ZEBRA: A Zero-Bit Robust-Accumulation Compute-In-Memory Approach for Neural Network Acceleration Utilizing Different Bitwise Patterns

Yiming Chen[1], Guodong Yin[1], Hongtao Zhong[1], Mingyen Lee[1], Huazhong Yang[1], Sumitha George[2], Vijaykrishnan Narayanan[3], and **Xueqing Li**[1†]

[1]Tsinghua University, [2]North Dakota State University, [3]Pennsylvania State University

†Email: xueqingli@tsinghua.edu.cn

Outline

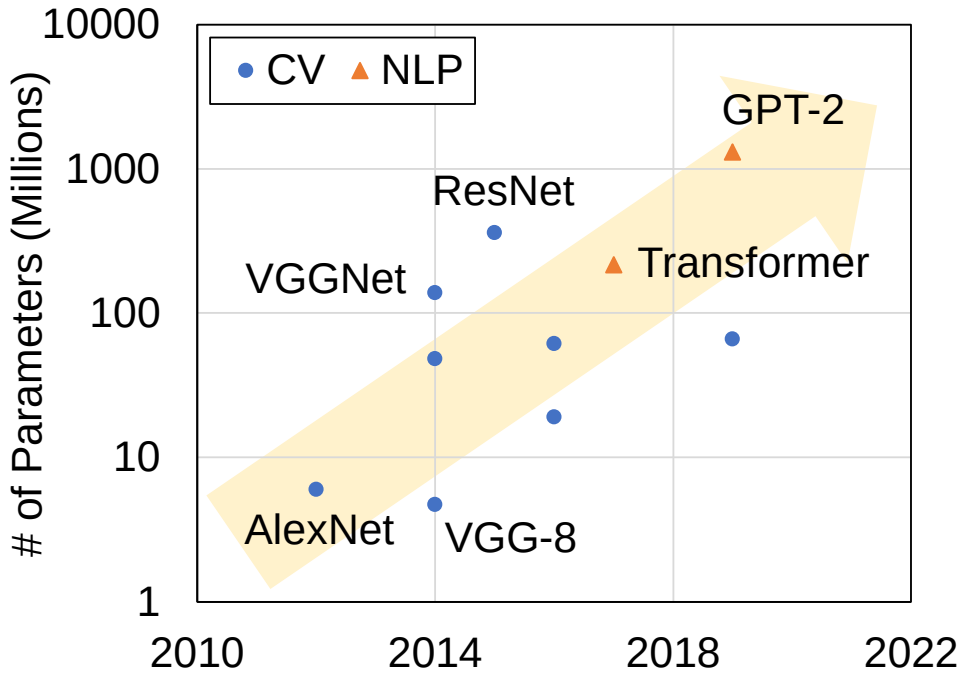
- Background
- Motivation
- Related Works
- Proposed Design
- Benchmark
- Conclusion

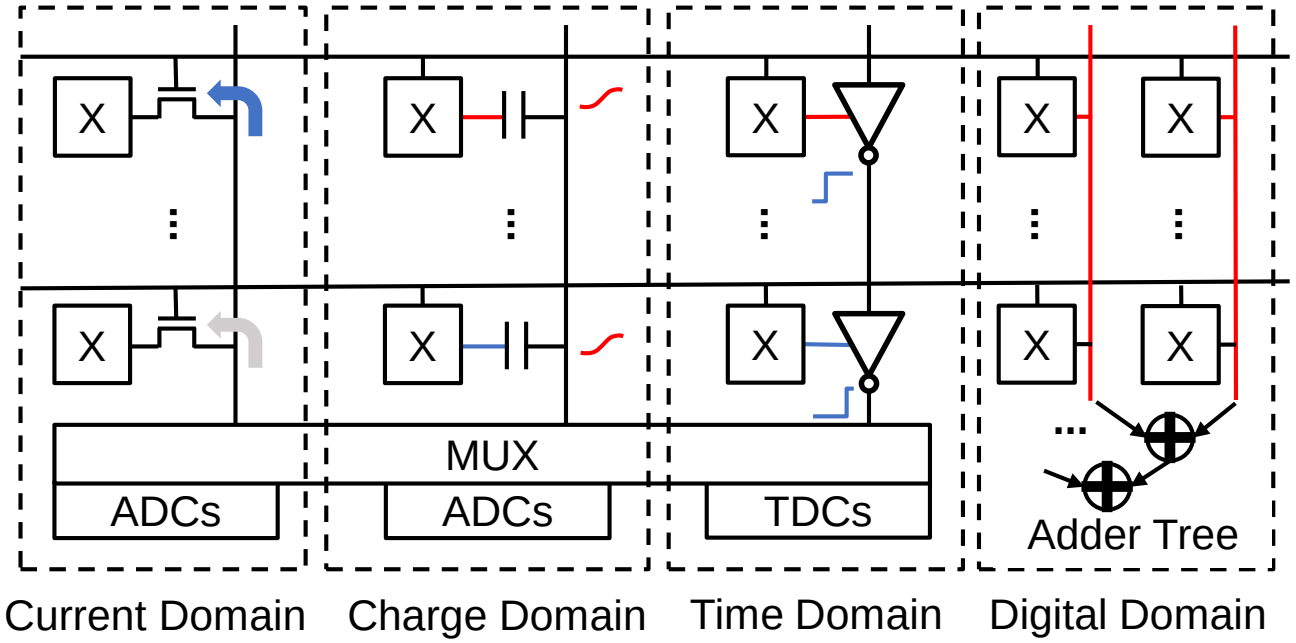
Outline

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

Background

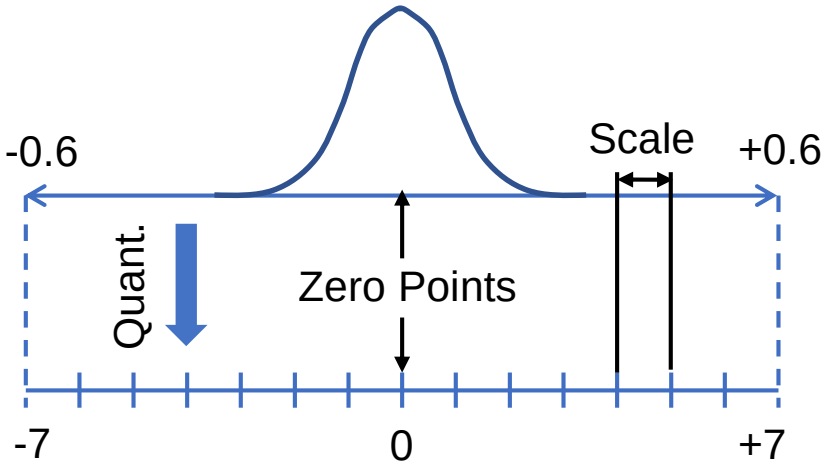
- Artificial intelligence (AI) enabled by deep neural networks (DNNs) has made significant breakthroughs.
- However, the large amount of memory access incur memory wall issue.
- **Compute-in-memory** is a promising technique to solve this problem.

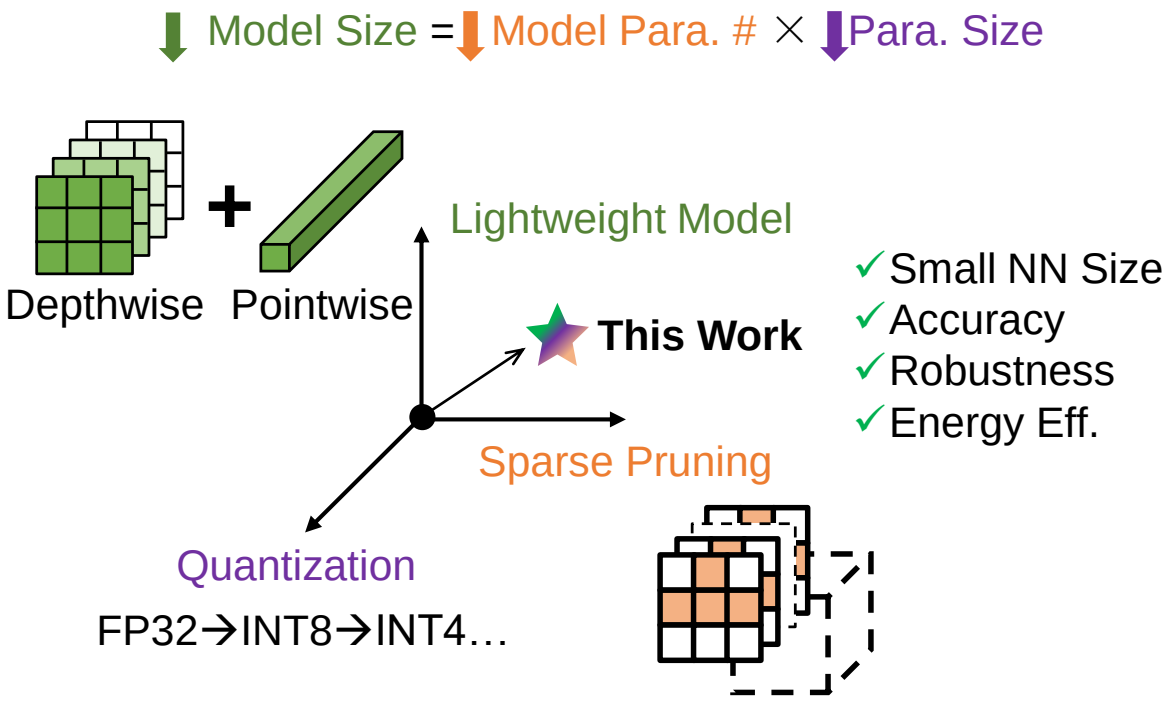




Background

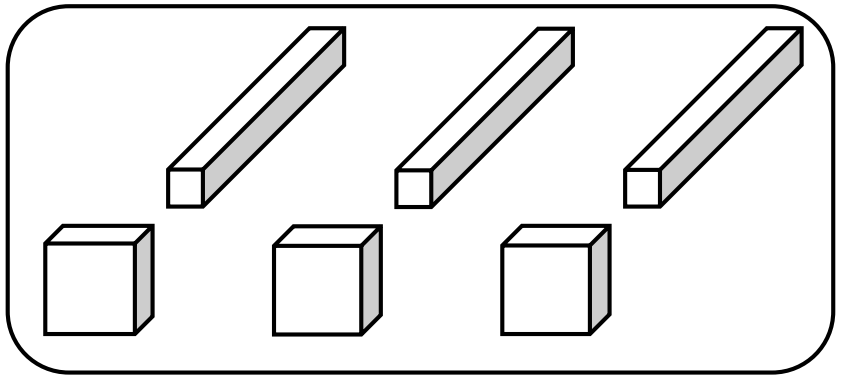
- Nevertheless, despite the high macro-level energy efficiency, there remain severe challenges to supporting **large models** on CIM.
- Several techniques have been developed to address this issue:
  - Lightweight model (MobileNet, etc.)
  - Layer-wise sparse pruning
  - Weight quantization



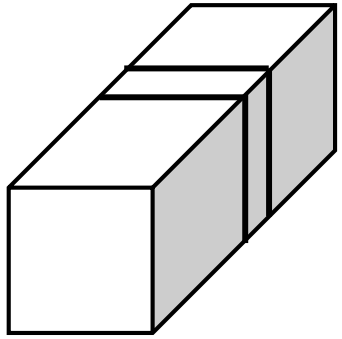


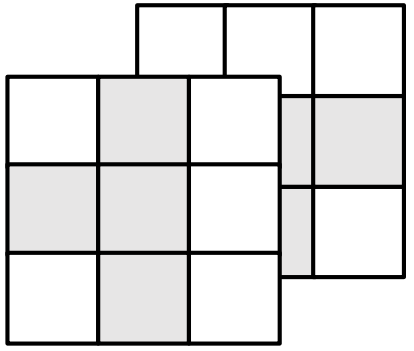
Background

- Lightweight model
  - MobileNet, ShuffleNet, ...
  - Challenging on analog-based CIM deployment due to lower redundancy.
- Sparse pruning
  - Unstructured sparsity methods: not well-suited for CIM arrays.
  - Structured sparsity methods: More feasible for energy-efficient CIM computations.



MobileNet





Pruning

Background

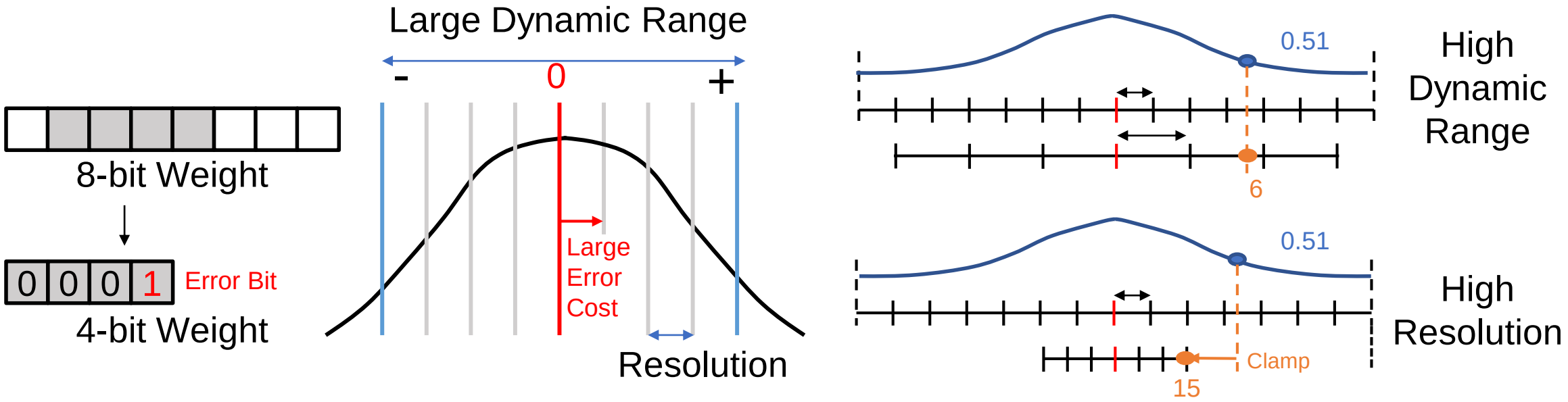
- Lightweight model
  - MobileNet, ShuffleNet, ...
  - Challenging on analog-based CIM deployment due to lower redundancy.
- Sparse pruning
  - Unstructured sparsity methods: not well-suited for CIM arrays.
  - Structured sparsity methods: More feasible for energy-efficient CIM computations.
- Quantization
  - Parameter-level approach
  - Vector-wise quantization: efficient 4-bit quantization over a wide range and delivers significant performance improvement on GPUs.
  - Challenges in irregular quantization scales

Outline

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

Motivation

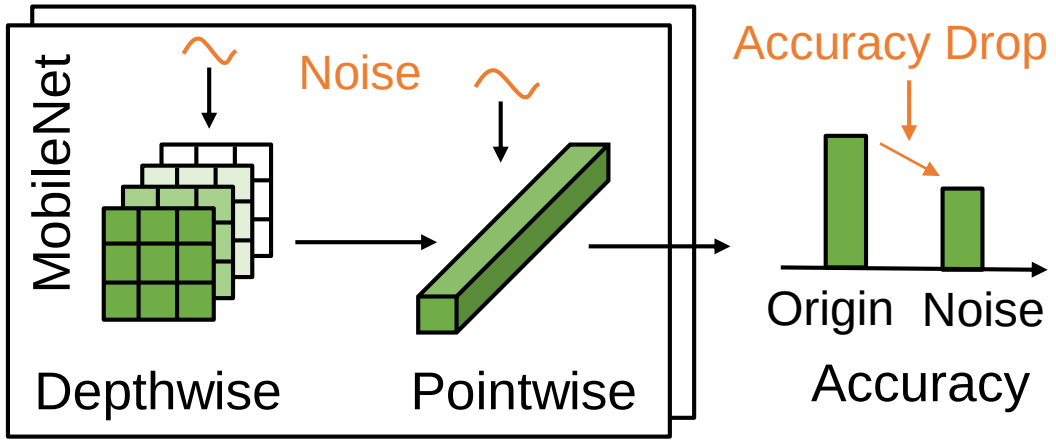
- Insight into low-bit quantization and bitwise structured sparsity in CIM:
  - Low-bit quantization can be modeled as a form of bitwise sparsity based on 8-bit quantization
  - There is a dilemma between the quantization dynamic range and resolution.

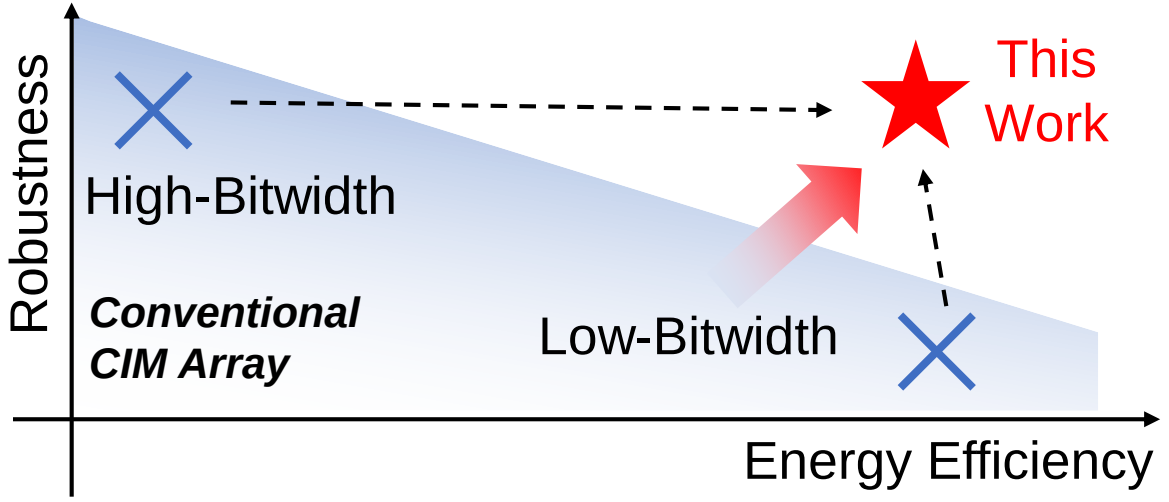


Dilemma between dynamic range and bit-error cost in low-bitwidth CIM

Motivation

- Insight into low-bit quantization and bitwise structured sparsity in CIM:
  - Accuracy drop due to reduced signal-noise rate (SNR) by lightweight quantized model
  - There is a trade-off in energy efficiency and robustness between high-bitwidth and low-bitwidth



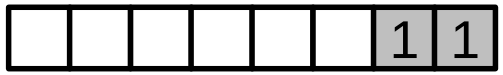


Dilemma between energy efficiency and robustness

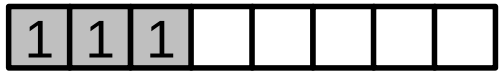
Motivation

- We ask: ***is it possible to introduce robust zero-bit compression on CIM to deploy a highly efficient lightweight model?***
  - ZEBRA is proposed with zero-bit patterns utilization and corresponding hardware-software co-optimization.

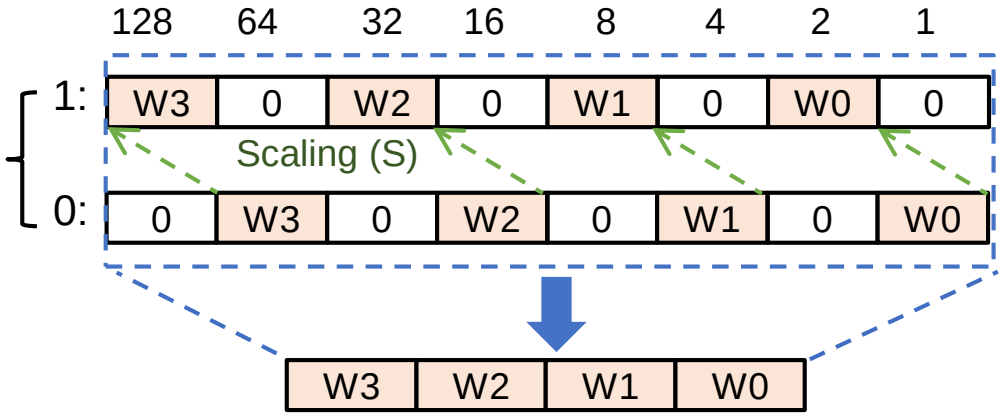
Resolution for small value



Dynamic Range for large value



Zero-bit Robust-Accumulation CIM: ***ZEBRA***



Contributions

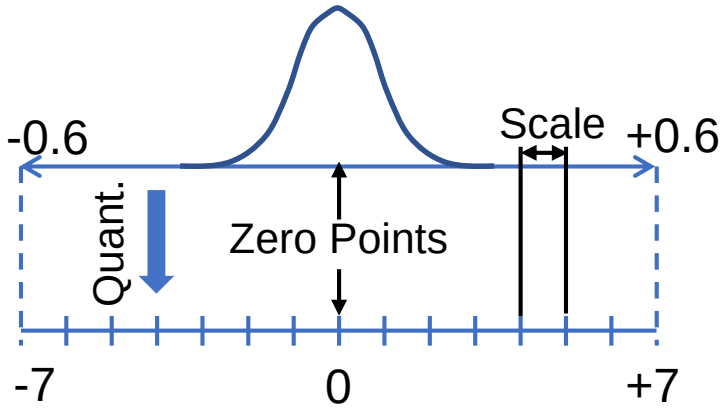
- The **key contributions** of **this work *ZEBRA***:
  - A robust low-bitwidth data encoding method enabled by value-adaptive zero-bit patterns.
  - A local computing unit supports multi-level input and weight multiplication.
  - Rich experiment results across application, macro, and system levels.

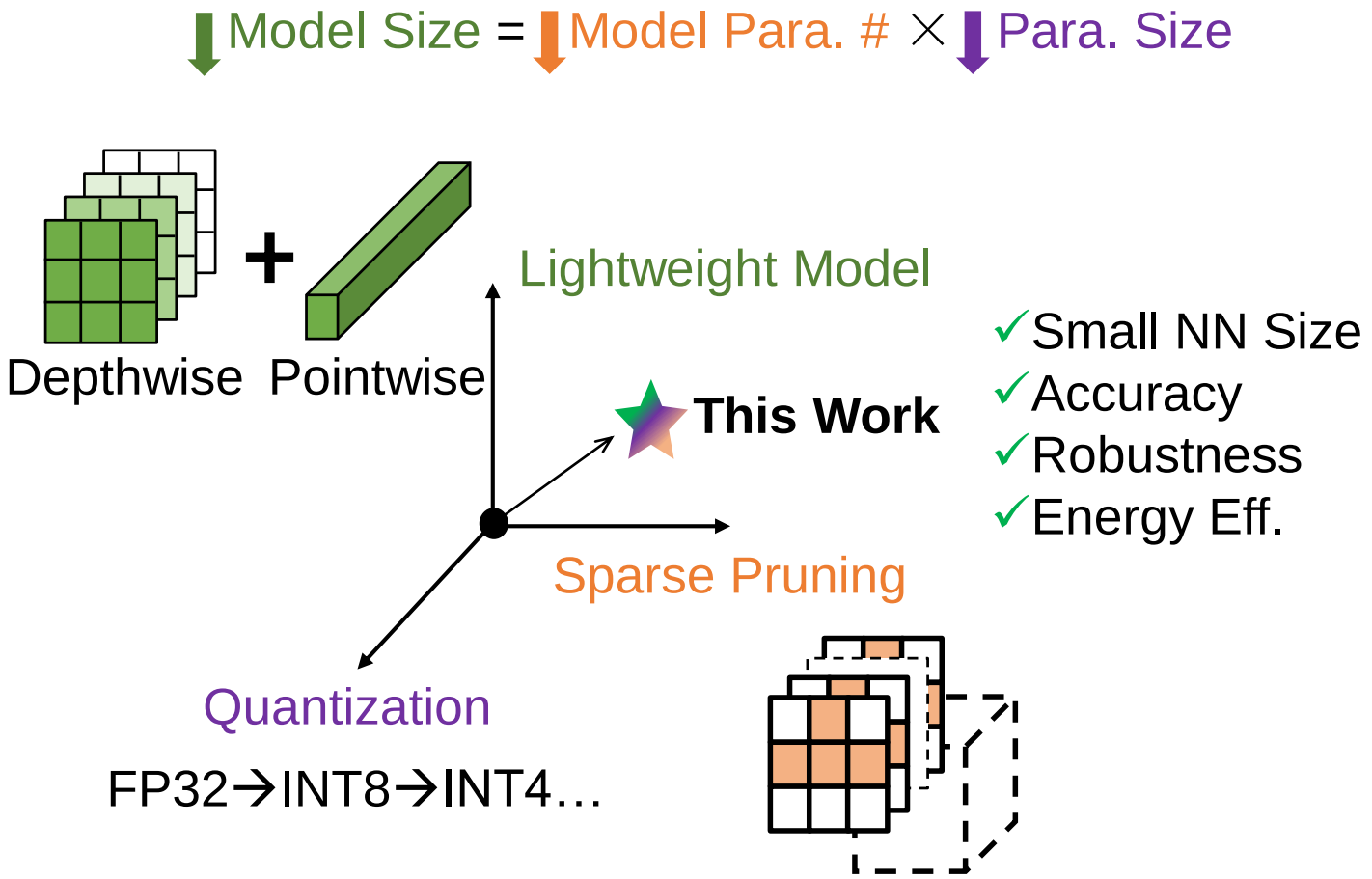
Outline

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

Related Works

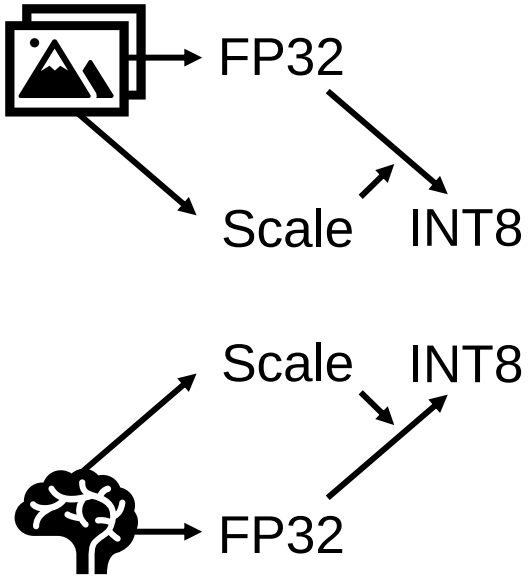
- Software-hardware co-optimization approaches for model compression
  - Quantization
  - Lightweight model
  - Sparsity utilization



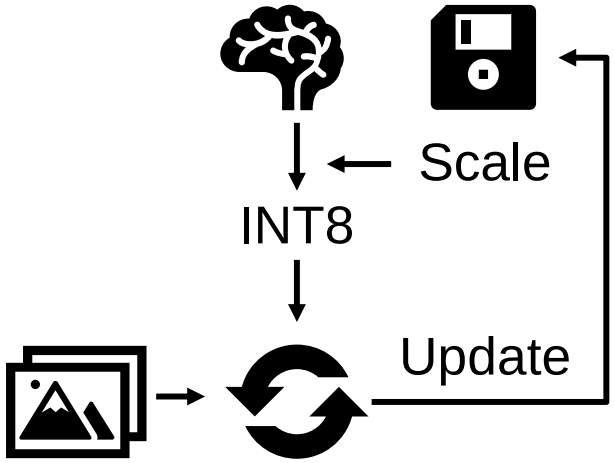


Related Works: Quantization

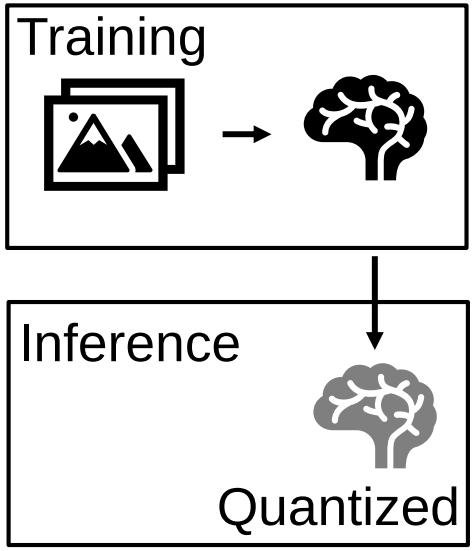
- Inference with integer weights and activations:
  - Dynamic quantization
  - Quantization-aware training (QAT)
  - Post-training quantization



Dynamic quantization



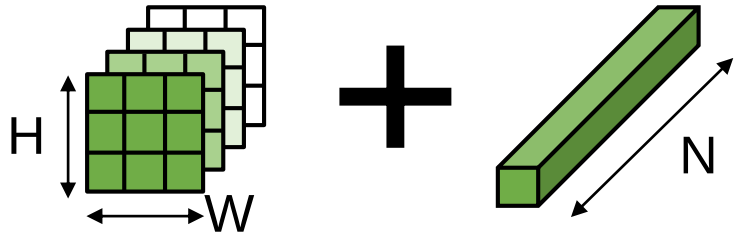
QAT



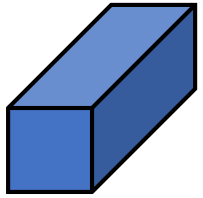
Post-training quantization

Related Works: Lightweight Models

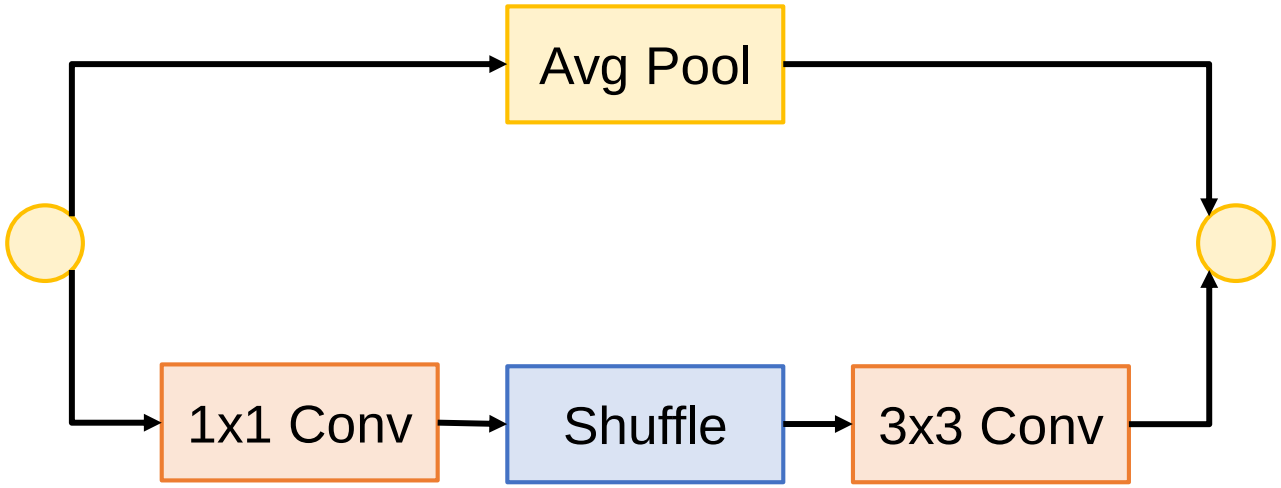
- Structure compress and weight reusing:
  - Depthwise separable convolution
  - Pointwise group convolution and channel shuffle



Bottleneck # of Para.: $N \times W \times H + N^2$

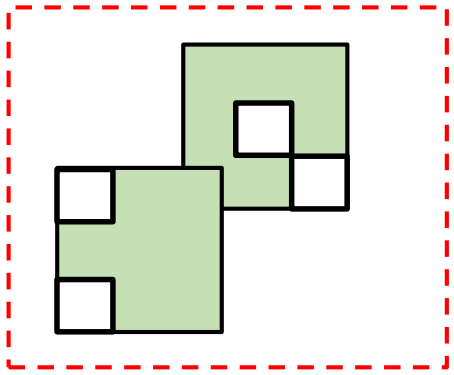


Conv2d # of Para: $N^2 \times W \times H$

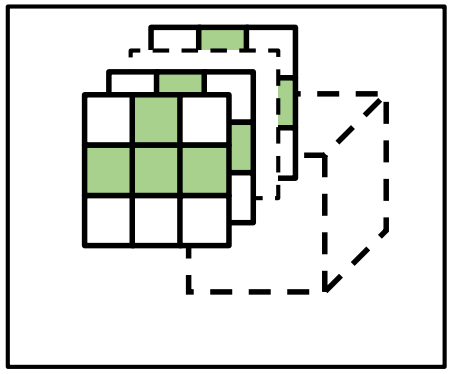


Related Works: Sparsity Utilization

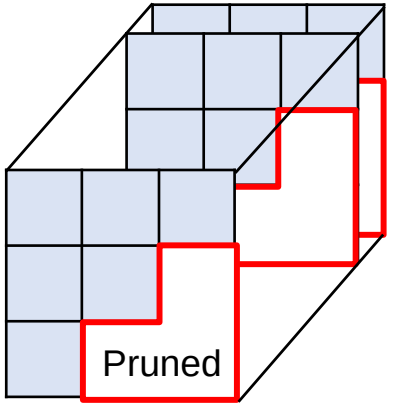
- Zero-skipping and weight pruning
  - Unstructured sparsity: CIM-unfriendly.
  - Structured sparsity: specific circuit design.



CIM-unfriendly
Unstructured sparsity



Structured sparsity



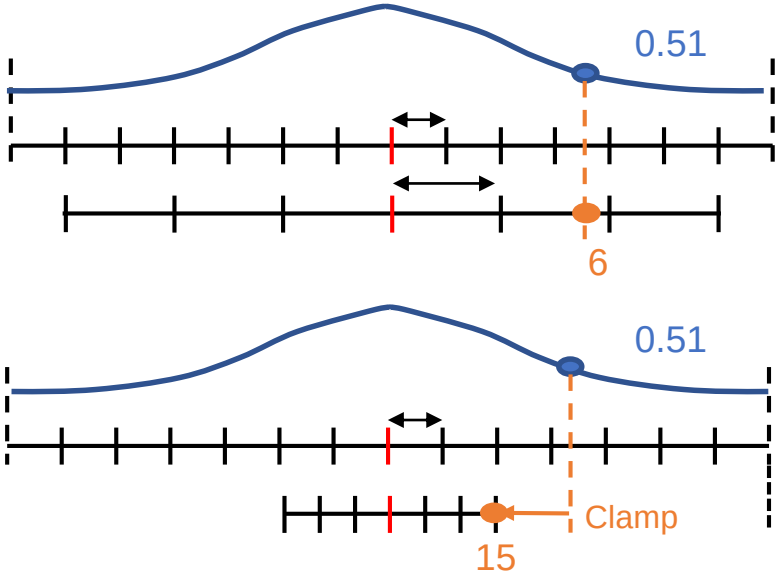
Weights pruning

Outline

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

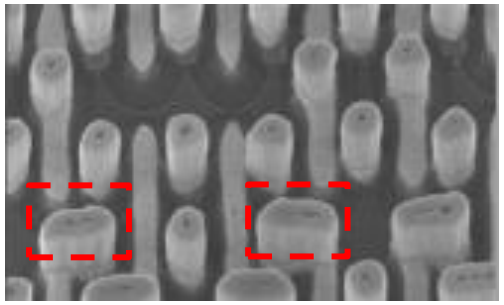
Proposed Design

- **Challenges** and **Opportunities** in Deployment of Low-bitwidth Quantized Neural Network on Compute-In-Memory (CIM):
  - **Conflicts** between **high dynamic range** and **high resolution** in low-bitwidth quantization for highly efficient inference.
  - **Noise** injected due to non-ideal factors or malicious attackers in CIM can result in significant accuracy degradation.

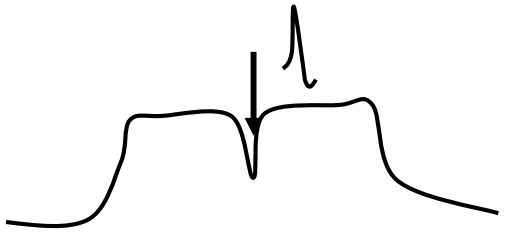


High Dynamic Range

High Resolution



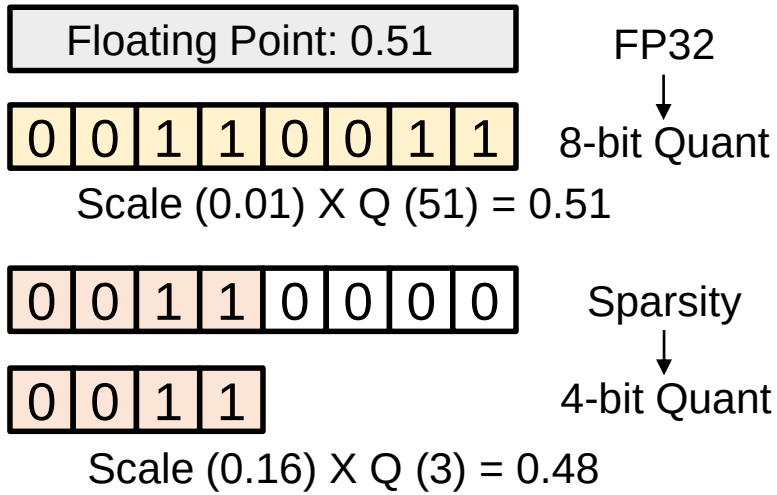
Non-ideal Factors

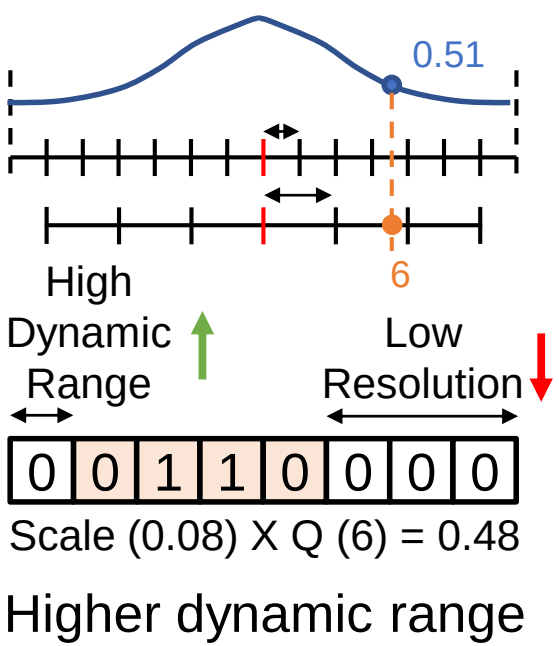


Noise Injection

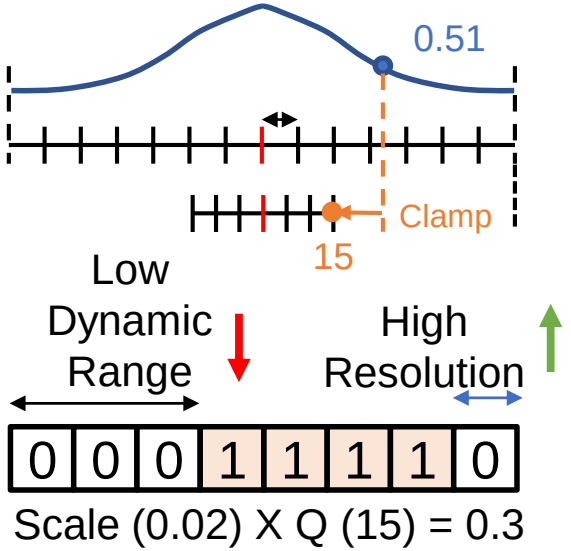
Proposed Design

- **Insight**: low-bitwidth quantization can be viewed as a form of structured sparsity by fixing specific bit locations as '0'.
- There is a **tradeoff** between the **dynamic range** and the **resolution** of quantized parameters:
  - Higher dynamic range
  - Higher resolution





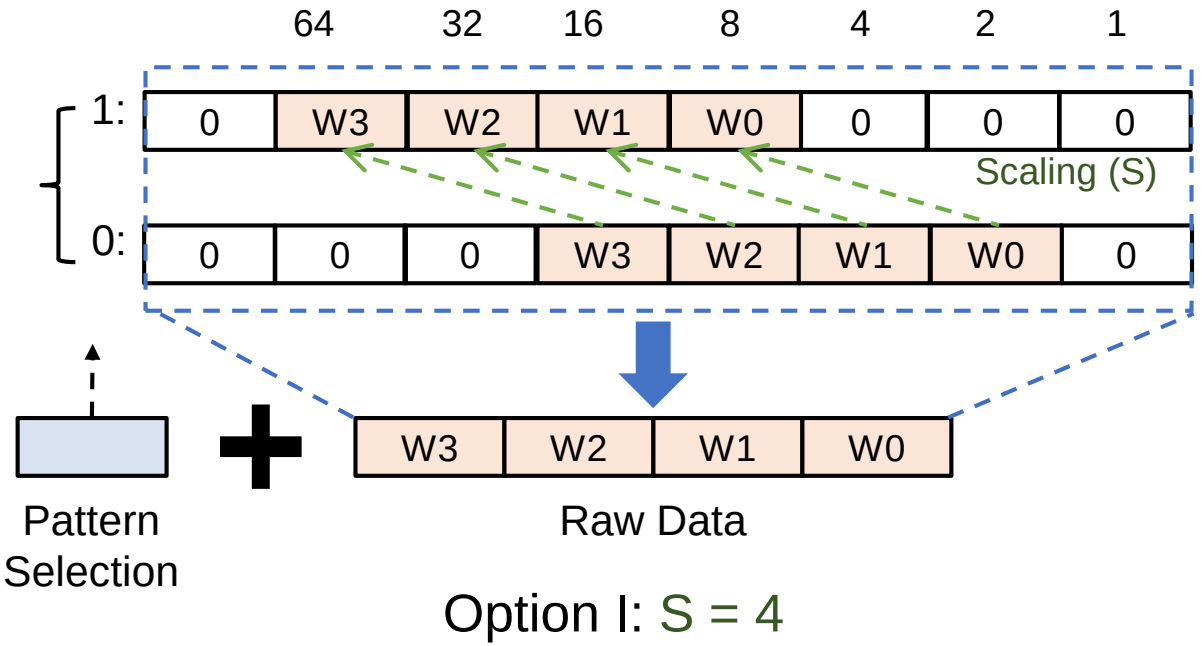
Higher dynamic range



Higher resolution

Proposed Design: Value-Adaptive Patterns

- **Option I**: Selecting between a high dynamic range pattern and a high-resolution pattern:
  - High dynamic range for large values
  - High resolution for small values



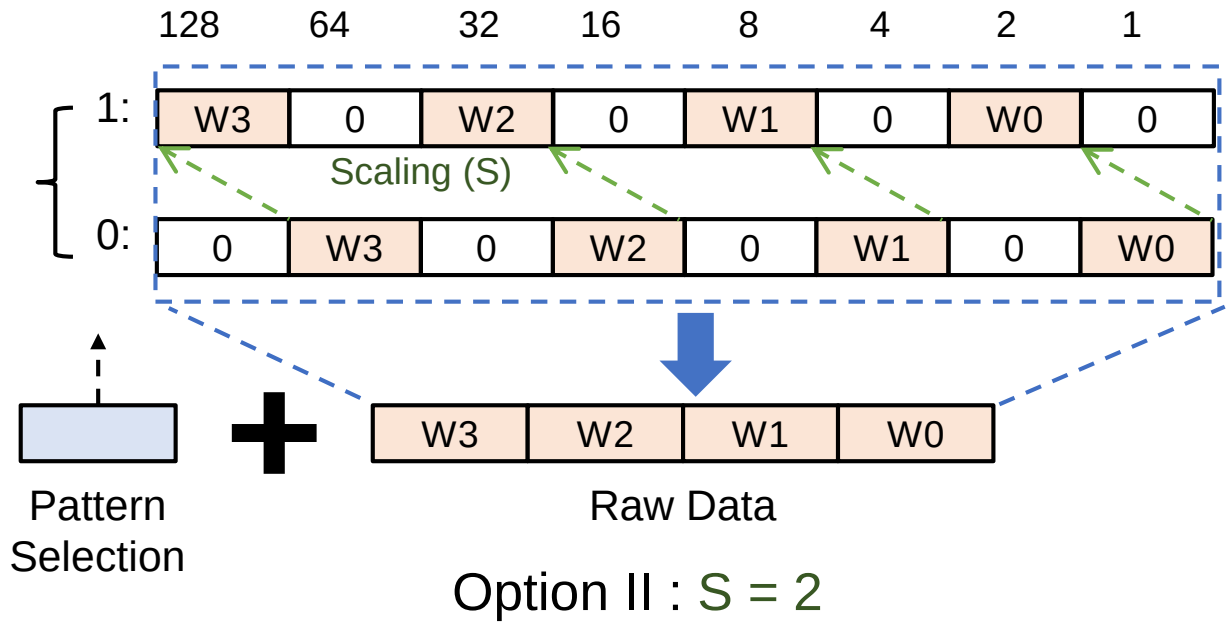
Option I: S = 4

High dynamic range pattern

High resolution pattern

Proposed Design: Value-Adaptive Patterns

- **Option II**: Alternating zero-bit patterns
  - Merge high dynamic range and high resolution in both patterns.
  - Scaling of the same bit is reduced from 4x to 2x.

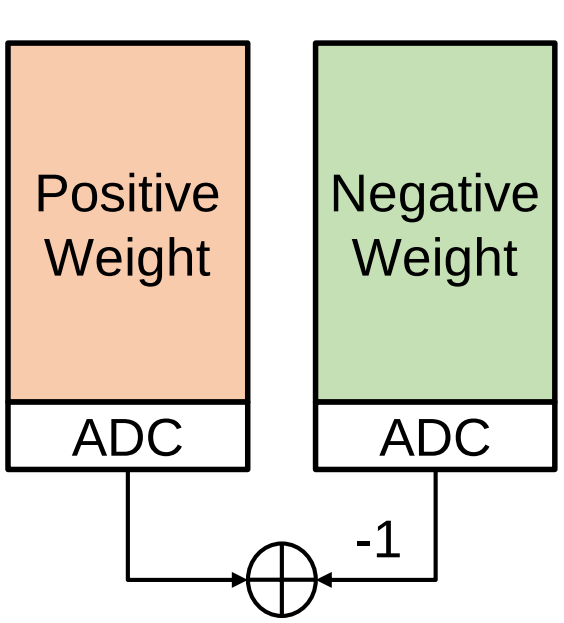


Option II : S = 2

Merge high dynamic range & high resolution in both

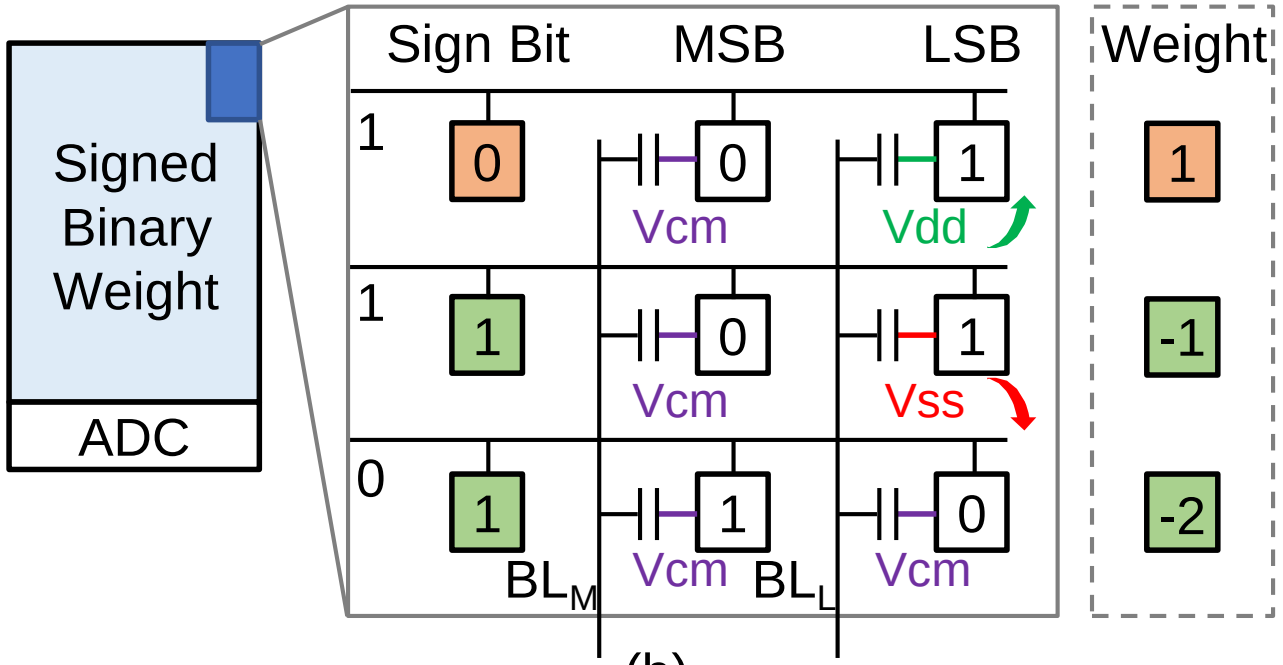
Proposed Design: Signed Binary Encoding

- Two possible implementations:
  - 2's complement encoding:
  - Signed binary encoding (**this work**):



(a)

Conventional Approach

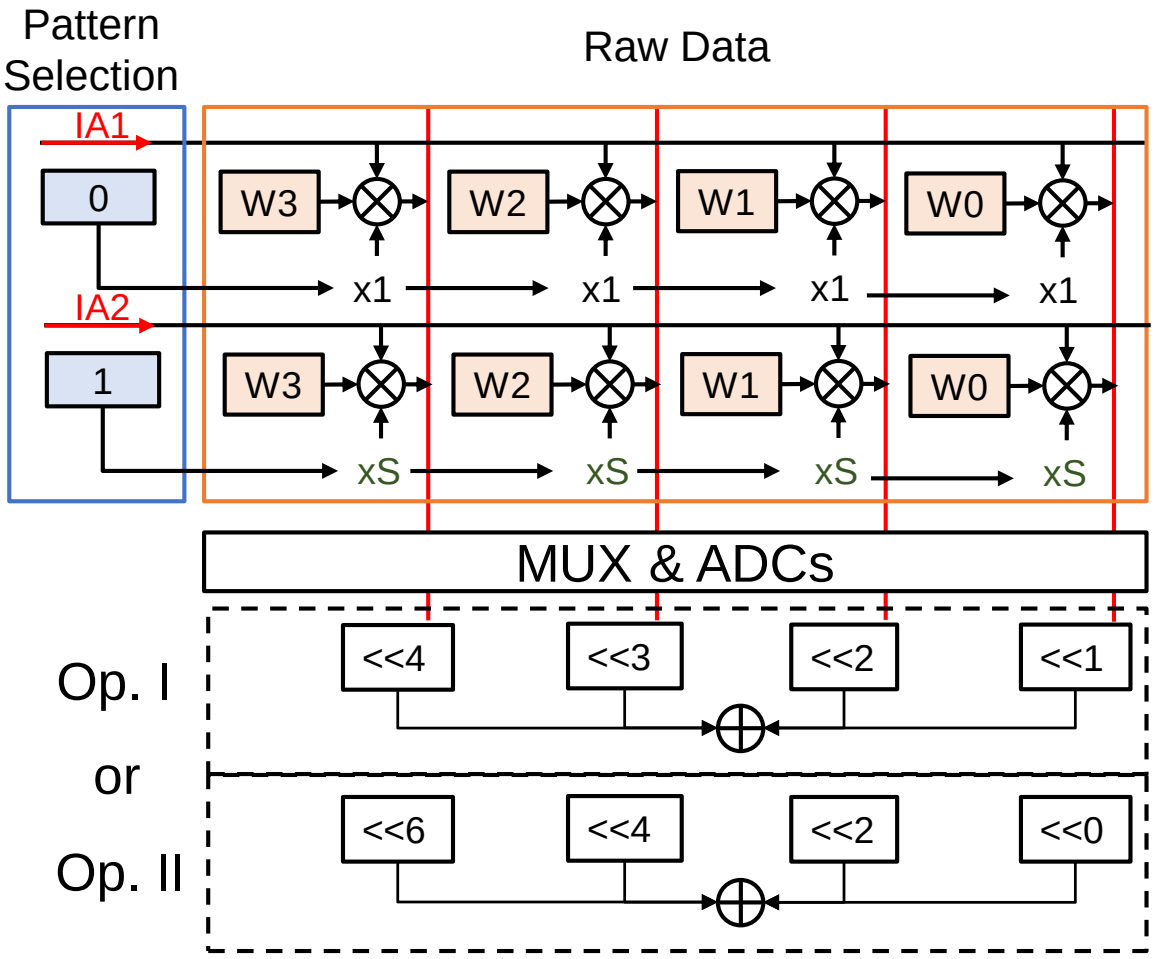


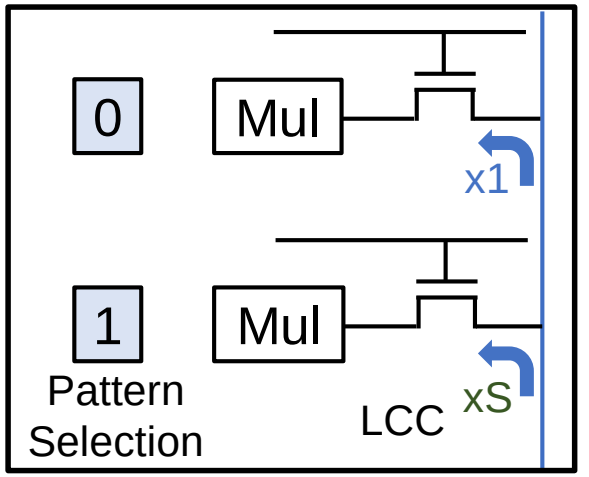
(b)

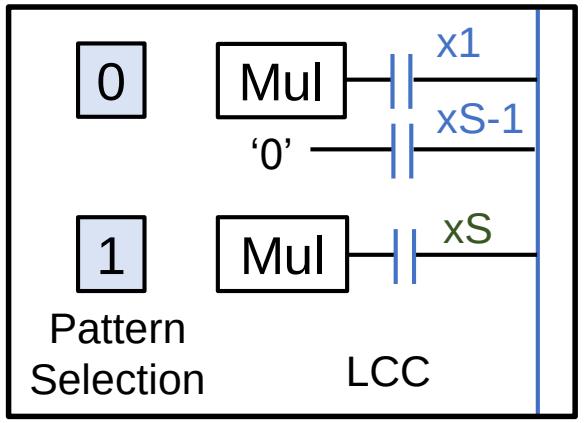
This Work

Proposed Design: Value-Adaptive Patterns

- Proposed value-adaptive sparsity patterns based on 8-bit quantization

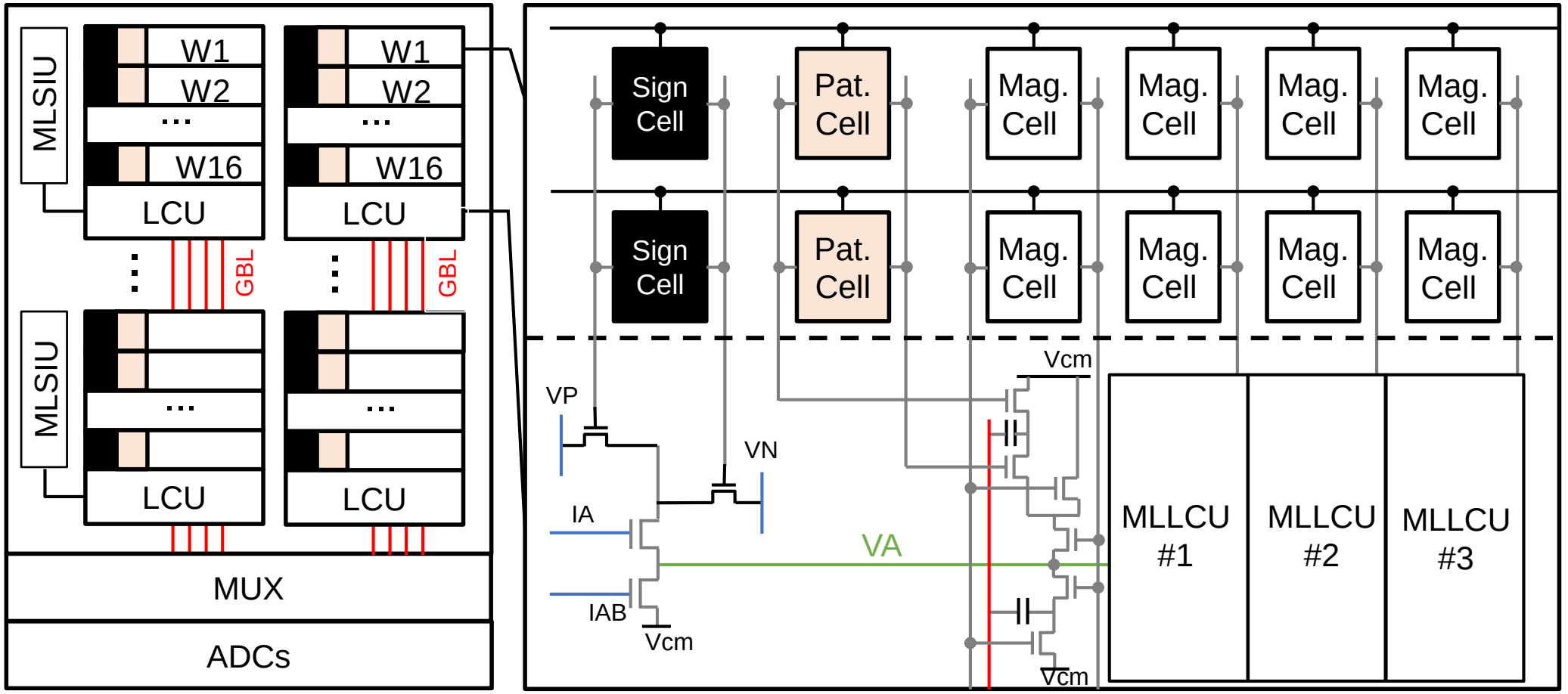






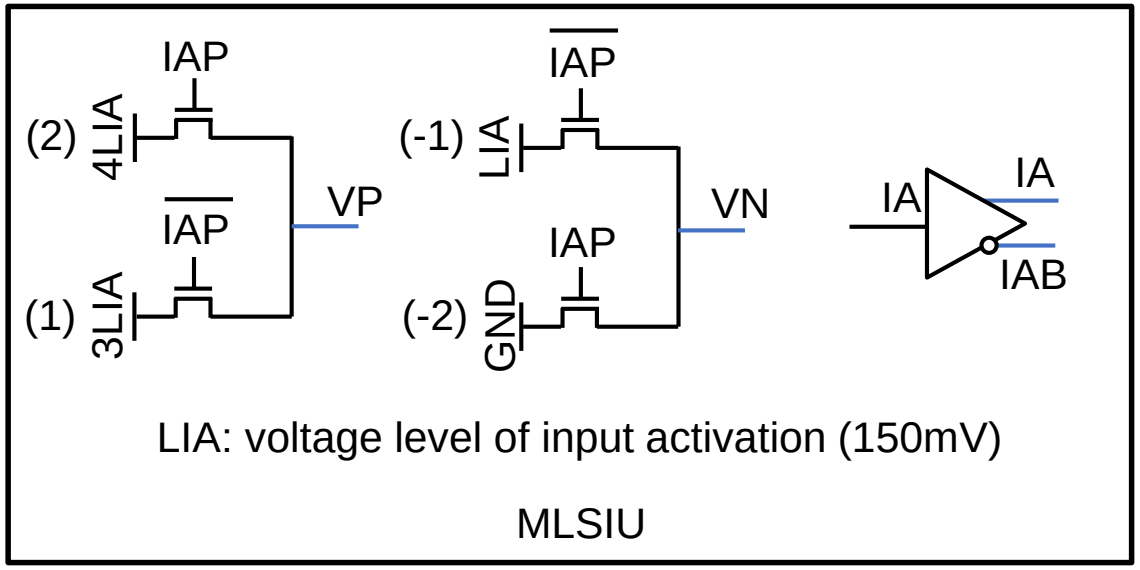
Proposed Design: Multi-Level Local Computing

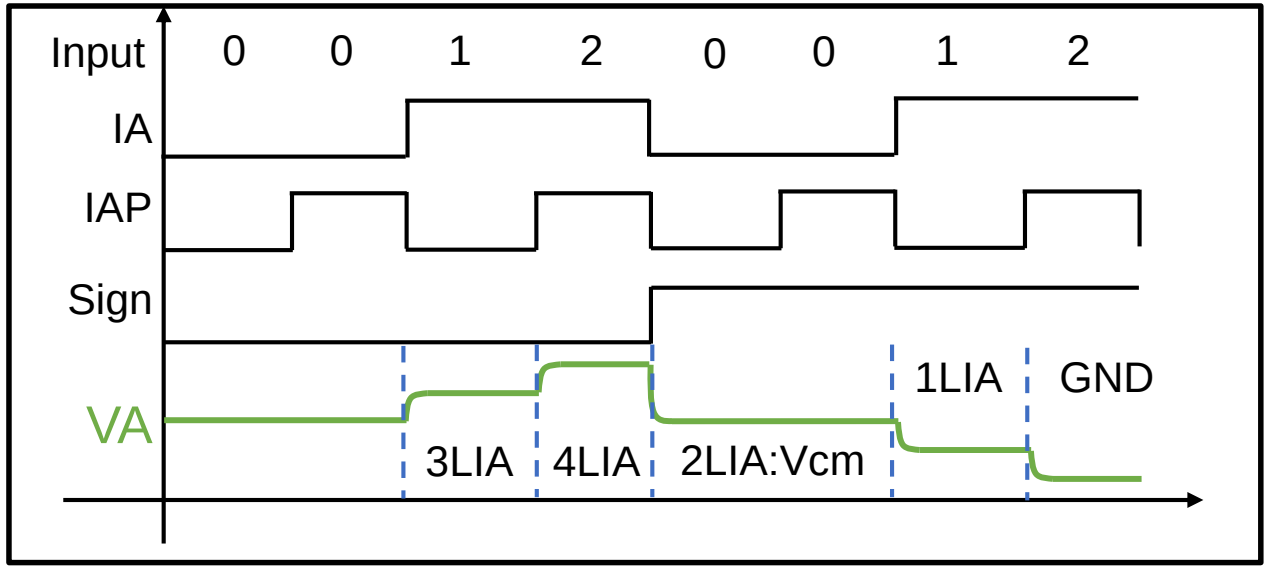
- Macrostructure:
  - Handling low-bitwidth weights of value-adaptive zero-patterns



Proposed Design: Multi-Level Local Computing

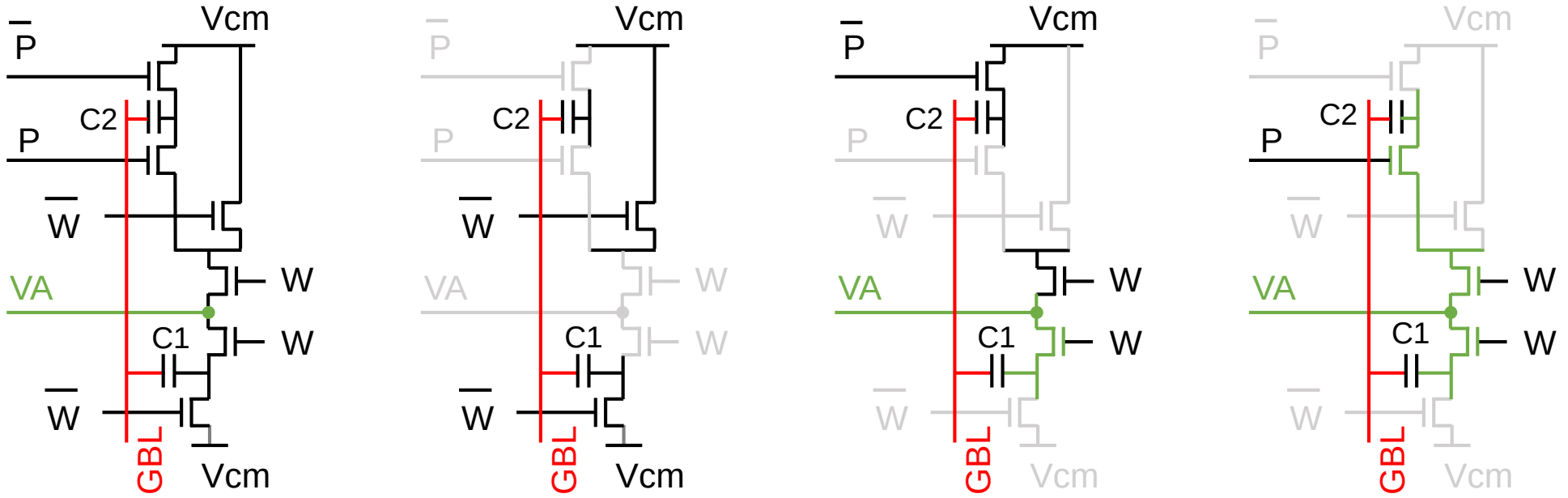
- Multi-level signed input unit circuits (**ML-SIU**):
  - Handling multi-level activations with value-adaptive zero-bit patterns.
  - Representing signed value ‘0’ with common-mode voltage.





Proposed Design: Multi-Level Local Computing

- Detailed structure and operation diagram of the proposed **ML-LCU**:
  - Weights bit (W)
  - Pattern selection bit (P): select between two different patterns
  - The partial sum result is involved with the global bit line by C1 and C2.

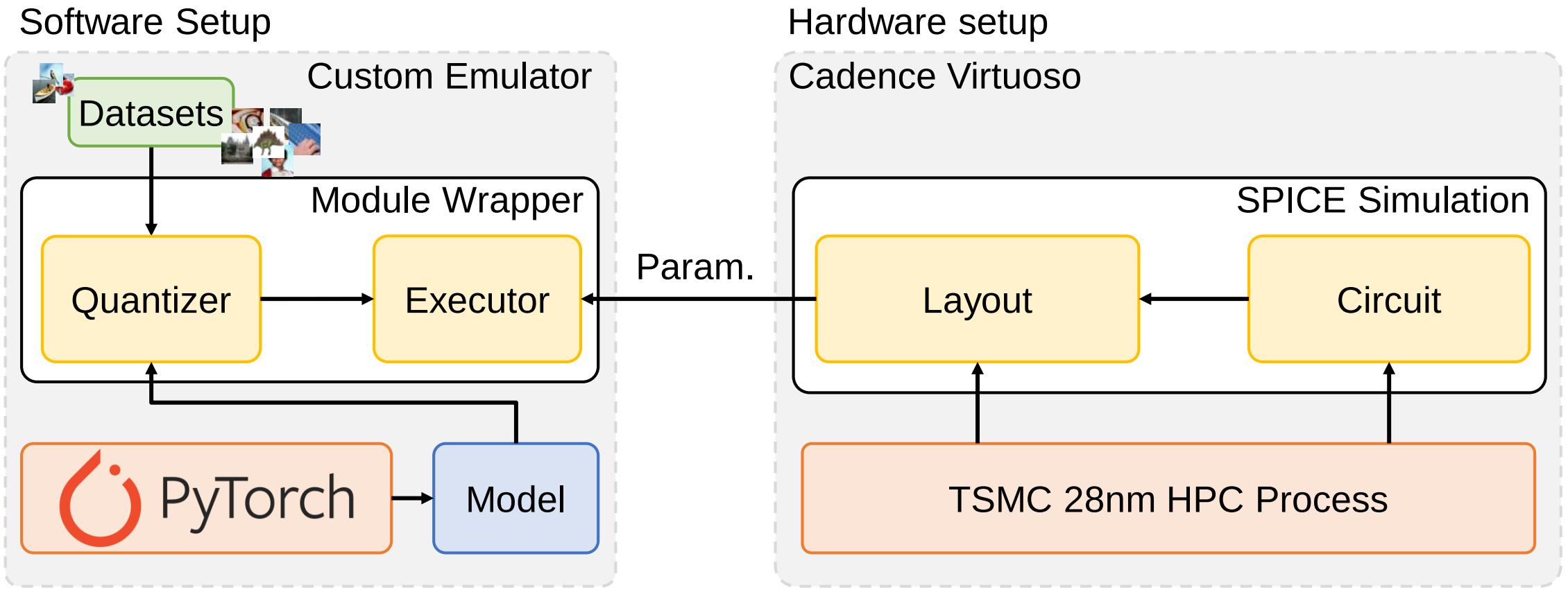


Outline

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

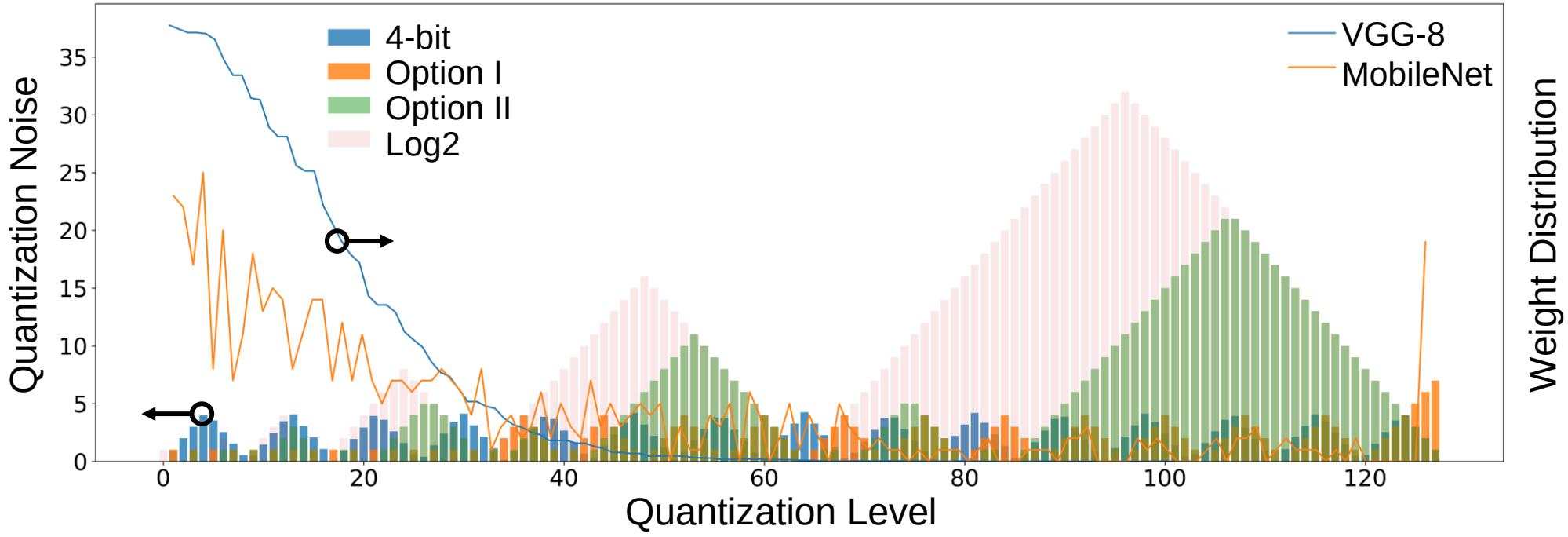
Benchmark

- Experiment setup:



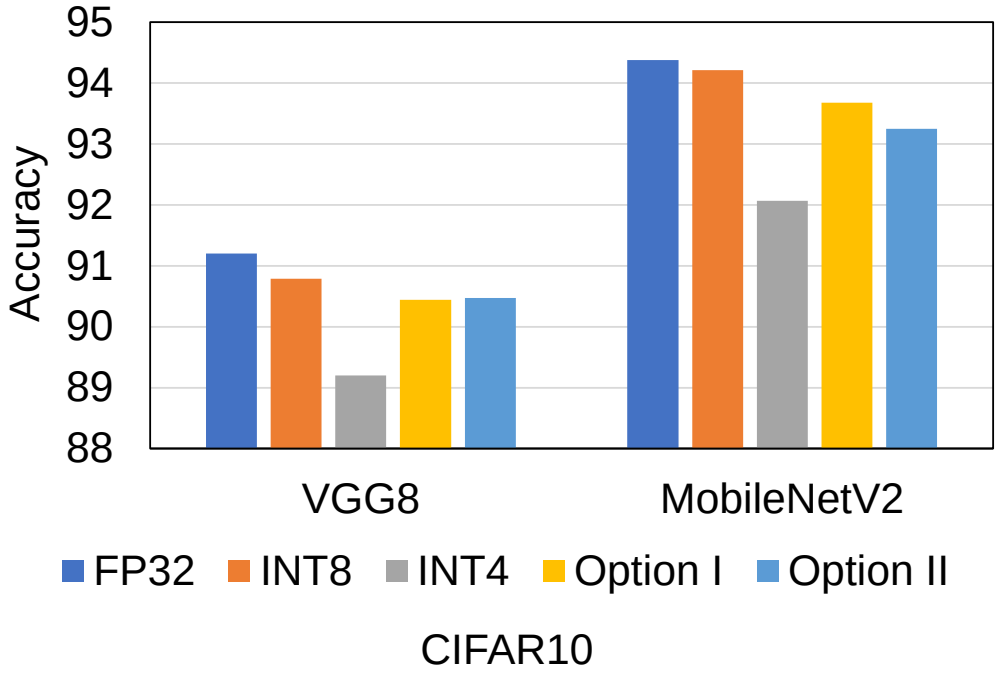
Benchmark

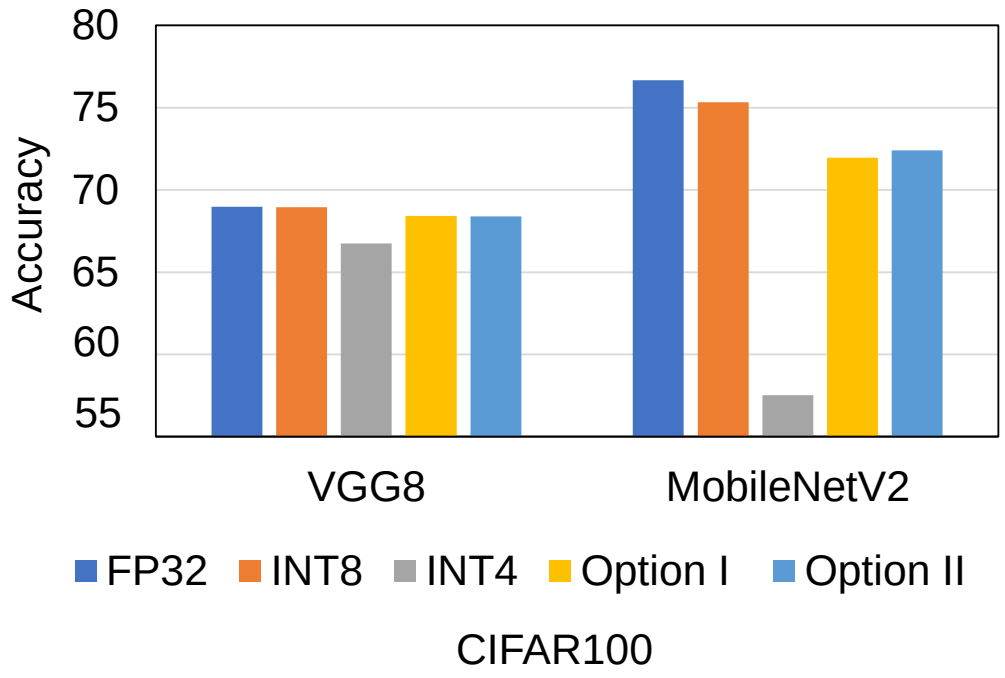
- Quantization noise analysis:
  - Typical weight distribution on VGG-8 and MobileNetV2
  - Noise introduced by:
    - 4-bit / log2
    - this work



Benchmark

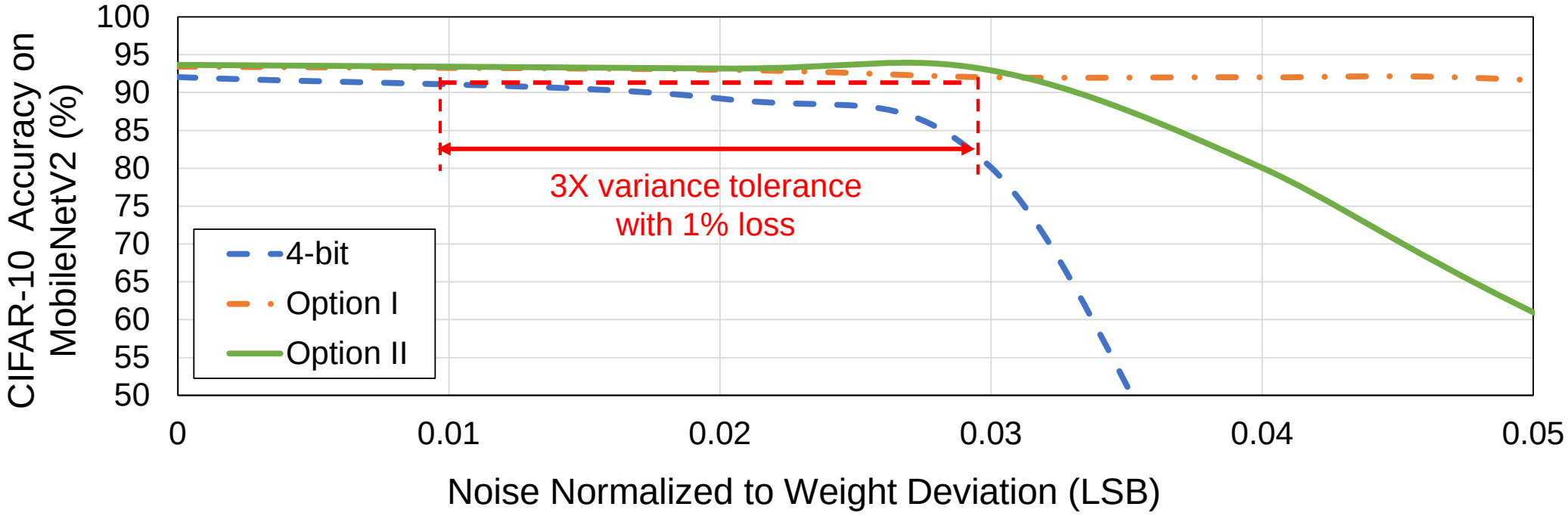
- Accuracy comparison on various datasets and models:
  - Achieves <1% accuracy drop in VGG-8 and <3% accuracy drop in MobileNetV2.
  - Compared with ~3x accuracy loss of the same bitwidth 4-bit quantization.





Benchmark

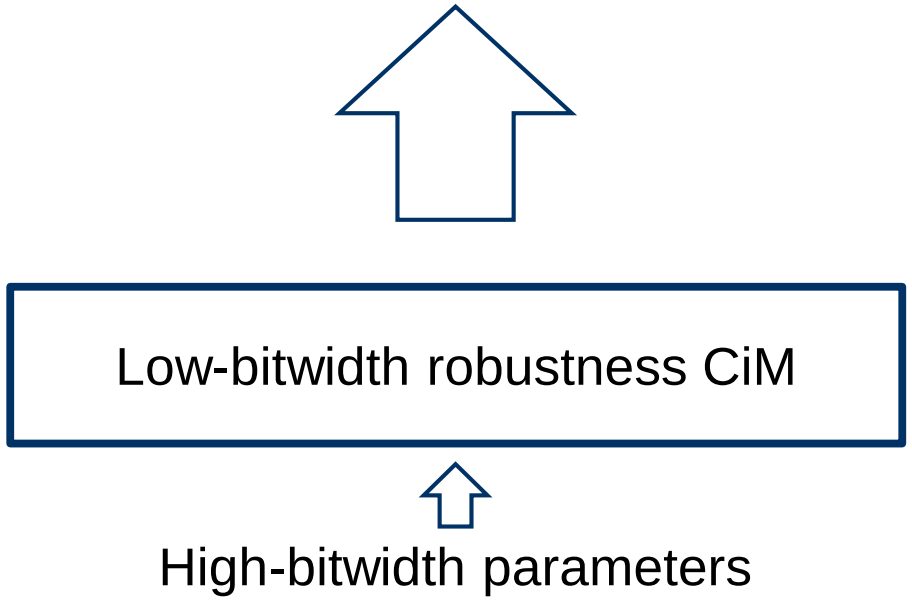
- Insights provided by robustness study:
  - Low-bitwidth quantization is more sensitive to disturbances.
  - Shrinking the dynamic range of quantization results in more outliers



Robustness enhancement on CIFAR-10 with MobileNet (4-8-8)

Benchmark

- Macro-level performance comparison
  - 20% higher parameter density
  - 3.1x higher energy efficiency
  - 2.9x higher area efficiency



Metrics	Charge-domain CIM Methodologies			
	Baseline[a]		*ZEBRA Option I*	*ZEBRA Option II*
Process	28 nm CMOS			
Area of LCC (μm^2)	1.04		3.58	2.17
Macro area (mm^2)	0.32		0.36	0.34
Precision (I-W)	4-4	8-8	8-8	8-8
Parameter density (M/mm^2)	0.97	0.50	0.58	0.61
Energy efficiency[b] (TOPS/W)	65.4	15.9	64.3	64.6
Area efficiency[b] ($GOPS/mm^2$)	390	95.25	346	367
Accuracy[c] under noise	80.1%	94.2%	93.7%	93.3%
SNR Robustness	No	Yes	Yes	Yes
VGG-8 support	😐	☺	☺	☺
MobileNet support	☹	☺	☺	☺
Robustness	☹	☺	☺	☺

[a.] Charge-domain local computing cell with 16 SRAM rows [18], [26]

[b.] OP is defined with the precision of the same column

[c.] Accuracy is obtained based on MobileNetV2 (4-8) under a variance of 0.03LSB

Outline

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

Discussion

- Compared with the original 4-bit quantization.
  - Area overhead: 37%
    - 2 extra bits: pattern selection and sign bits
    - ML-LCU and ML-SIU
  - Energy efficiency overhead: 2%
    - 1.8~1.9x ML-LCU and ML-SIU energy overhead
- Future work: Reconfigurable ZEBRA:
  - Multiple options
  - Hybrid encoding scheme

Conclusion

- Proposed ZEBRA architecture
  - High-robustness charge-domain CiM architecture
  - Value-adaptive zero-bit patterns
  - Dealing with the bottleneck of deploying a low-redundancy NN on analog CIM.
- Features:
  - Capability of tolerating noise due to mismatches, non-ideal interfaces, noise, etc.
  - 2.9X higher area efficiency and 3.1X higher energy efficiency.

Thank You

Yiming Chen[1], Guodong Yin[1], Hongtao Zhong[1] , Mingyen Lee[1] , Huazhong Yang[1] , Sumitha George[2] , Vijaykrishnan Narayanan[3], and **Xueqing Li[1†]**

[1]Tsinghua University, [2]North Dakota State University, [3]Pennsylvania State University

†Email: xueqingli@tsinghua.edu.cn