computing-density SRAM CiM

Computing density

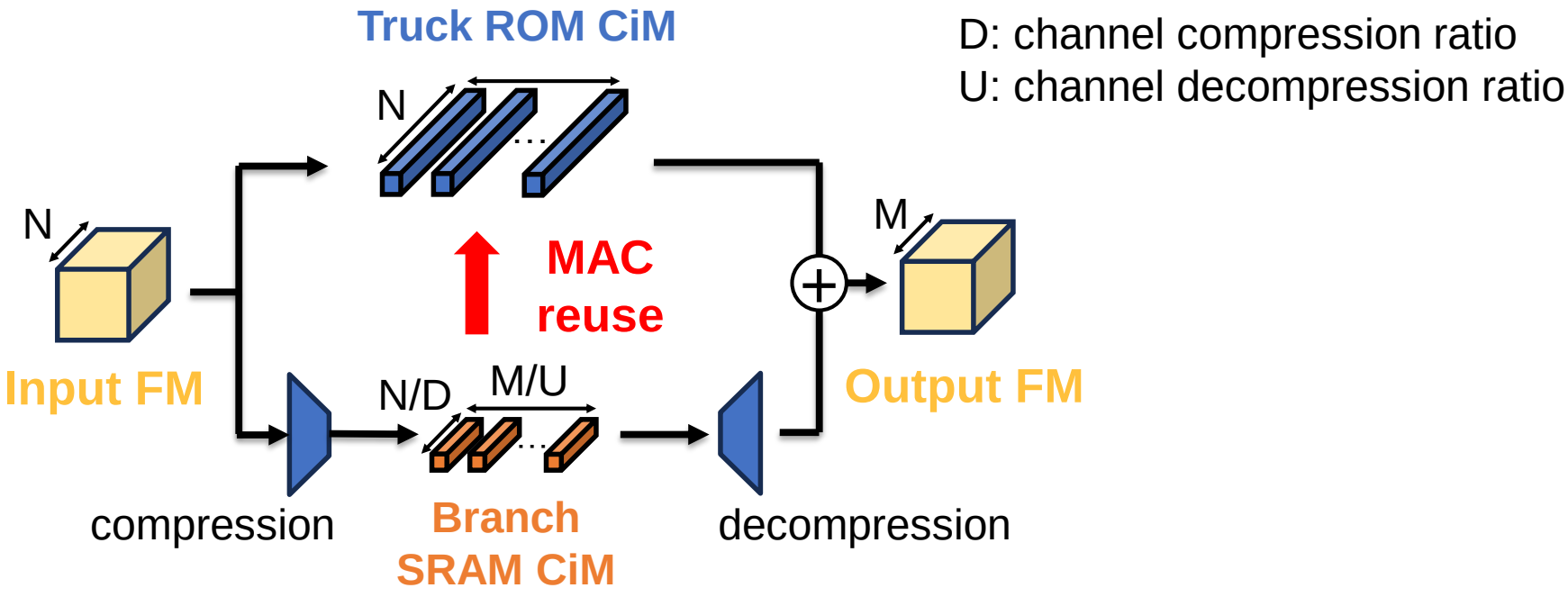
Memory density





Motivation

- YOLoC demonstrates a new concept of cutting off off-chip parameters loading with large-capacity ROM CiM and finetuning to various tasks
- Further area reduction by MAC reusing has not to be exploited



Contributions

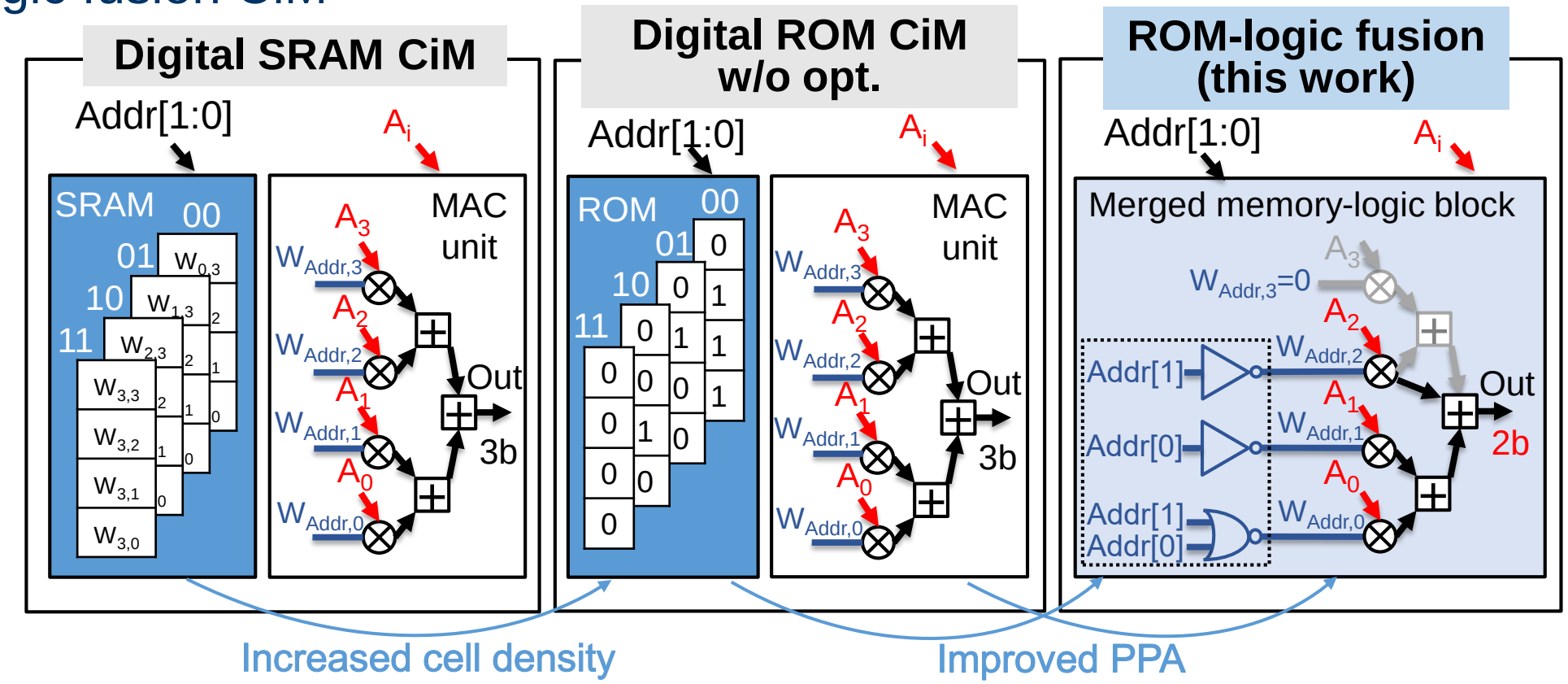
- How to further improve the density of **ROM CiM**?
- The **key contributions** of **this work *DCiROM***:
  - A ROM-logic fusion design approach that achieves both high memory density and high computing density by greatly simplified ROM circuit and adder tree
  - Two methods that achieve low extra circuit overhead flexibility of DCiROM on different datasets by ROM CiM resource reusing
  - A 65nm DCiROM chip that has built-in with all weights of ResNet-56, achieving experimentally measured ultra-high 2.06 TOPS/mm^2 computing density and 487 Kb/mm^2 memory density on end-to-end inference task

Outline

- Background
- Motivation
- **Proposed Design**
- Measurement
- Conclusion

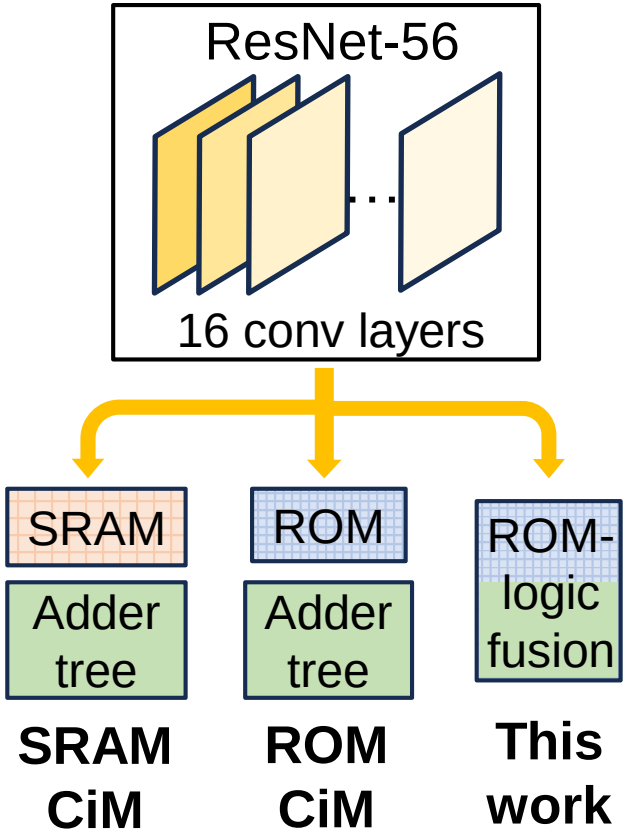
Proposed Design

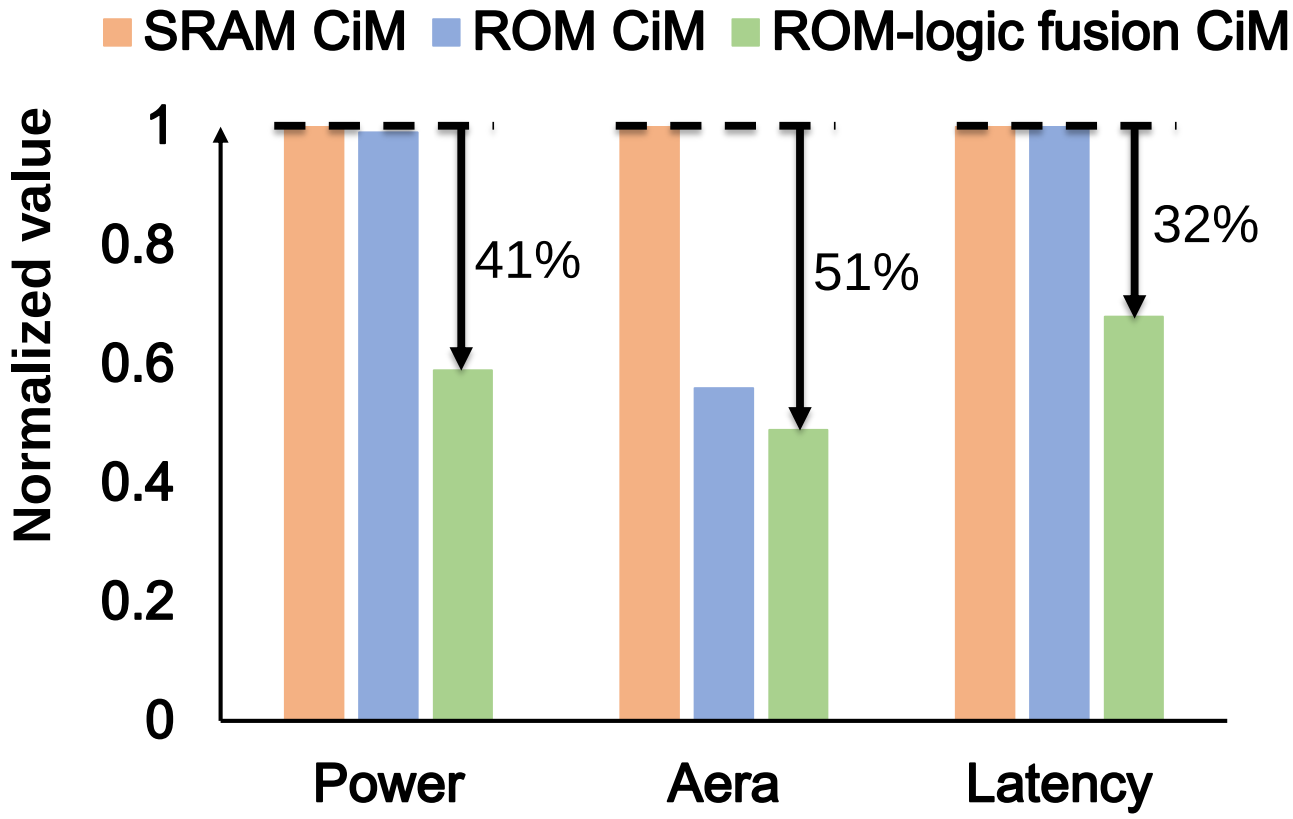
- Proposed DCiROM: A synthesizable **ROM-logic fusion** CiM design approach
  - Density enhancement process from general digital SRAM CiM to ROM-logic fusion CiM



Proposed Design

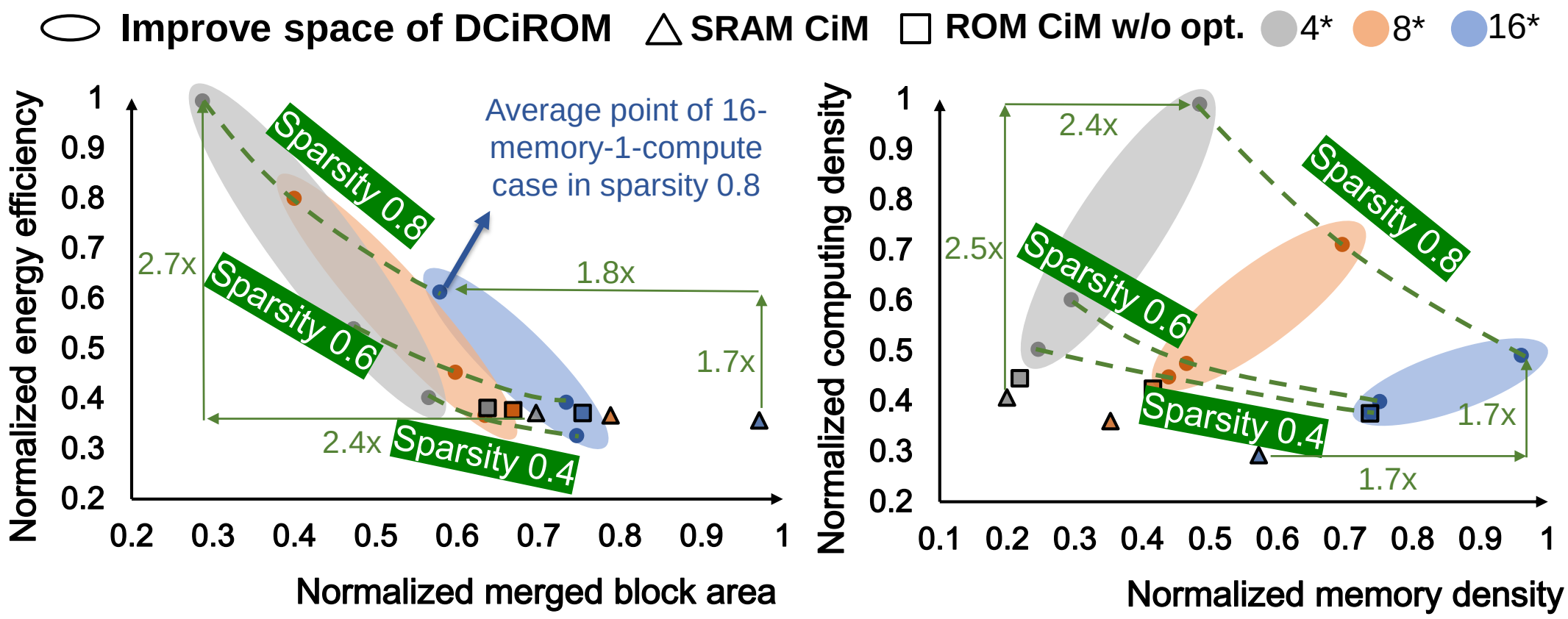
- Implementation of DCiROM on ResNet-56 convolutional layers
  - PPA comparison of SRAM CiM, ROM CiM and ROM-logic fusion CiM





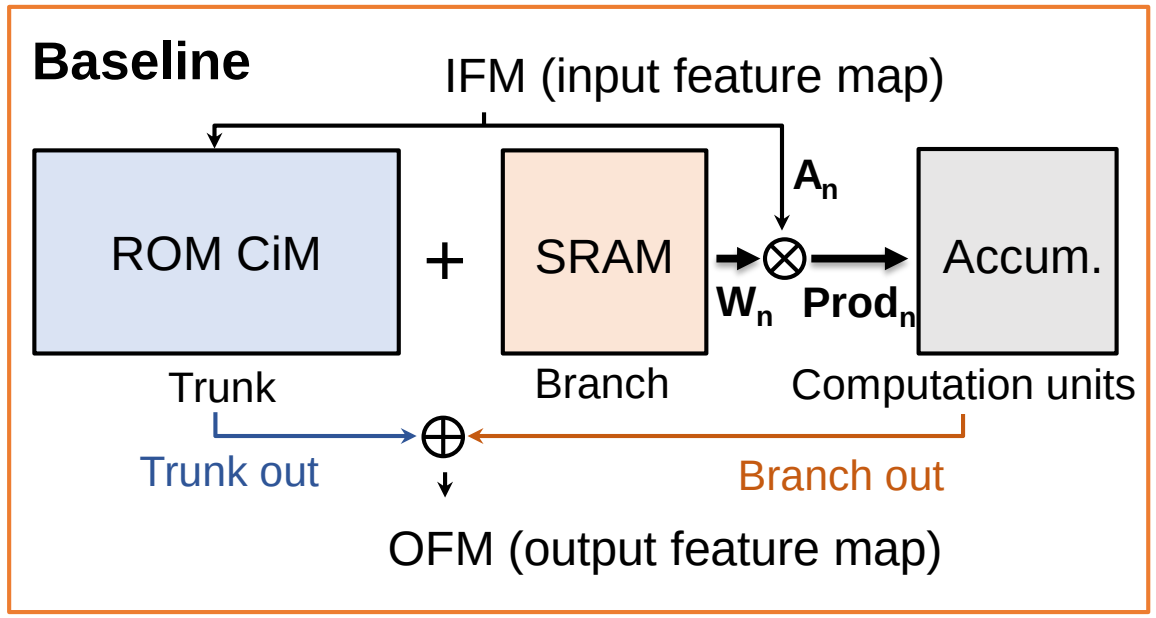
Proposed Design

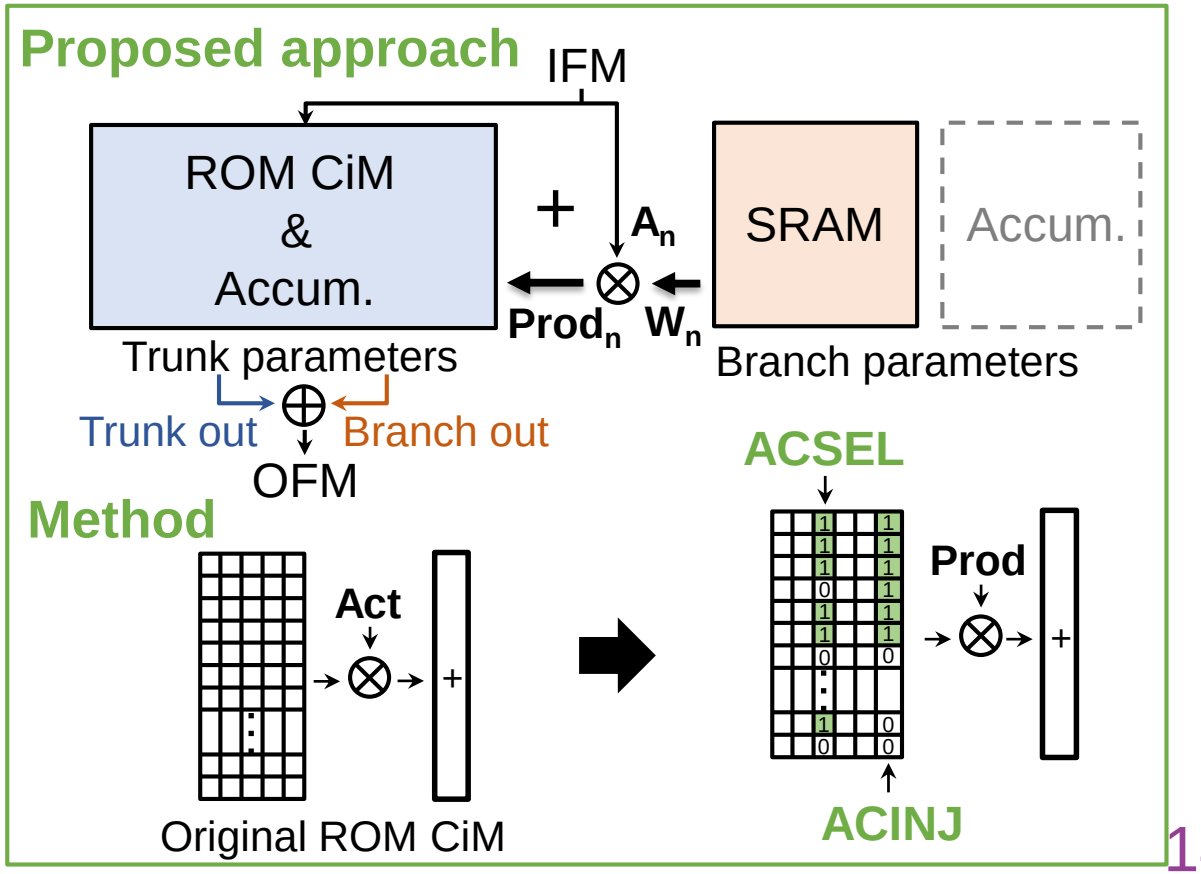
- PPA improvement space of ROM-logic fusion CiM
  - *__Memory-compute ratio__: Select 1 column from a 4/8/16-column memory block to perform MAC operation



Proposed Design

- Two methods of ROM CiM reusing: **ACSEL** and **ACINJ**
  - ACSEL: Select a column including sufficient 1's for accumulating
  - ACINJ: Add a redundant column including enough 1's to ROM

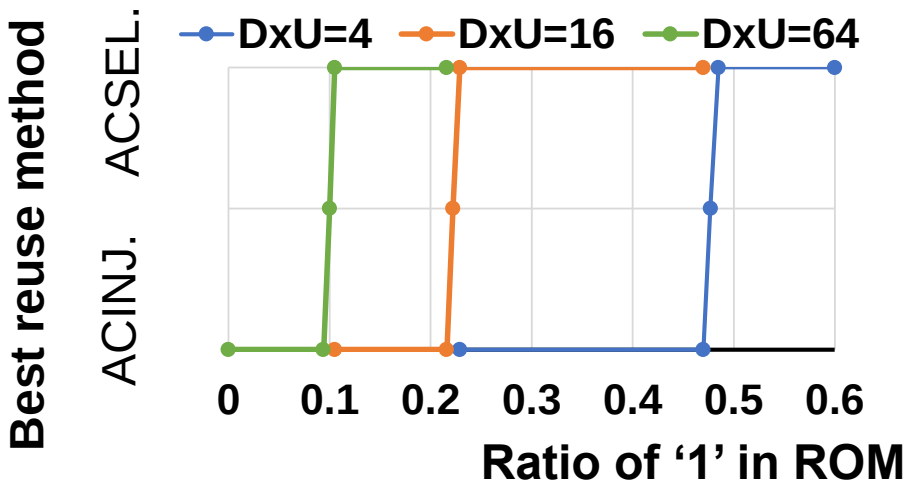


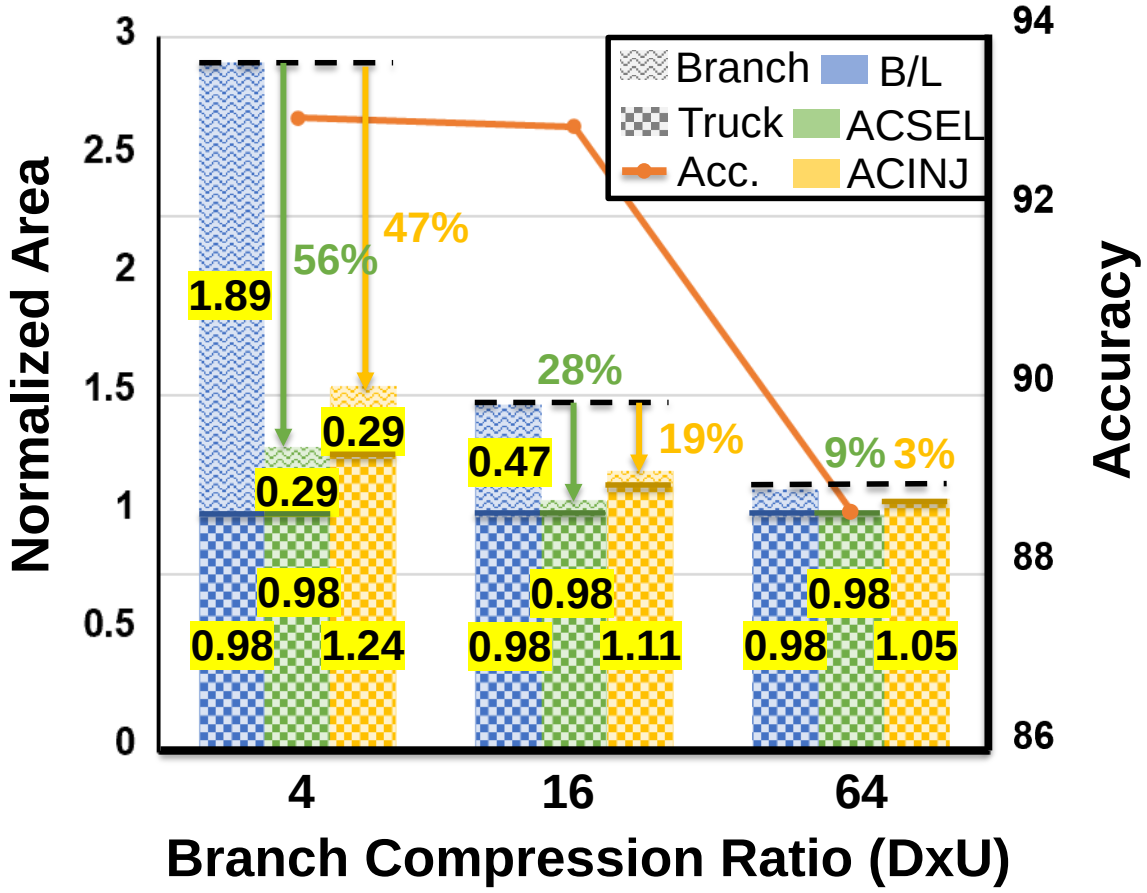


Proposed Design

- ROM CiM reuse evaluation: **Feasibility** and **area reduction**
  - ACSEL/ACINJ reduce 56%/47% area overhead at most

ResNet-18 conv2_x	
Convolution layers	4
Truck size	(64,3,3,64)
D/U	2~8
Branch size	(8,3,3,8)~(32,3,3,32)



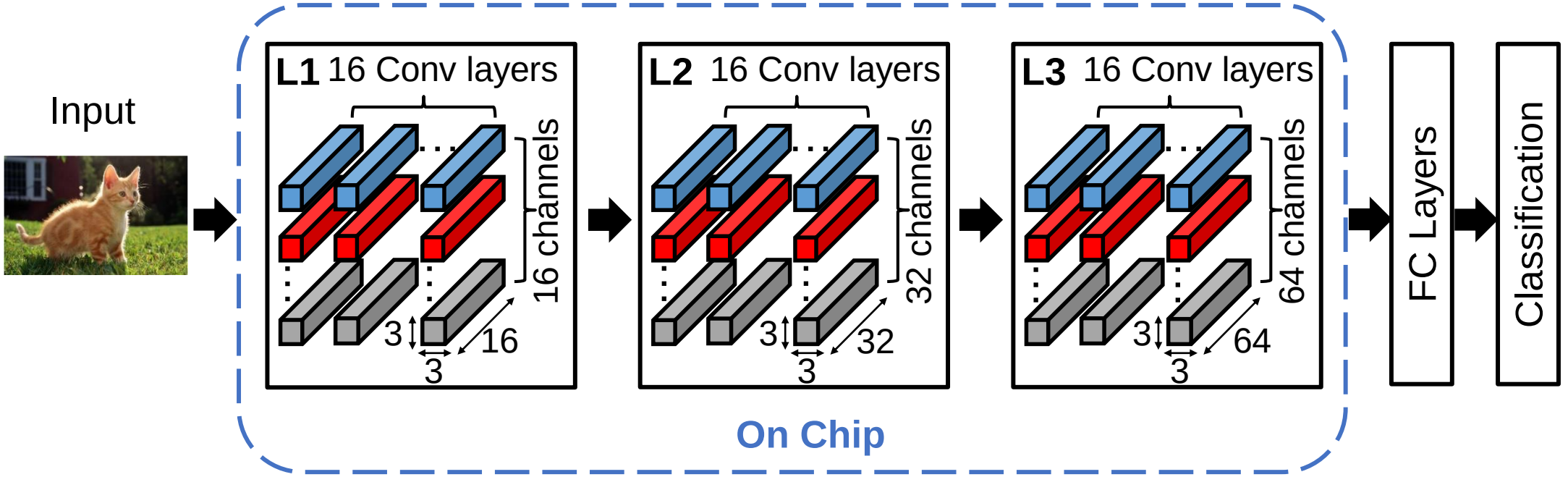


Outline

- Background
- Motivation
- Proposed Design
- **Measurement**
- Conclusion

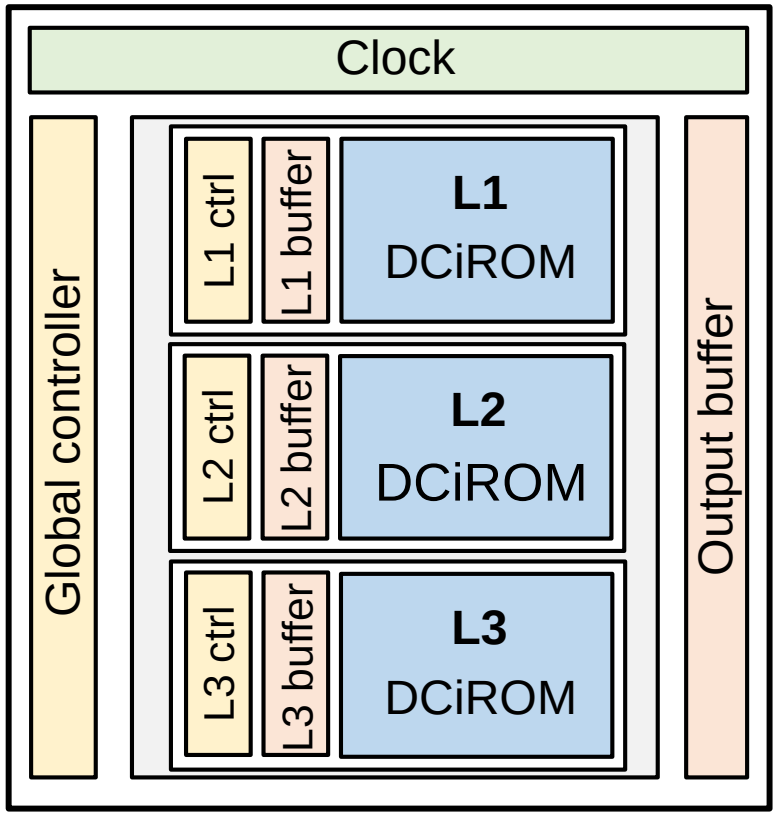
Measurement

- Structure of **ResNet-56**
  - The precision of input/weight is 8bit/4bit
  - All convolutional layers are mapped on chip

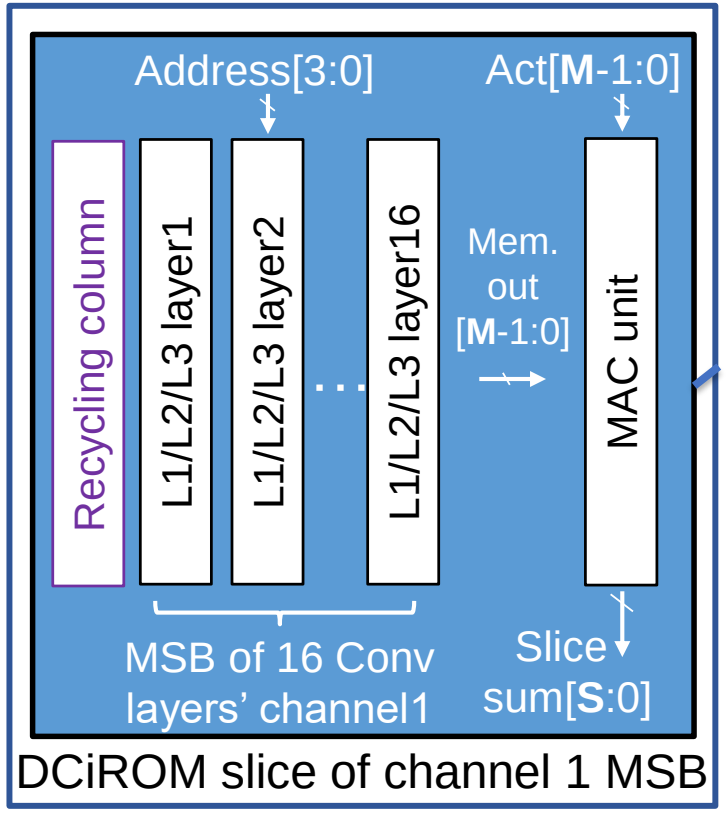


Measurement

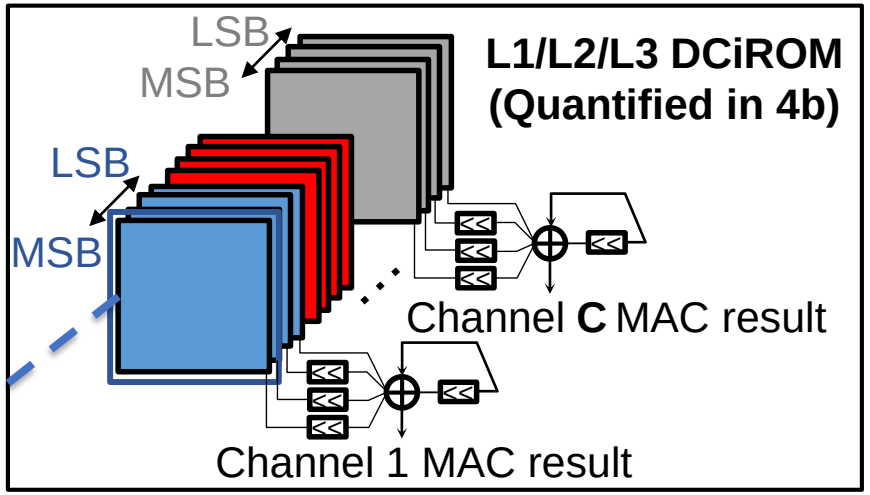
- DCiROM chip architecture
  - High efficiency ROM-logic fusion synthesis

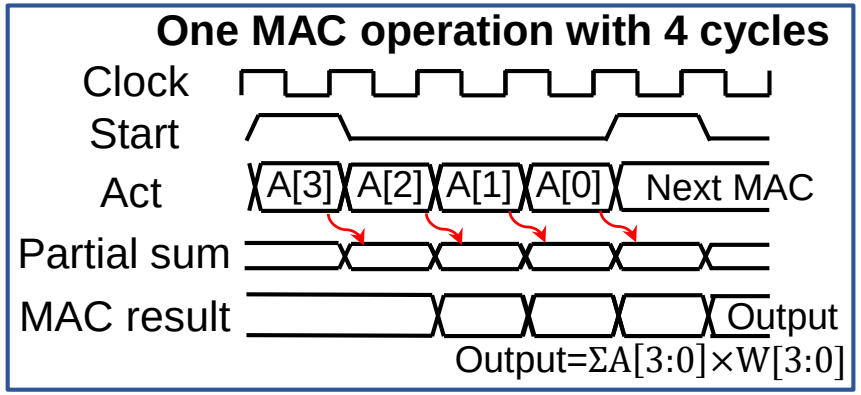


Chip architecture



DCiROM slice of channel 1 MSB

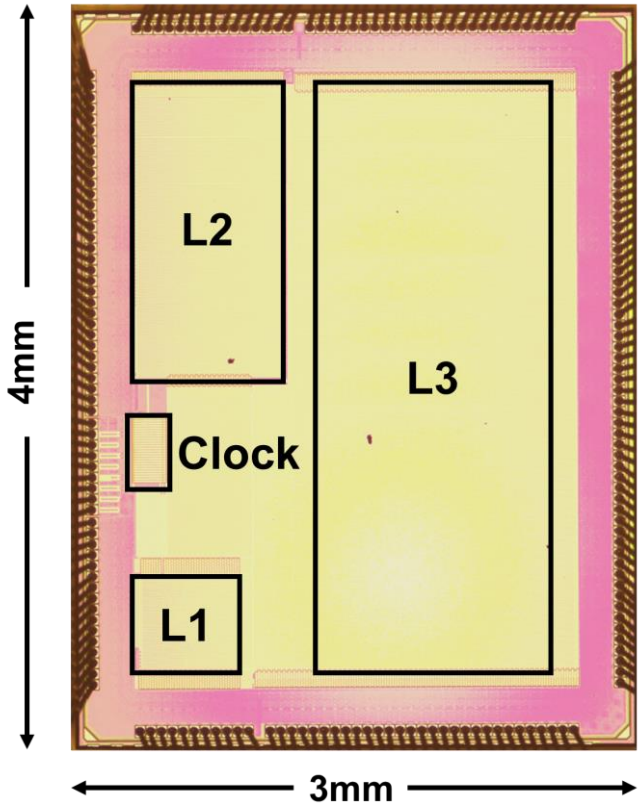




Measurement

Die micrograph and chip summary

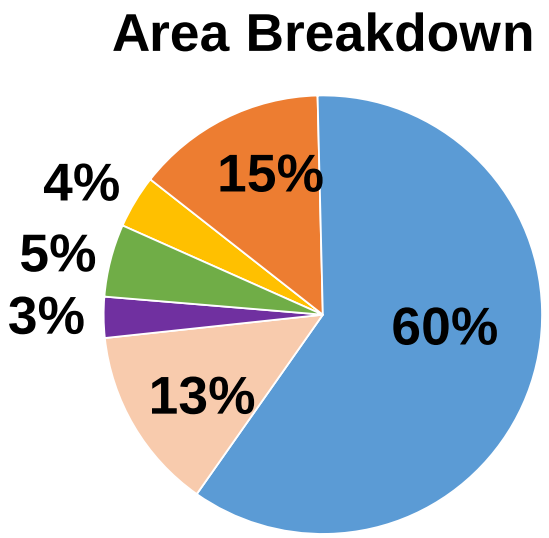
- FoM: Memory density x Computing density

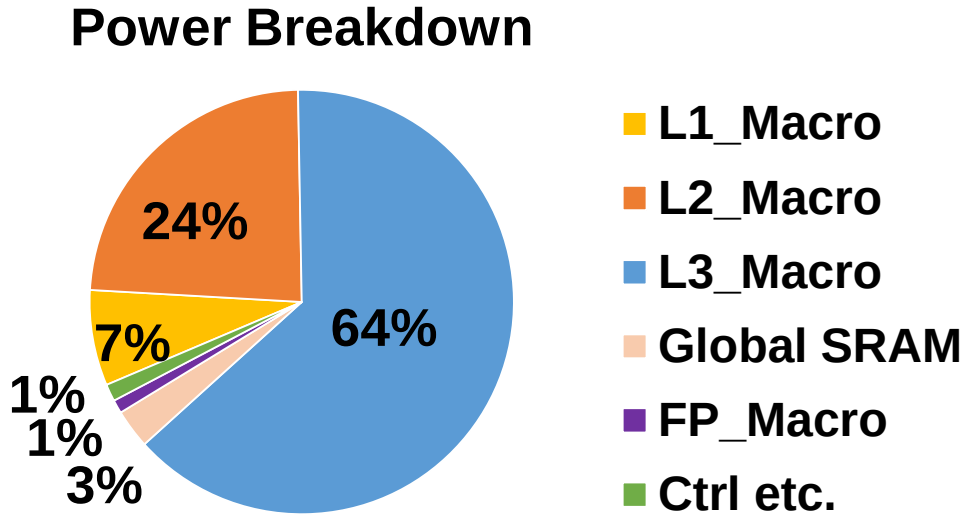


	L1		L2		L3	
Capacity (Kb)	144		576		2304	
Voltage (V)	0.6	1.2	0.6	1.2	0.6	1.2
Energy efficiency (TOPS/W)	34.8	8.6	35.6	9.0	39.2	9.1
Memory density (Kb/mm^2)	620		495		479	
Computing density (TOPS/mm^2)	0.85	3.3	0.57	2.4	0.51	1.6
FoM (TOPS/mm^2·Kb/mm^2)	527	2046	282	1188	244	766

Measurement

- **Area/Power breakdown** and **inference accuracy**
  - More than 80% area/power to compute
  - Less than 3% accuracy loss compared with FP32 baseline

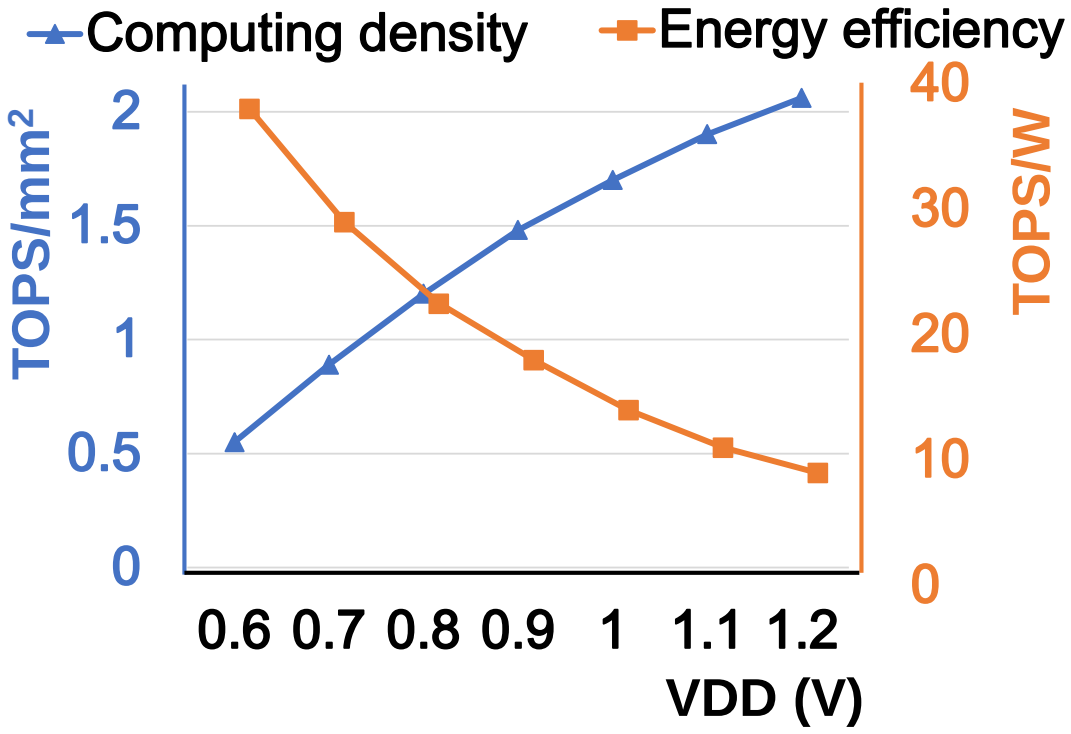


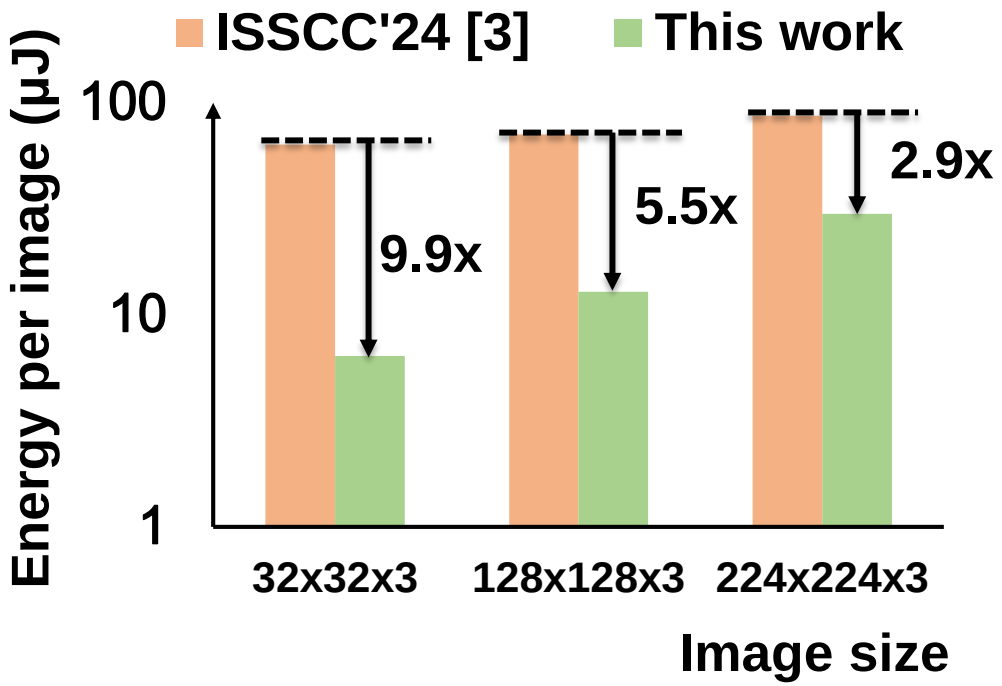


	CIFAR-10	MNIST	Fashion-MNIST
Baseline (FP32)	93.39%	99.64%	92.65%
This work (4b/8b)	91.23%	99.38%	90.10%

Measurement

- **Task-level** chip measurement results
  - Energy efficiency and computing density at different voltages
  - Inference energy consumption compared with SRAM CiM





Comparison with SOTA CiM works

- Achieving high computing density (> 30x than ROM CiM), high Memory density (> 4x than SRAM CiM) by ROM-logic fusion, and breaking through the SOTA works trade-off

	ISSCC'24 [3]	ISSCC'24 [4]	CICC'24 [5]	This work		JSSC'23 [10]	CICC'24 [7]
Technology	65nm	28nm	22nm	**65nm**		28nm	28nm
CiM operation	Analog SRAM	Hybrid SRAM	Digital SRAM	**Digital ROM**		Analog eDRAM	Analog ROM
Capacity (Kb)	80	192	128	**3024**		9600	22528
Voltage (V)	0.6~1.1	0.7~0.95	0.6~0.8	**0.6**	**1.2**	1.1	0.7~1.1
Energy efficiency (TOPS/W)*	255	16.9~37.6	4.2~11.1	**38.0**	**9.0**	21.49	14.9~31.2
Memory density (Kb/mm²)*	31	18	115	**487**		181	1656
Computing density (TOPS/mm²)*	0.78	1.02	0.21~0.28	**0.55**	**2.06**	0.19	0.030~0.059
FoM (TOPS/mm²·Kb/mm²)	24	1085	3629~5000	**268**	**1004**	1975	2856~5713
FoM (TOPS/mm²·Kb/mm²)*	24	18	24~32	**268**	**1004**	34	49~98

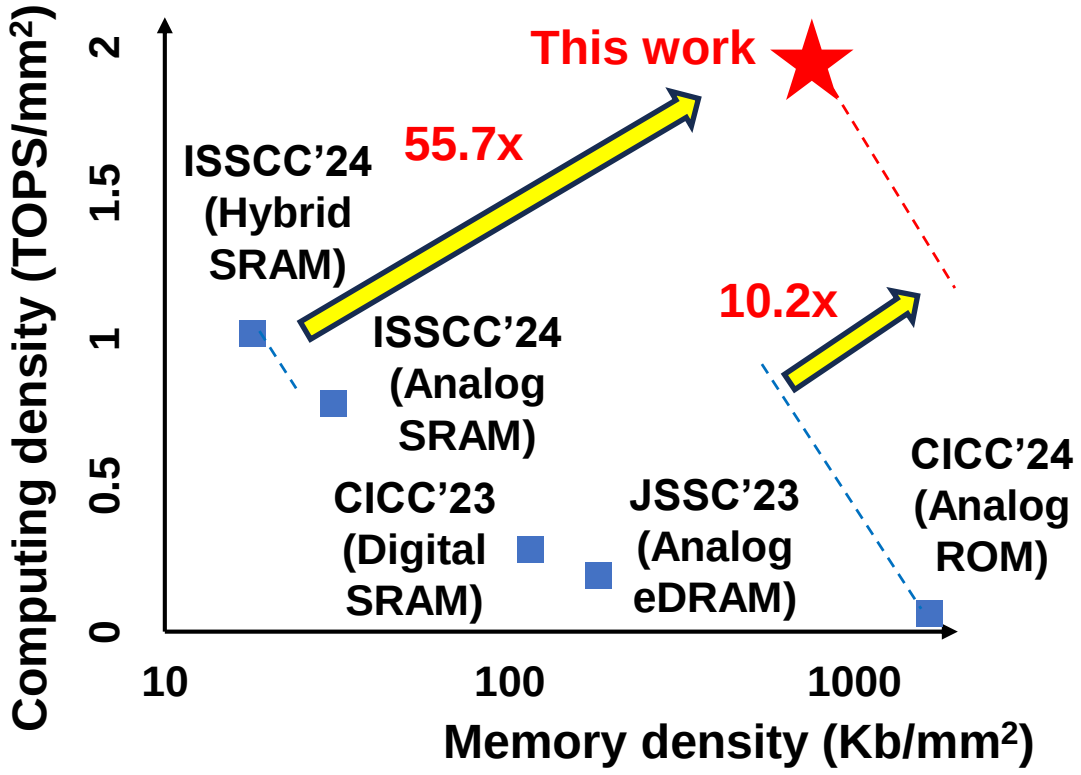
*Normalized to 65nm

Outline

- Background
- Motivation
- Proposed Design
- Measurement
- **Conclusion**

Conclusion

- Highlight of DCiROM
  - Ultra-high density FoM (computing density x memory density)



*Normalized to 65nm.

Conclusion

- Proposed DCiROM design approach
  - High memory density and high computing density
  - Less area overhead to realize flexibility
- Features:
  - A fully digital ROM with local computing units supports 10.2x-55.7x density FoM of SOTA CiM works
  - Task evaluation shows 9.9x system-level energy efficiency improvement over SRAM CiM
  - Reduce 53%-85% YOLoC branch area overhead through ROM CiM reusing

Thank You

Tianyi Yu, Tianyu Liao, Mufeng Zhou, Xiaotian Chu, Guodong Yin, Mingyen Lee, Yongpan Liu, Huazhong Yang, and **Xueqing Li**[1†]

[1]Tsinghua University

†Email: xueqingli@tsinghua.edu.cn