3D-METRO: Deploy Large-Scale Transformer Model on a Chip Using Transistor-Less 3D-Metal-ROM-Based Compute-in-Memory Macro

Yiming Chen*, Xirui Du*, Guodong Yin, Wenjun Tang, Yongpan Liu, Huazhong Yang, and **Xueqing Li[1†]**

[1]Department of Electronic Engineering, LFET/BNRist, Tsinghua University

[†]Email: xueqingli@tsinghua.edu.cn

* These authors equally contributed to this work

Outline

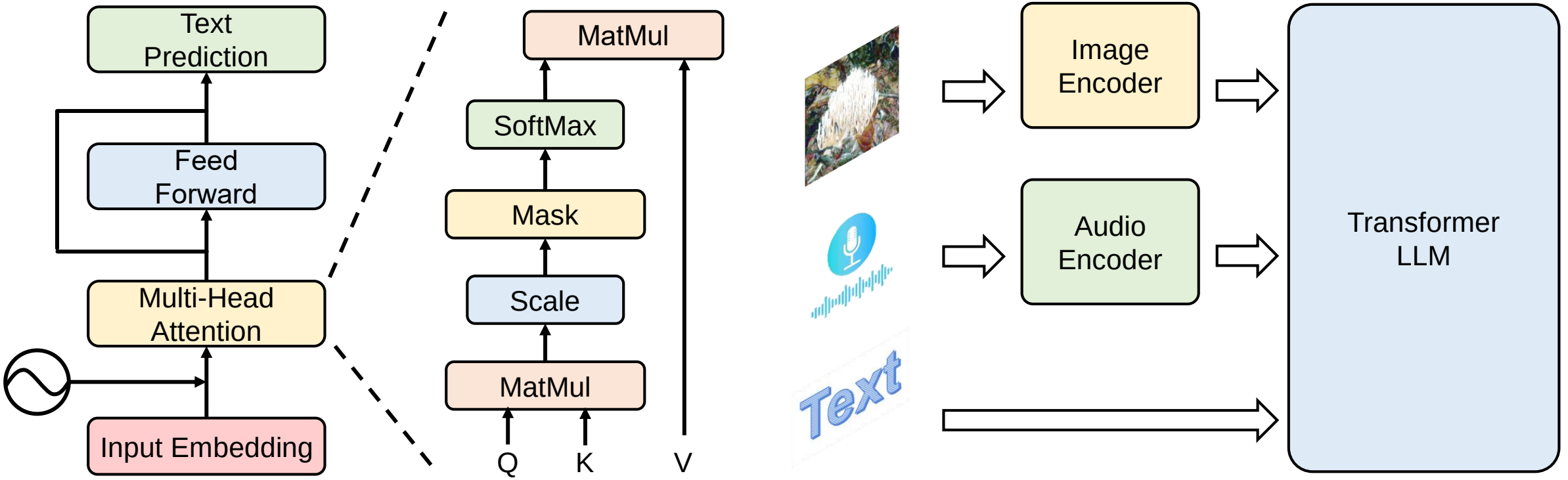
- Background
- Motivation
- Proposed Design
- Benchmark
- Conclusion

Outline

- **Background**
- Motivation
- Proposed Design
- Benchmark
- Conclusion

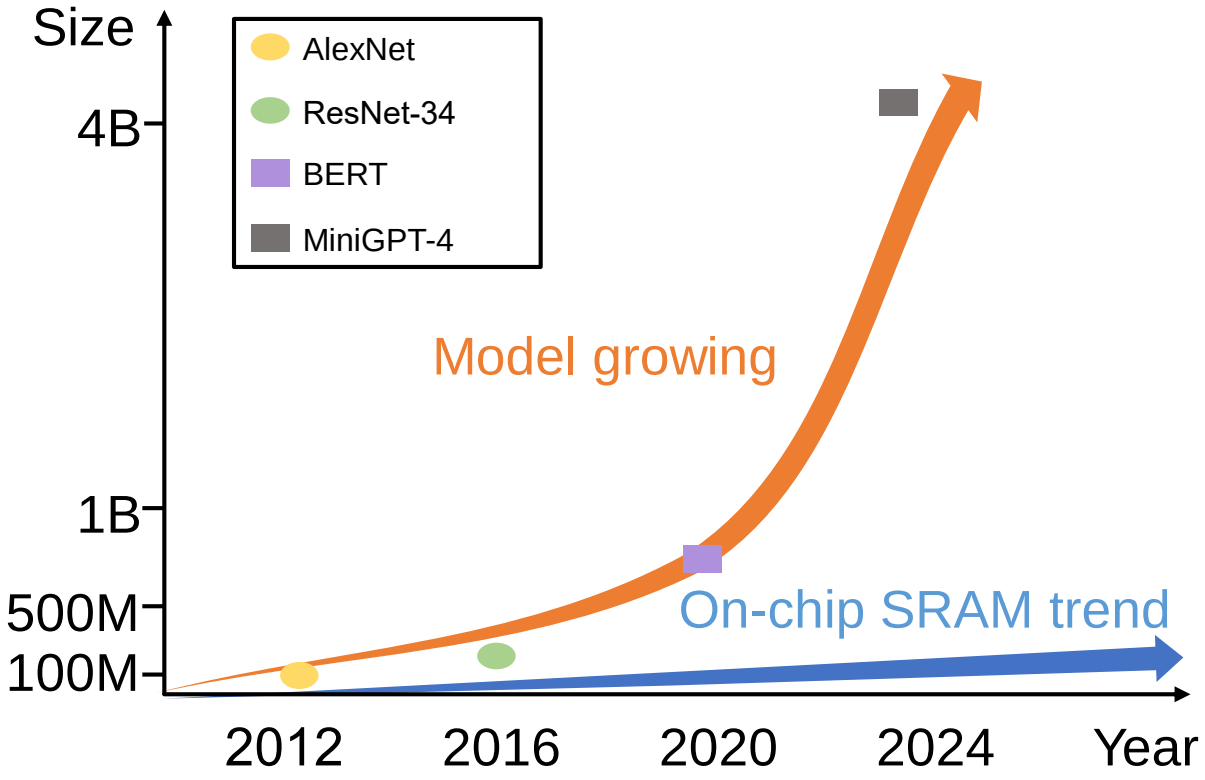
Background

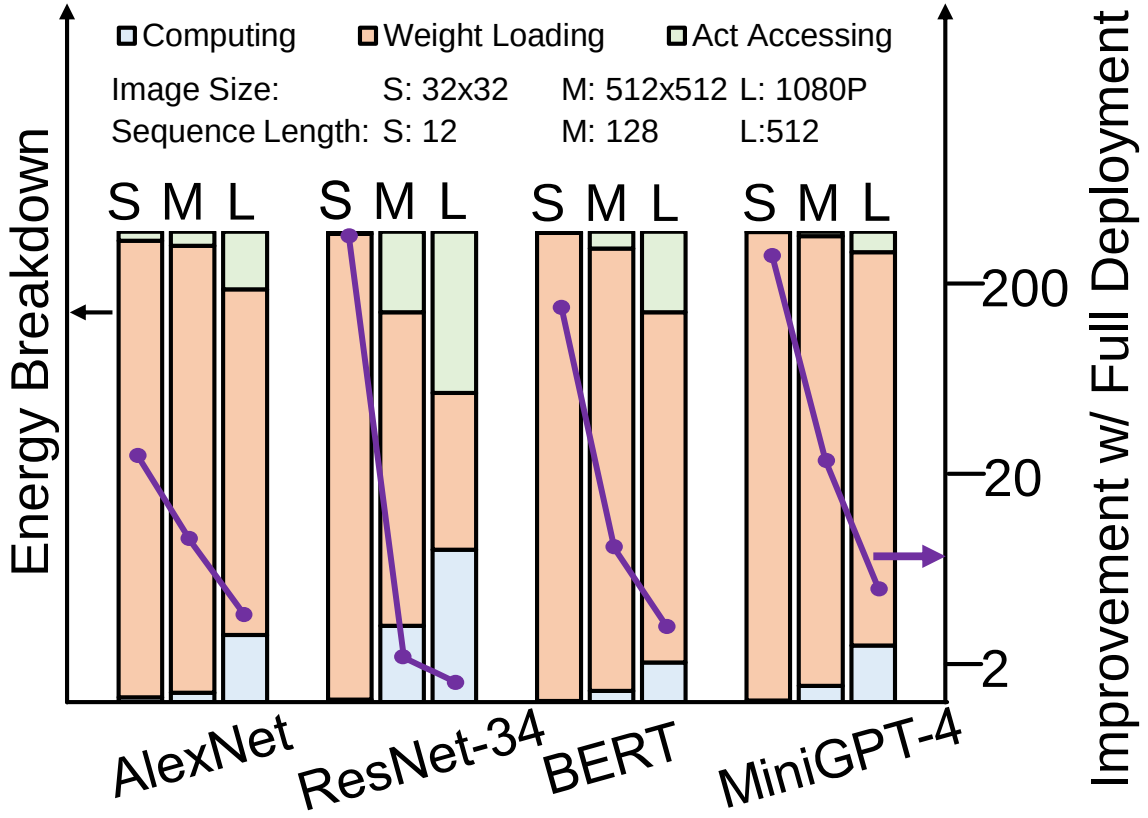
- Large language models (LLM) based on the Transformer (A. Vaswani et al., 2017) are blooming in both CV and NLP.
- Furthermore, multimodal large model, such as GPT-4 (J. Achiam et al., 2023) and MiniGPT-4 (D. Zhu et al., 2023), performs human-like AI in various applications.



Background

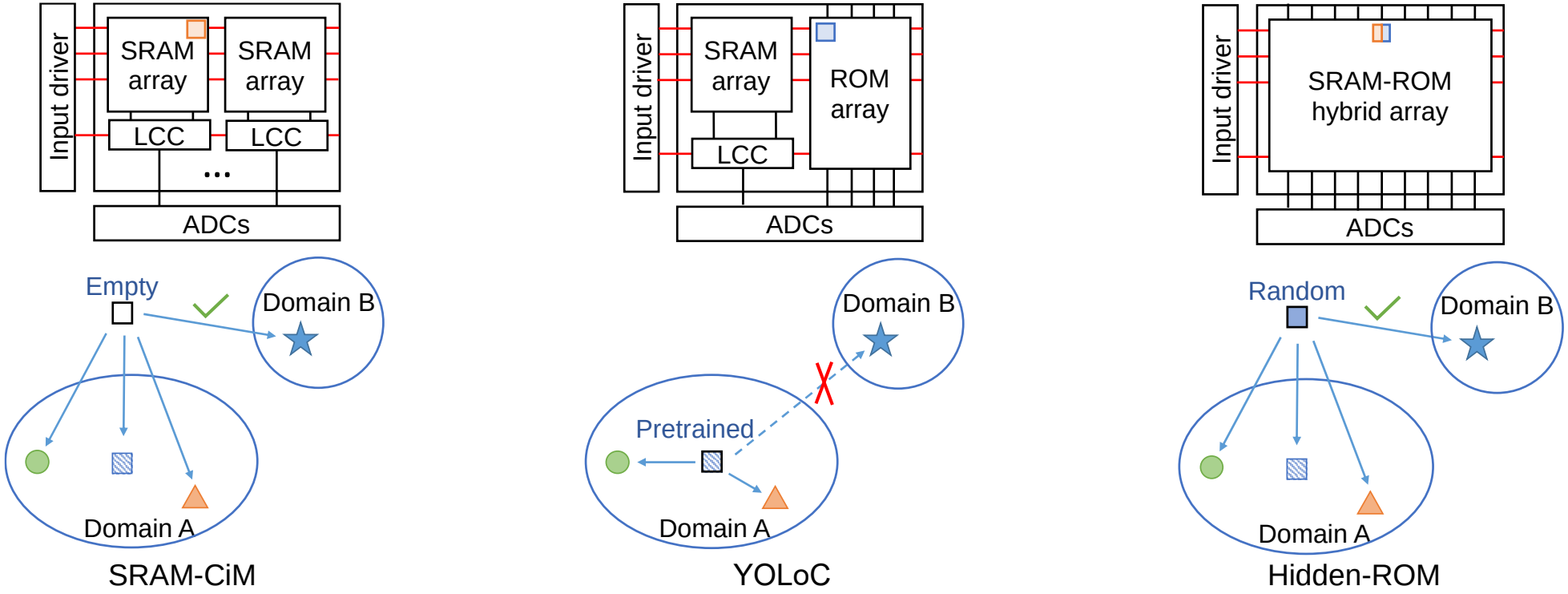
- SRAM-CiM, especially high-density designs, enables the in-memory attention mechanism.
- However, in short-sequence scenarios such as the edge side, the limited on-chip density still results in a serious weight-dumping overhead.





Background

- Recently, a high-density ROM-CiM (G. Yin et al., 2023) has been proposed to address the limited density challenges of SRAM-CiM.
- 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.



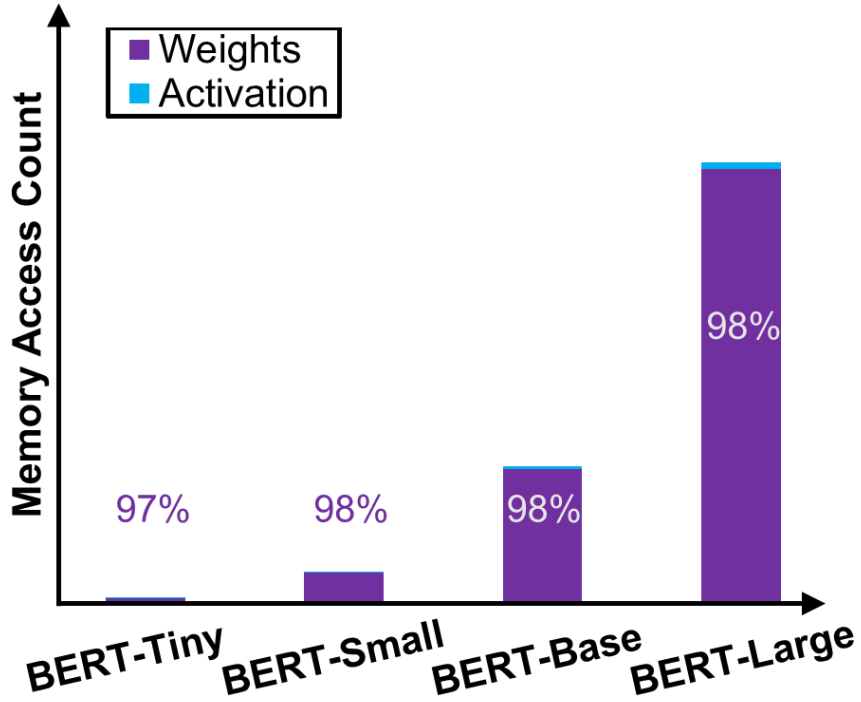
Source: Y. Chen *et al.*, *ICCAD'22*.

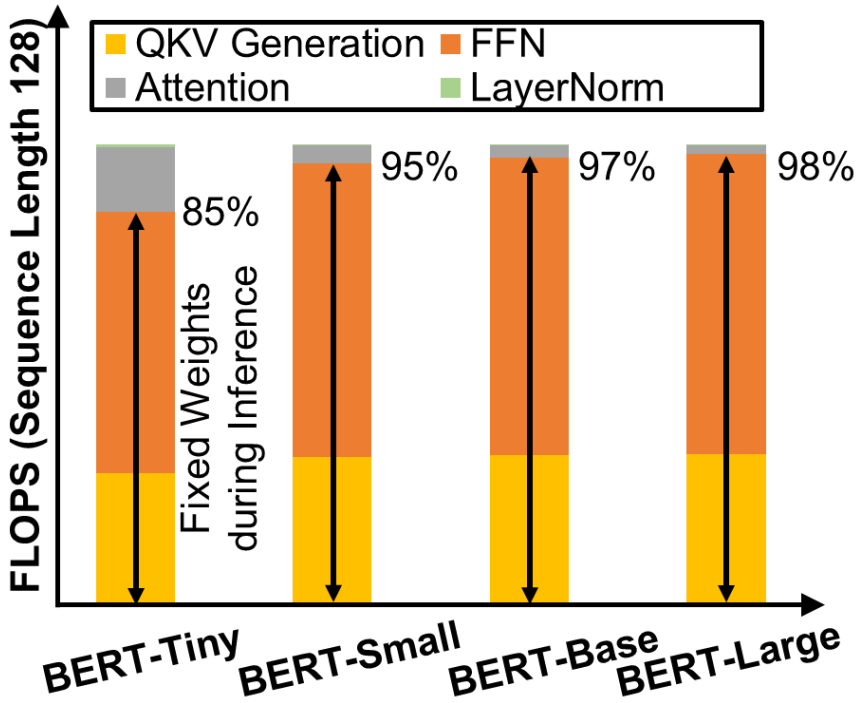
Outline

- Background
- **Motivation**
- Proposed Design
- Benchmark
- Conclusion

Motivation

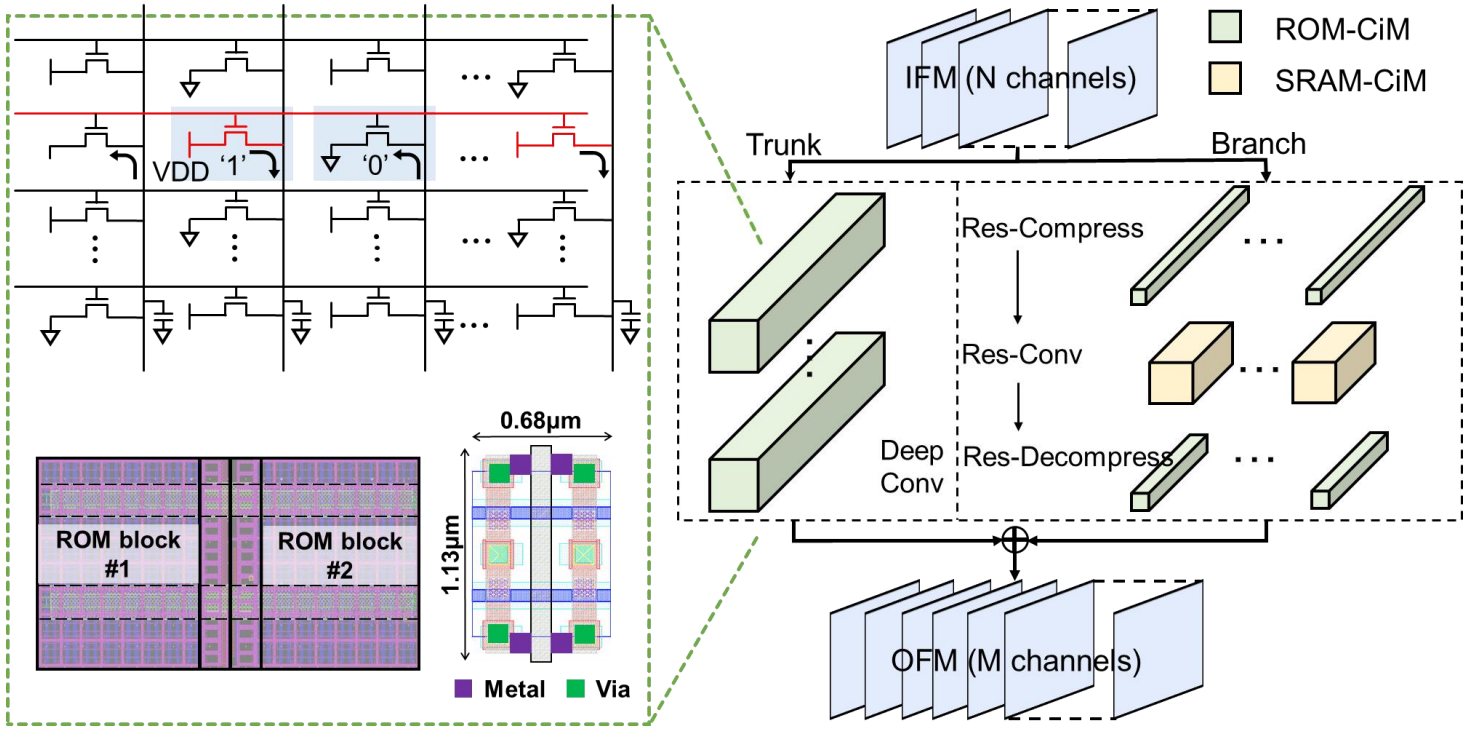
- Beyond RNN and LSTM, Transformer is currently the most dominant architecture for NLP and multimodal tasks.
- In short-sequence scenarios, the weight access dominates the memory access, while the weight-stationary MVM dominates the computing operations.

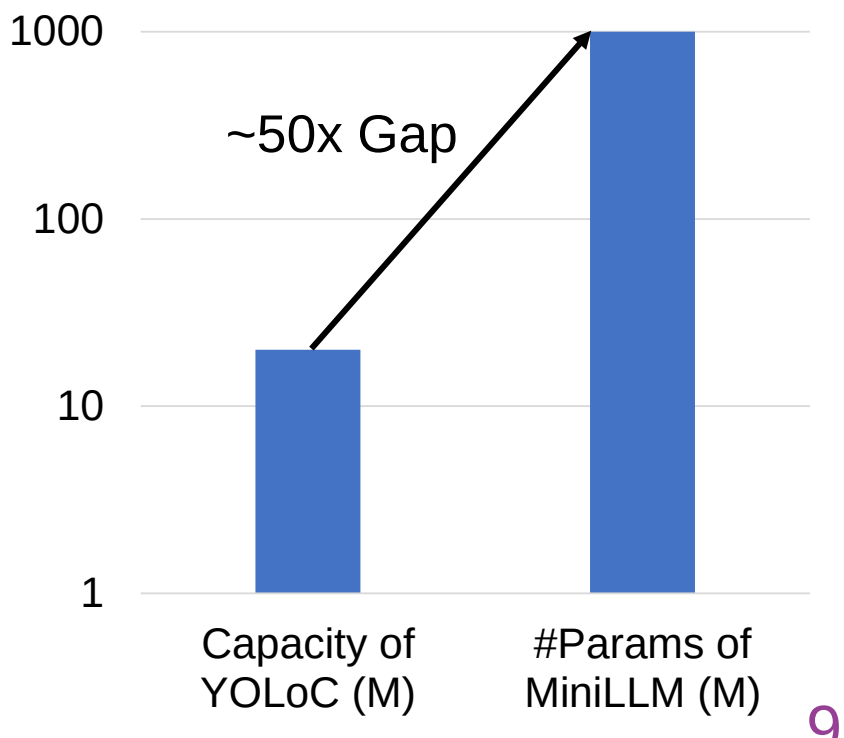




Motivation

- YOLoC (Y. Chen et al., 2022) demonstrates a new concept of cutting off off-chip parameter loading with large-capacity ROM-CiM and finetuning to various tasks.
- However, the density of conventional transistor-based ROM-CiM is still limited when it comes to even lite LLM, such as MiniGPT-4.





Contributions

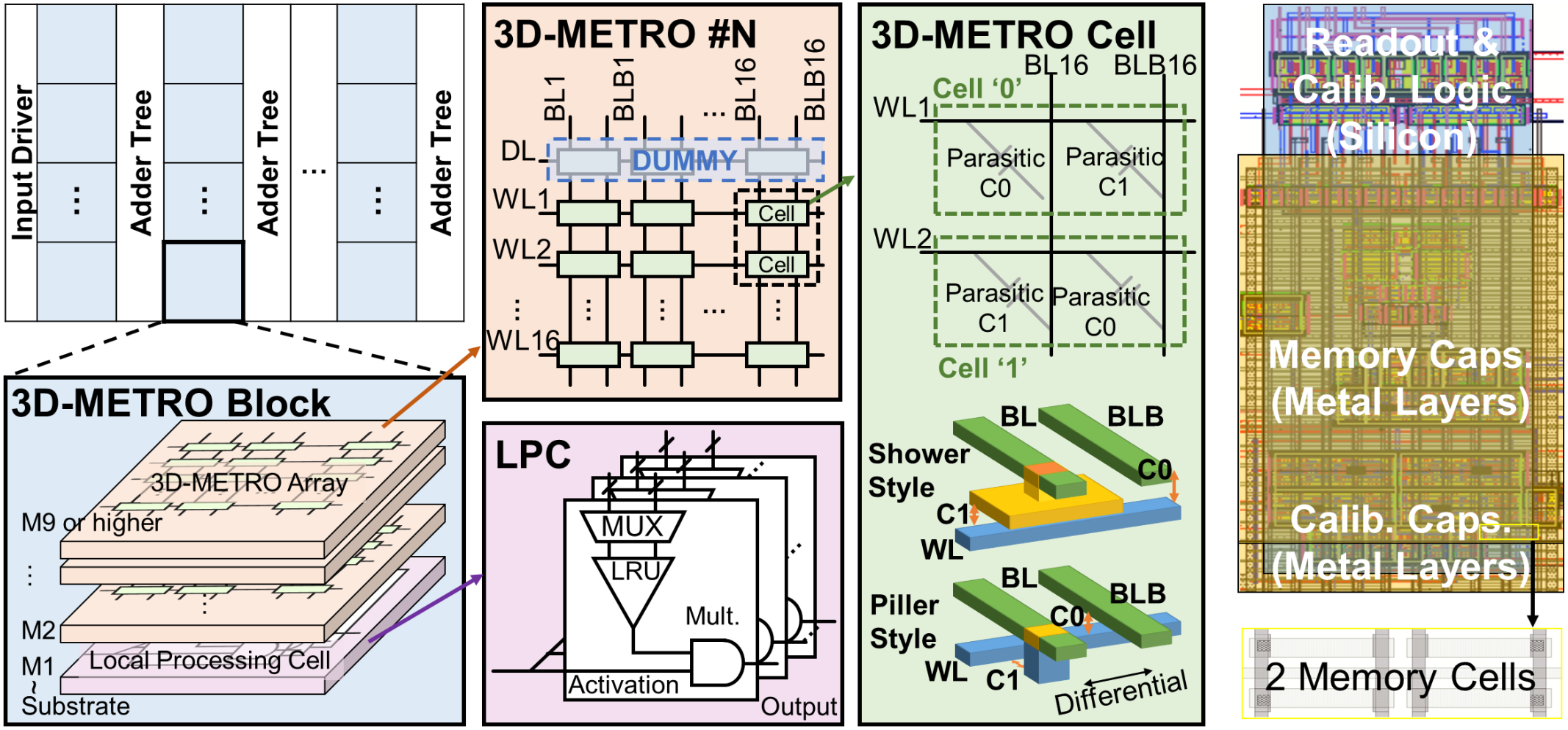
- How to further improve the density of **ROM-CiM**?
- The **key contributions** of **this work *3D-METRO***:
  - Transistor-less 3D-metal-ROM-based CiM with potentially 10x higher density than conventional ROM-CiM and 100-200x higher than SRAM-CiM.
  - Local recovering unit for counteracting the inter-column interference due to high-density arrays.
  - Potential of full deployment of a large language model on a $3cm^2$ chip of 28nm CMOS process with 28x energy efficiency improvement thanks to terminating off-chip weight loading.

Outline

- Background
- Motivation
- **Proposed Design**
- Benchmark
- Conclusion

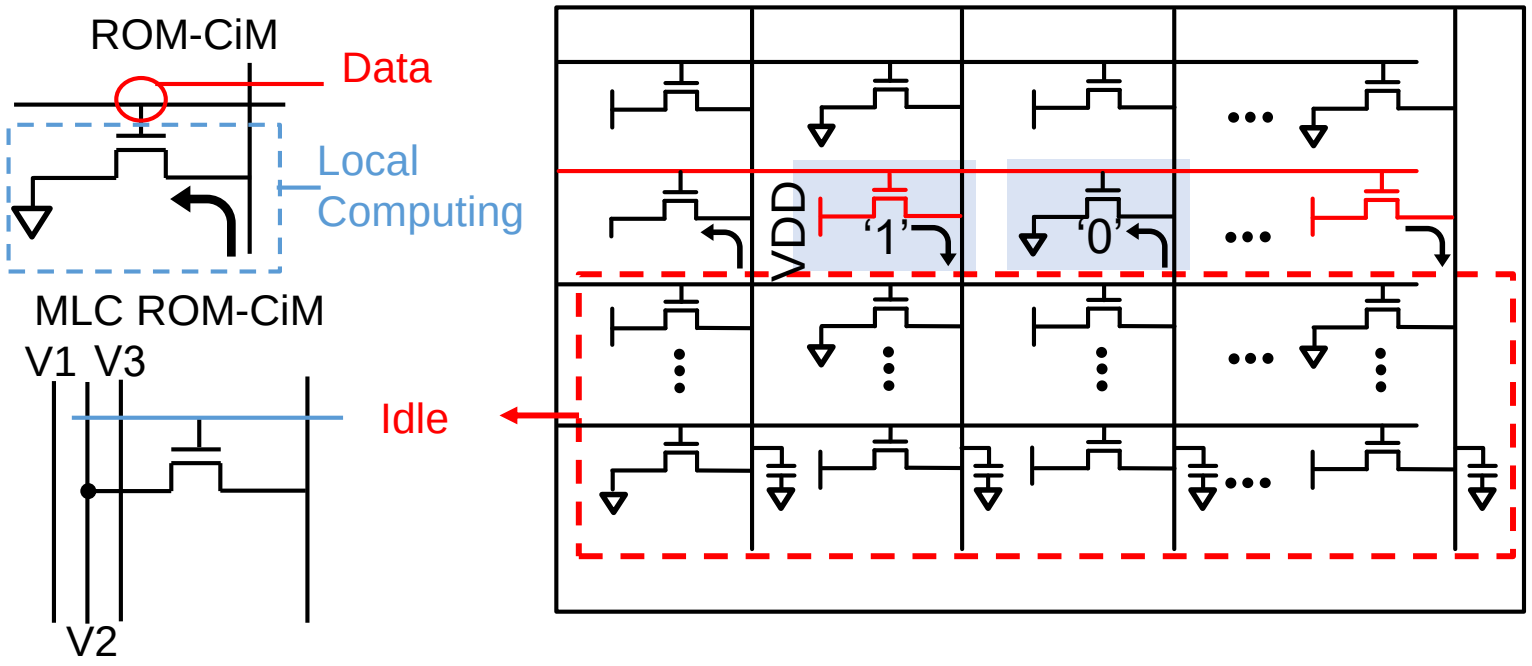
Proposed Design

- Proposed 3D-METRO contains **Transistor-Less-Metal-ROM-based CiM** Blocks, **Local Processing Cell**, and **Adder Tree** for computing

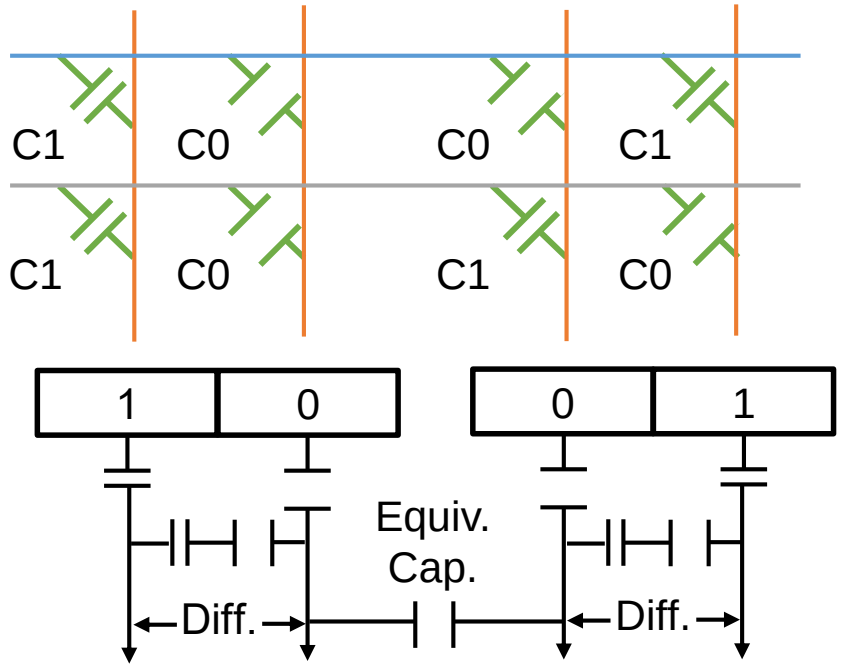


Proposed Design

- Conventional transistor-based ROM-CiM cell depends on the connection of the gate and the wordline, which does not rely on the transistor.
- We propose the Metal-ROM methodology, which uses only metal layers to construct the whole ROM array and corresponding differential readout.



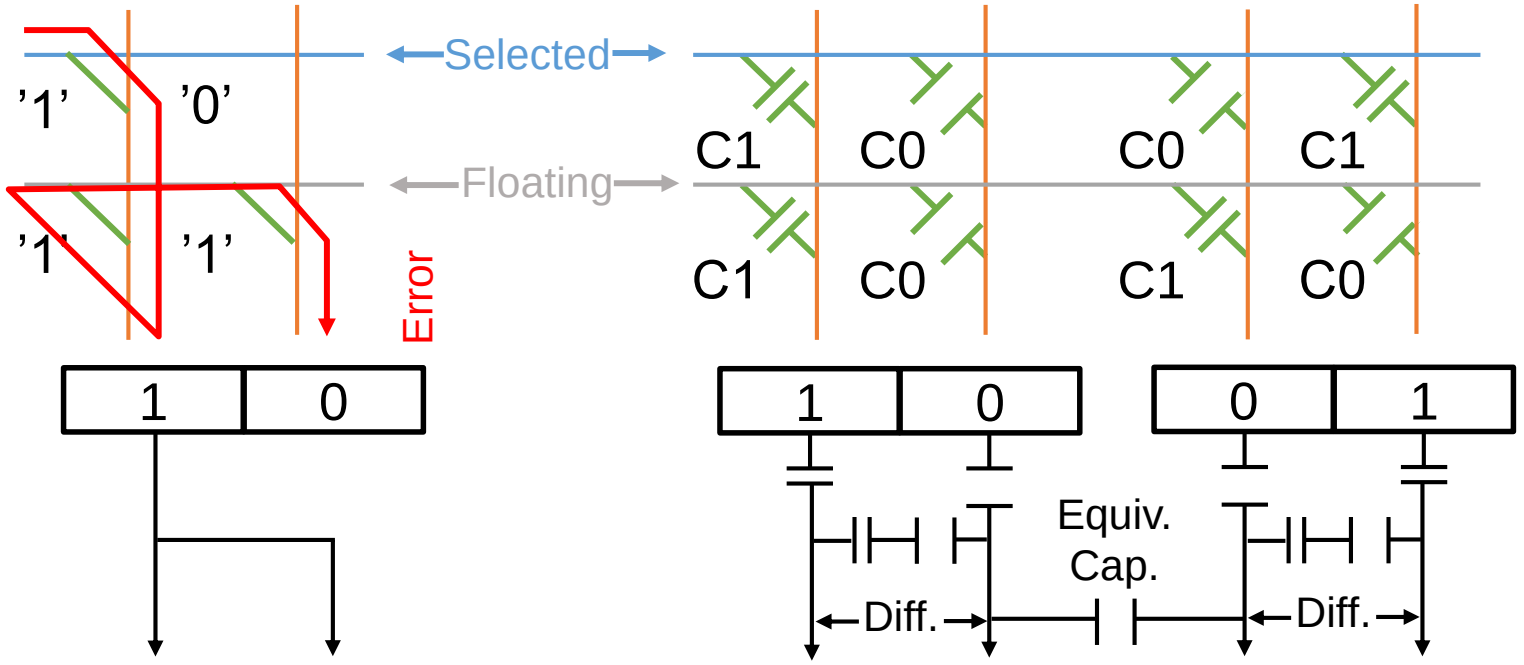
Conventional Transistor-based ROM-CiM



Metal-ROM

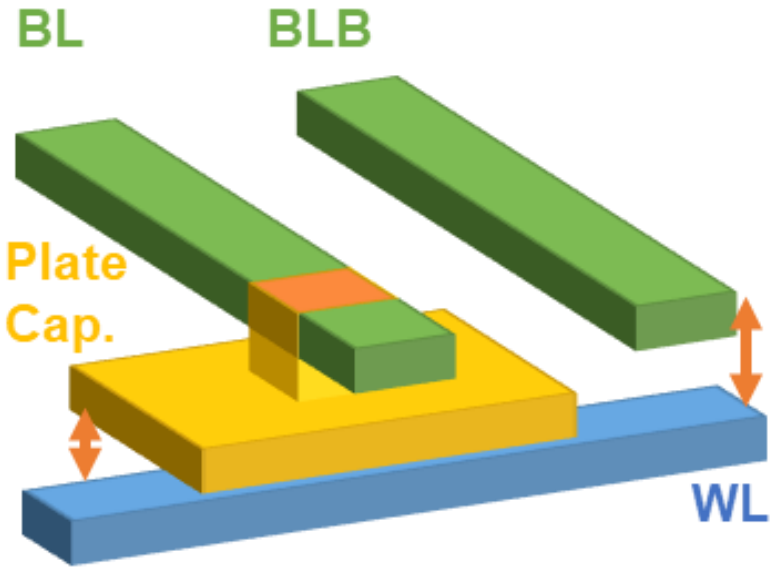
Proposed Design

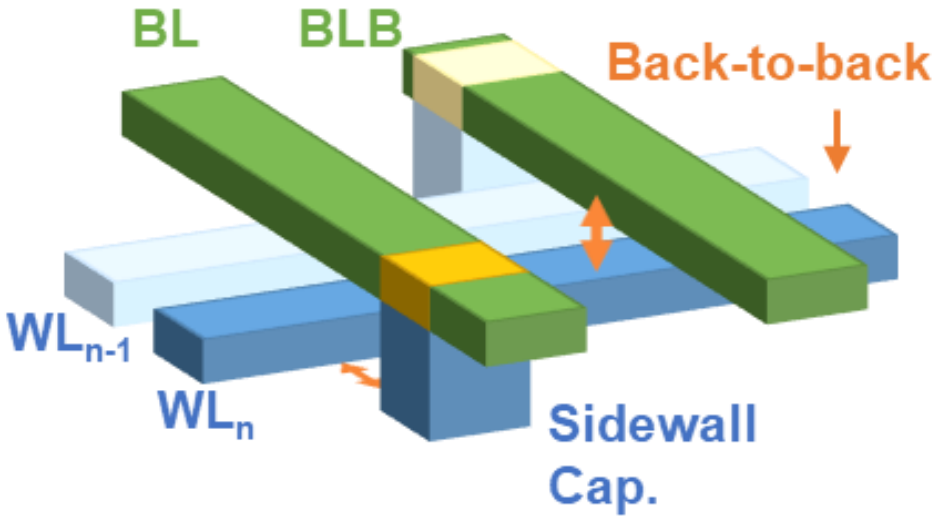
- However, there will be bit errors due to a short connection in special condition of transistor-less ROM-CiM.
- **Capacitive coupling scheme**: breaking the “no access controller, no correct read” by introducing differential parasitic capacitance C0 and C1.
- When only the selected WL is set to high, the data can be read by sensing the difference caused by C0 and C1 between the BL and BLB.



Proposed Design

- Two possible implementation: **Shower Type** and **Pillar Type**
- **Shower**: Doubling the capacitance of state ‘1’ by halving the distance.
- **Pillar**: Utilizing the sidewall capacitor to enhance the capacitors.

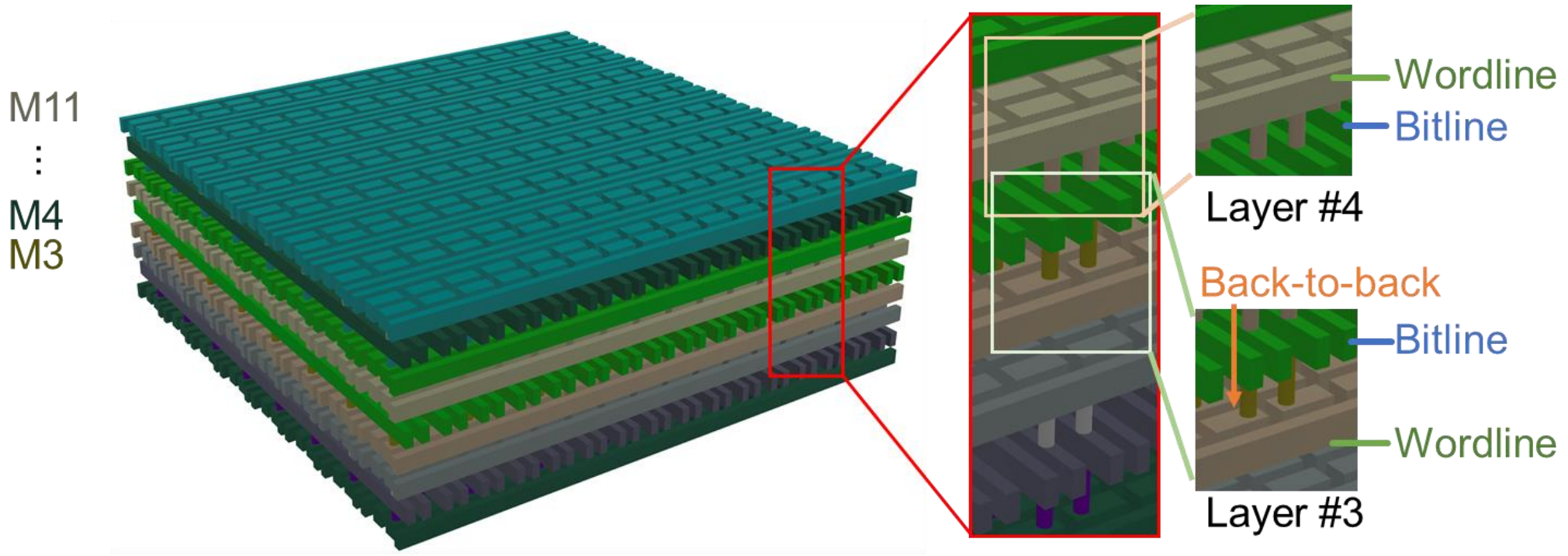




Shower		Pillar	
‘1’	‘0’	‘1’	‘0’
6.6 aF	2.8 aF	20 aF	3.2 aF

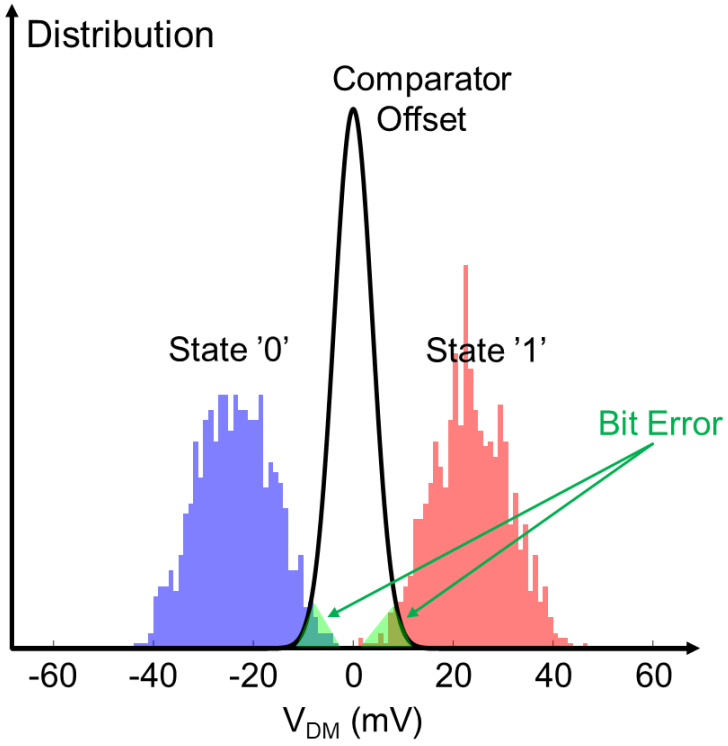
Proposed Design

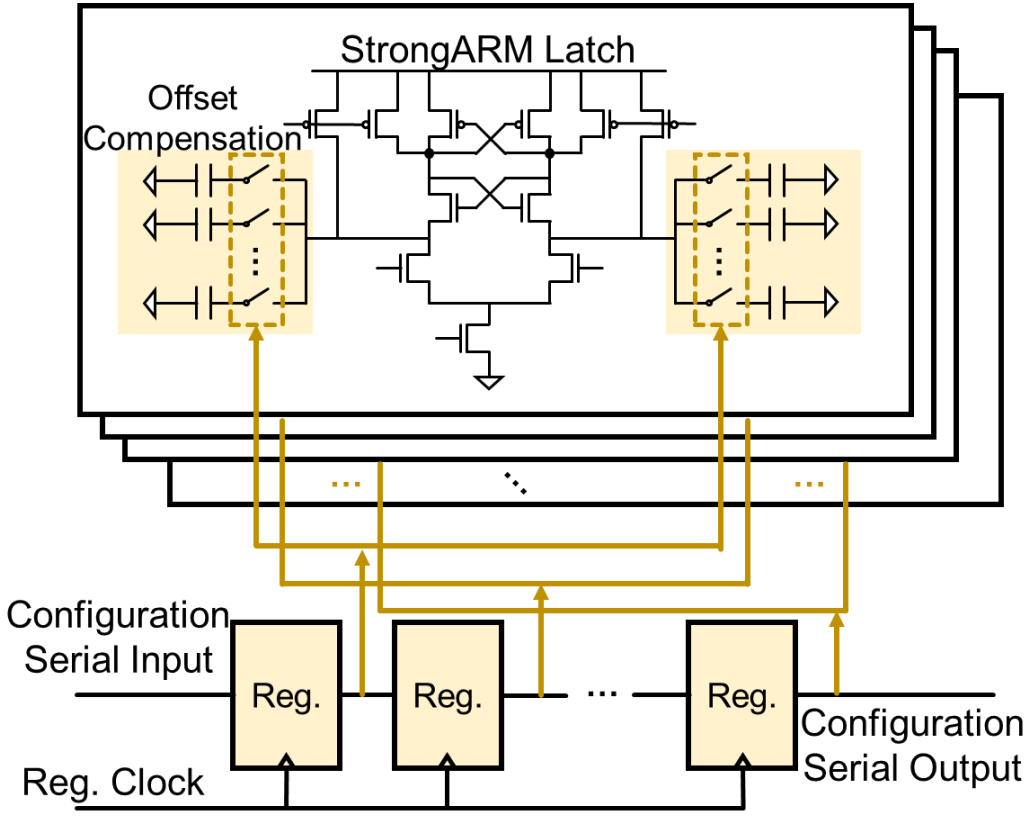
- Metal-based 3D stacking on mature CMOS



Proposed Design

- Limitation and Tradeoff: **Local Recovering Unit** (LRU)
  - Bit error due to the coupled parasitic capacitors and process variation of SA.
  - LRU: StrongARM latch structure and related offset compensation units to finetune the fixed offset of the SA for recovering the dirty data.





Outline

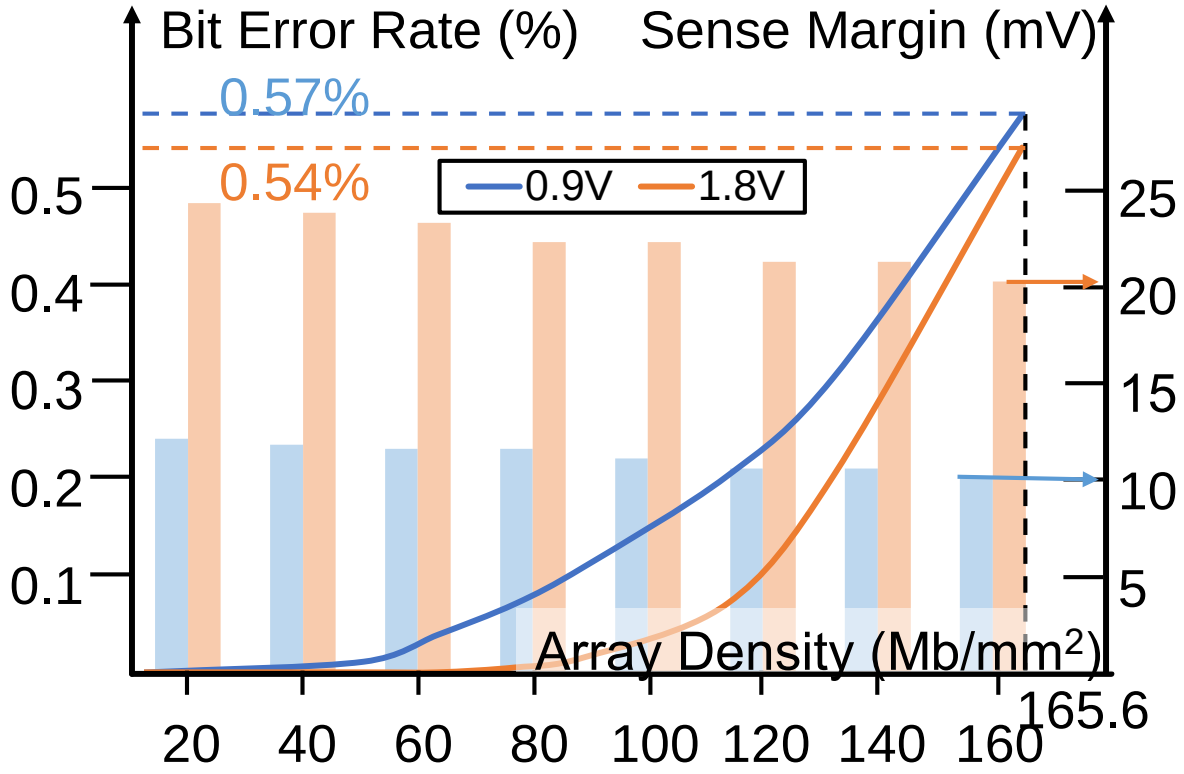
- Background
- Motivation
- Proposed Design
- **Benchmark**
- Conclusion

Benchmark

- **Experiment Setup**
- Hardware
  - 13-metal 28nm CMOS process. SRAM from CACTI. Baseline is based on HBM2.
  - Monte-Carlo simulation for evaluating the bit error rate due to process variation.
- Software
  - Model: BERT, and ViT
  - Dataset: SST-2, AG_NEWS, TinyImageNet, and CIFAR-100
- System
  - Data transmission for weight data and intermediate result is included.
  - Custom simulator based on the data by fvcore and the macro-level simulation.

Benchmark

- Reliability and Trade-off Analysis
- Tradeoff between **memory density** and **bit error rate** in variations considering PVT corners, Monte-Carlo, and inter-column interference.



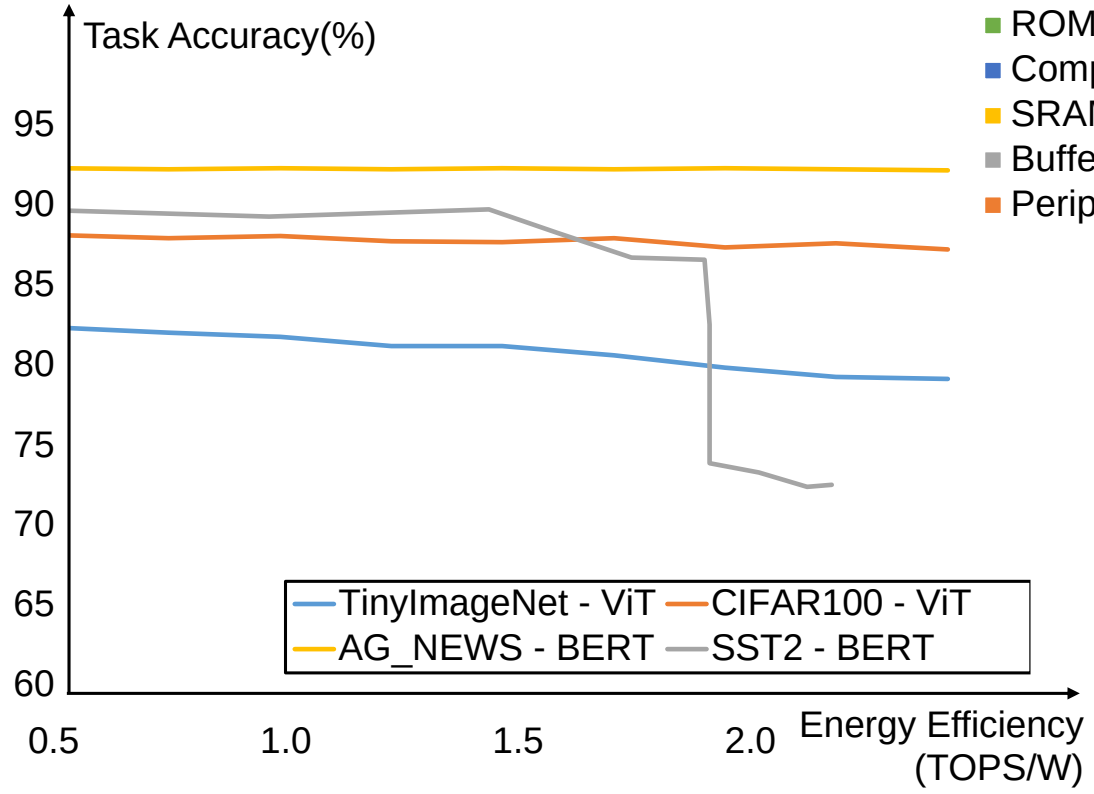
Size of Block (bit)

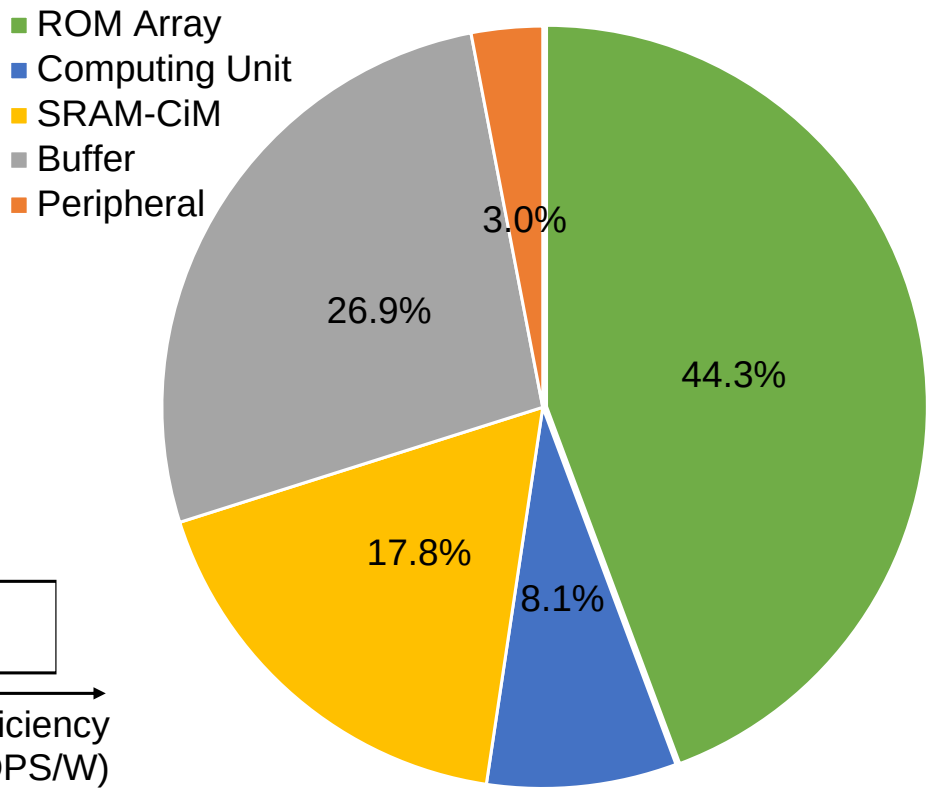
64	128	256	512	1024	Offset of Readout (mV/σ)
0.12	0.14	0.17	0.29	0.94	2.0
0.20	0.24	0.37	0.72	2.0	2.5
0.34	0.43	0.65	1.3	3.6	3.0
0.51	0.65	1.1	2.1	5.3	3.5
0.71	0.95	1.6	3.1	7.5	4.0
1.0	1.4	2.3	4.4	9.6	4.5

Plot of Bit Error Rate (%)

Benchmark

- **Accuracy** and **energy efficiency** tradeoff of various tasks
  - Improve the parasitic capacitors to enhance the task accuracy.
  - Use a more varied but more energy-efficient design for the low-bit weights to balance energy efficiency and accuracy.





Benchmark

- **Macro-level** Performance Comparison
  - ❑ Breaking through the traditional trade-off.
  - ❑ Achieving ultra-high density by 3D-Metal stacking (> 100x).
  - ❑ Making full use of the transistor layers to achieve 5.8x higher area efficiency.

Metrics	This Work	ROM-CiM JSSC'23 [11]	SRAM-CiM JSSC'23 [19]
Process	28 nm CMOS		
Array Density (Mb/mm^2)	165.6	16.4	<1.0
Models Support	Large	Medium/Large	Large
Flexibility	LoRA	ReBranch	Fully
Energy Efficiency Per Parameter (EBOPS/W·mm^2)	31.0 (12x)	23.6/2.8 (9.2x/1.1x)	2.56 (1x)
Area Efficiency (GOPS/mm^2)	1283 (5.8x)	119 (0.53x)	221 (1x)

Benchmark

- **System-level** Metrics Comparison on LLM
  - The SRAM-CiM is the fine-tuning weights in LoRA.
  - The SRAM buffer is used to store the intermediate data on-chip completely.

Metrics		This work	SOTA CiM [19]
Process		28 nm	
Chip Area (cm^2)		3.0	0.068
	ROM-CiM (cm^2)	1.55	-
	SRAM-CiM (cm^2)	0.53	0.019
	SRAM Buffer (cm^2)	0.80	0.015
On-chip Weights (Mb)		3342	0.19
On-chip Weight Density (Mb/cm^2)		1160 (414x)	2.8
Energy per Token (w/ off-chip data fetching) (μJ/Token)		222.5	6146 (28x)
Task Flexibility		Finetuning	Transformer

Outline

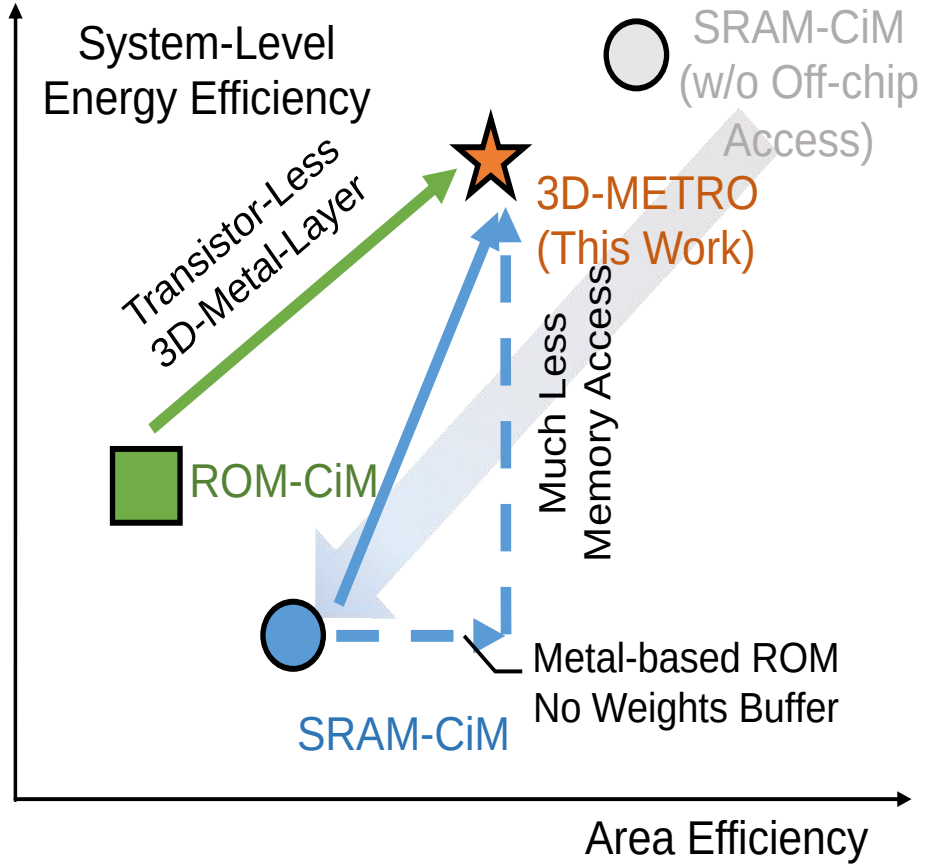
- Background
- Motivation
- Related Works
- Proposed Design
- Benchmark
- **Conclusion**

Discussion

- Reliability
  - Repressed random effects with the proposed recovery scheme by differential capacitor pairs.
- Flexibility
  - Enabled by parameter finetuning technique, such as LoRA.
- Heat Dissipation
  - 3D stacking by metal on mature CMOS process
  - No multilayer silicon stack with poor heat dissipation
  - 3D Stack for low power density ROM memory.
- Overhead
  - Peripheral MUXs
  - MUXs could be fully deployed beneath 6 stacked arrays with 16x16 size.

Conclusion

- Proposed 3D-METRO architecture
  - Ultra-high-density compute-in-memory structure.
  - Transistor-less 3D-Metal-ROM layers.
- Features:
  - A transistor-less ROM with local computing and local recovering units potentially supports 100x-200x density of SRAM-CiM.
  - Task evaluation shows 28x system-level energy improvement over SRAM-CiM
  - Only 10% area efficiency overhead and <1% accuracy loss in BERT and ViT.



Thank You

Yiming Chen*, Xirui Du*, Guodong Yin, Wenjun Tang,
Yongpan Liu, Huazhong Yang, and **Xueqing Li[1†]**

[1]Department of Electronic Engineering, LFET/BNRist, Tsinghua University

[†]Email: xueqingli@tsinghua.edu.cn