BFP-CIM: Data-Free Quantization with Dynamic Block-Floating-Point Arithmetic for Energy-Efficient Computing-In-Memory-based Accelerator

ASP-DAC 2024

*<u>**Cheng-Yang Chang**</u>, Chi-Tse Huang, Yu-Chuan Chuang, Kuang-Chao Chou, An-Yeu (Andy) Wu

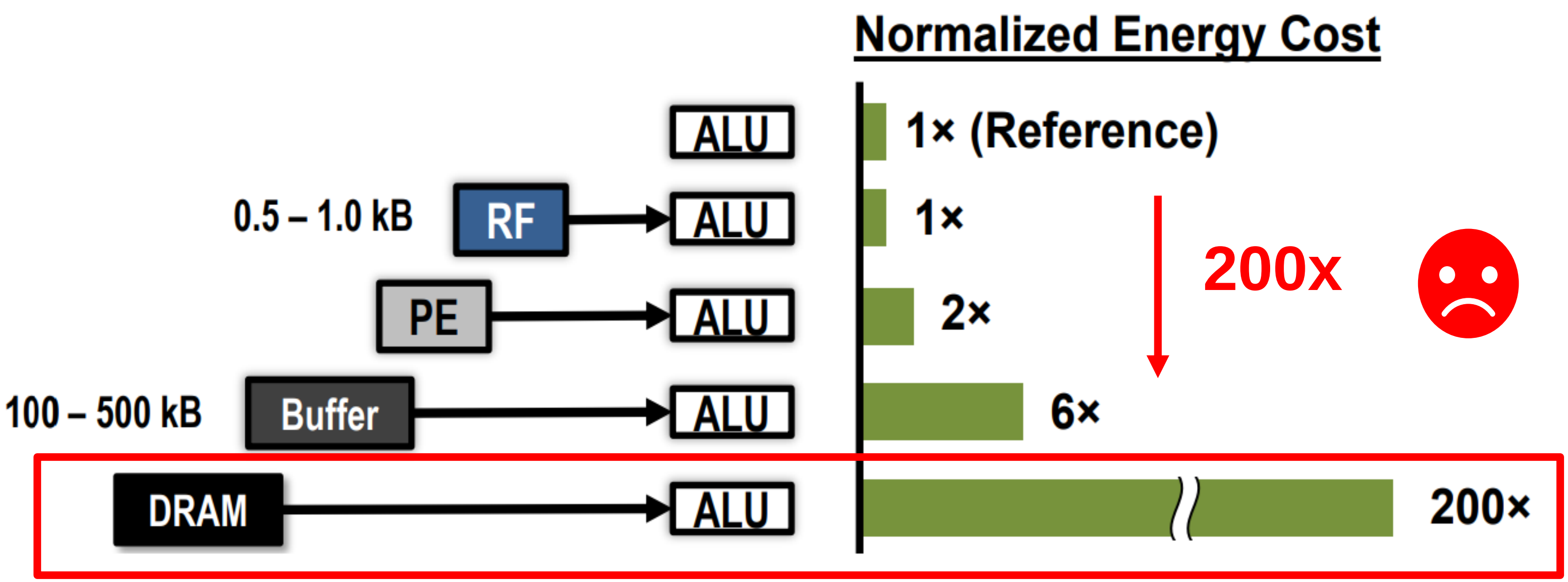
Graduate Institute of Electronics Engineering, National Taiwan University

Outline

- Background
- Introduction of Computing-in-Memory (CIM)
  - Related Works
- Proposed BFP-CIM Technique
  - Motivational Experiments
  - Nonlinear Quantization with Range-aware Rounding (RAR)
  - Dynamic Block-Floating-Point Arithmetic (BFP)
- Experiment
- Summary

Background – The Memory Wall

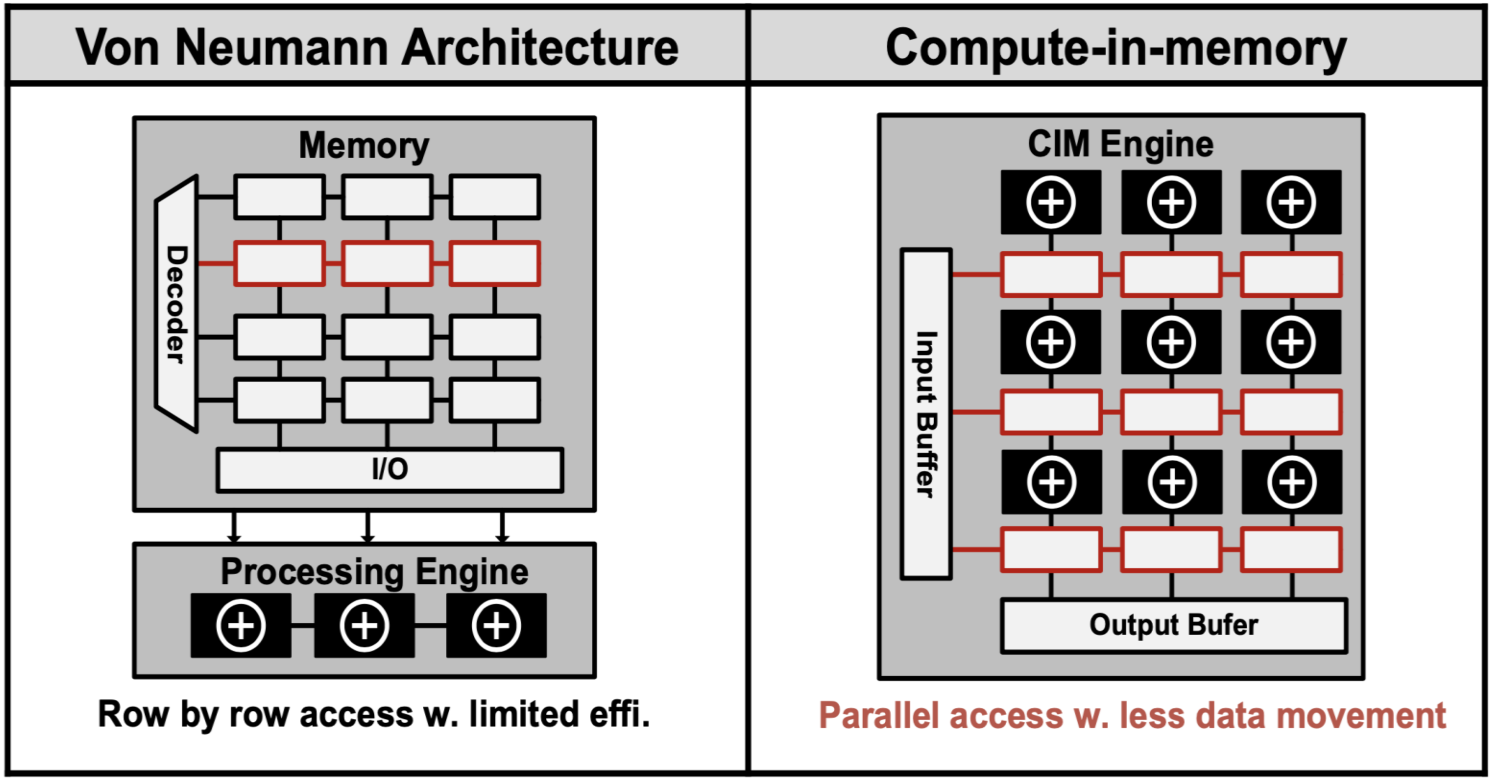
- **Von Neumann Architecture**
  - Separated memory and processing elements (PEs)
  - **Data movement** consumes more energy than computing
- DNNs require excessive energy consumption due to **<u>memory access</u>**



[Chen, ISCA'16]

Introduction of CIM

- CIM <u>embeds computation in memory</u>, thus reducing data movement and providing large bandwidth
  - **No explicit data read**

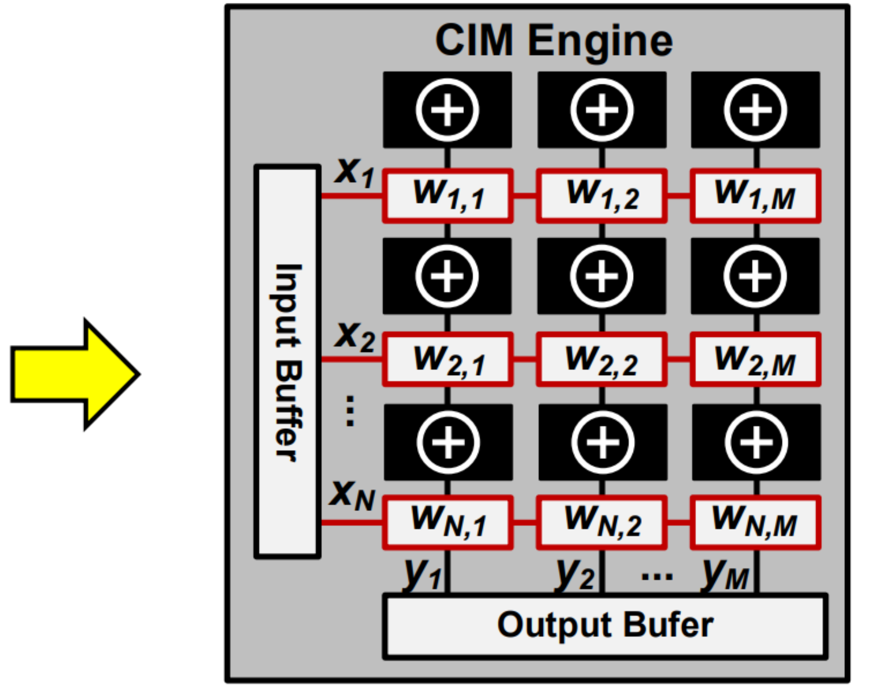


[Y. He, ISSCC'23 7.3]

Introduction of CIM (Cont'd)

- CIM-based **matrix-vector-multiplication (MVM)**
  - Extreme spatial architecture with 2-D data reuse

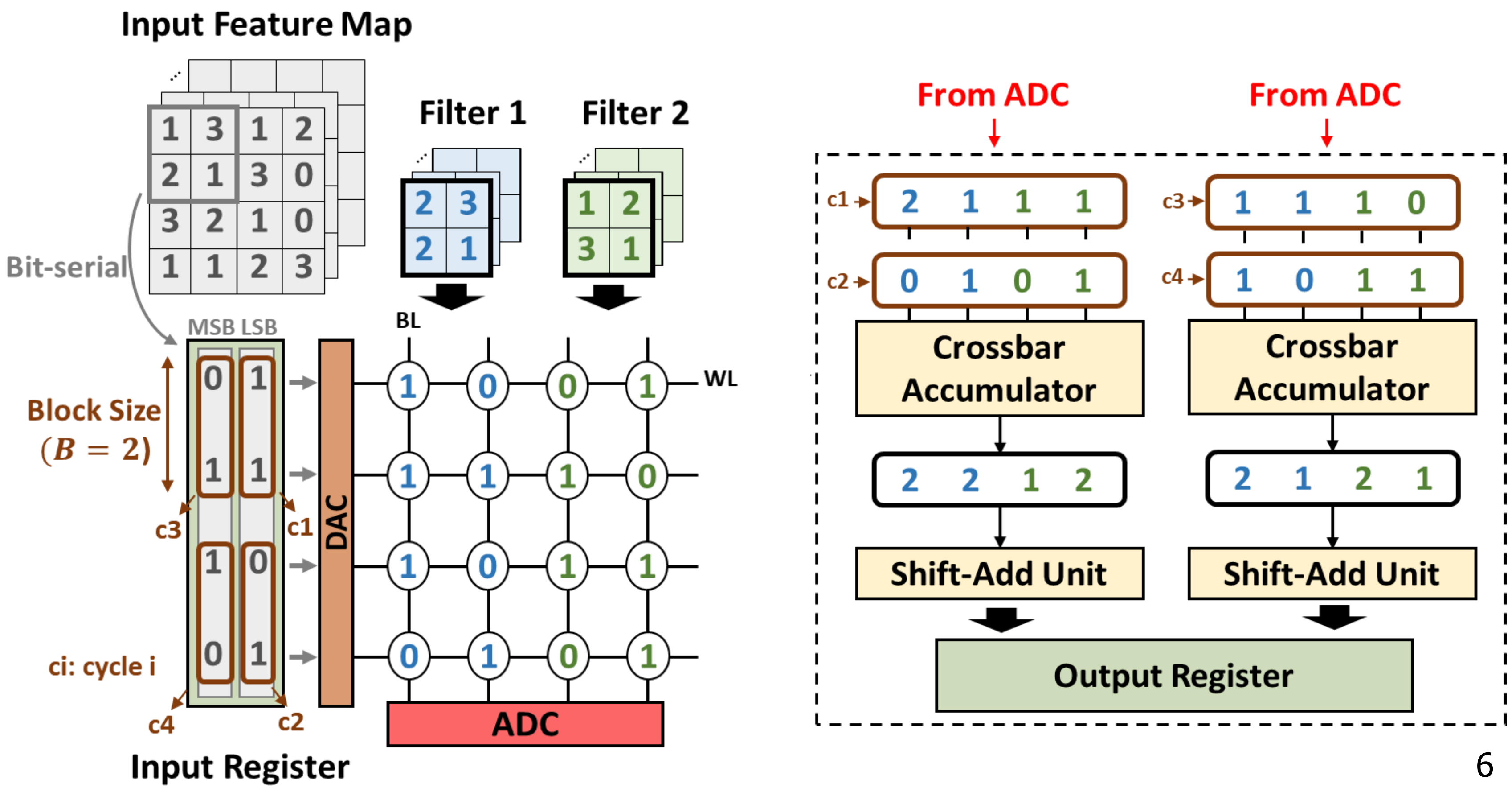
$$\begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} w_{1,1} & \cdots & w_{1,N} \\ \vdots & \ddots & \vdots \\ w_{M,1} & \cdots & w_{M,N} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$



[Y. He, ISSCC'23 7.3]

Introduction of CIM (Cont'd)

- Mapping DNN workloads to CIM architecture
  - In most prior works, feature map and filter weights are **vectorized** and computed in a <u>bit-serial</u> manner

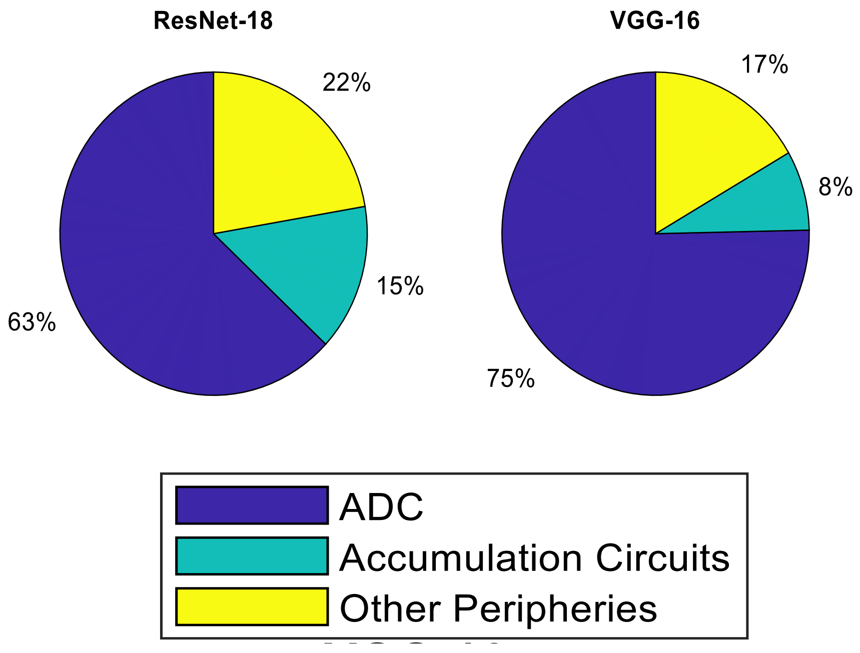


ADC Overhead of CIM (Cont'd)

- CIM energy breakdown using **DNN+NeuroSim**
  - ADCs account for **>60%** energy consumption

[X. Peng, IEDM'20]

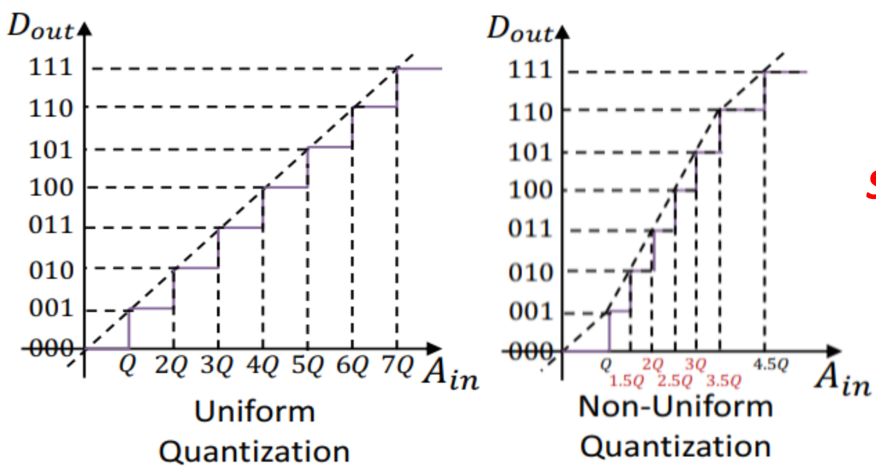
Simulation Settings	
Dataset	**CIFAR-10 / ImageNet**
Model	**ResNet18 / VGG-16**
Device	**ReRAM**
Array Size	**128×128**
Block Size	**16**
ADC	**4-bit**



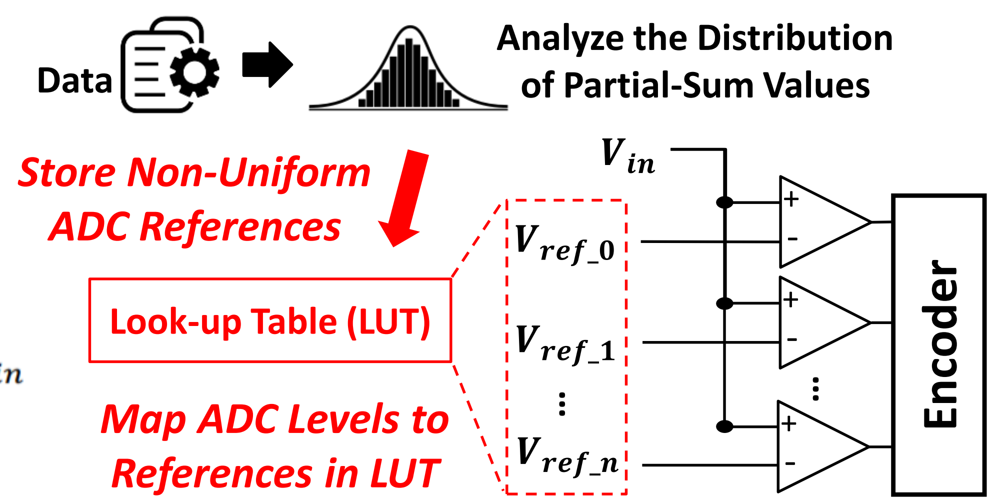
Quantization for CIM: Lower the number of ADC access by reducing the **bit-width of input activations**

Related Works – CIM Quantization

- Non-uniform Quantization
  - Rely on calibration data
  - Require ADCs with **<u>non-uniform</u>** reference voltage levels

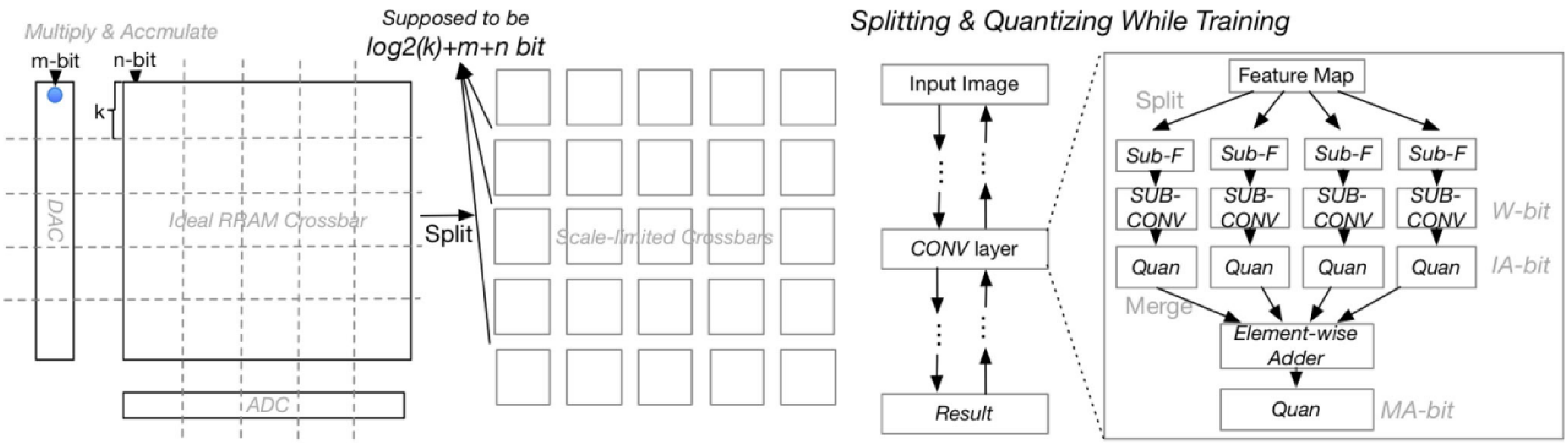


[H. Sun, ASP-DAC'20]



Related Works – CIM Quantization

- Quantization-aware Training
  - The training/fine-tuning process is time-consuming

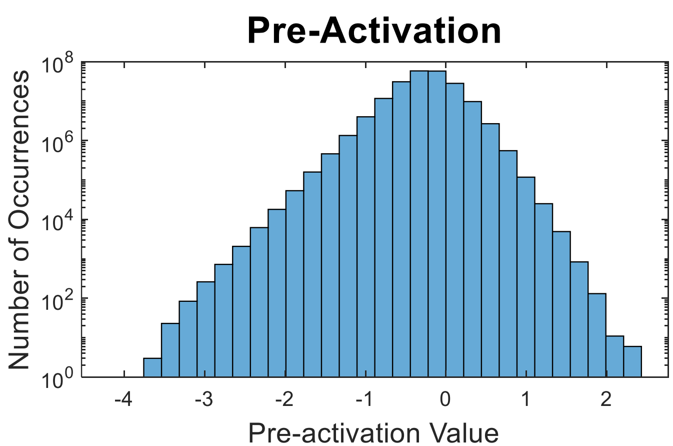


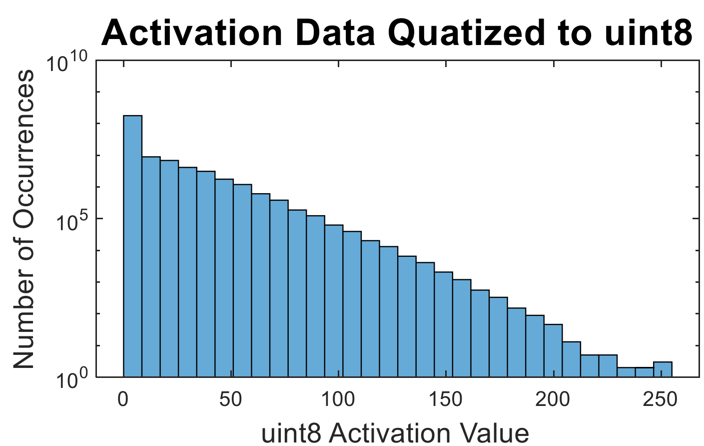
[Y. Cai, TCAD'19]

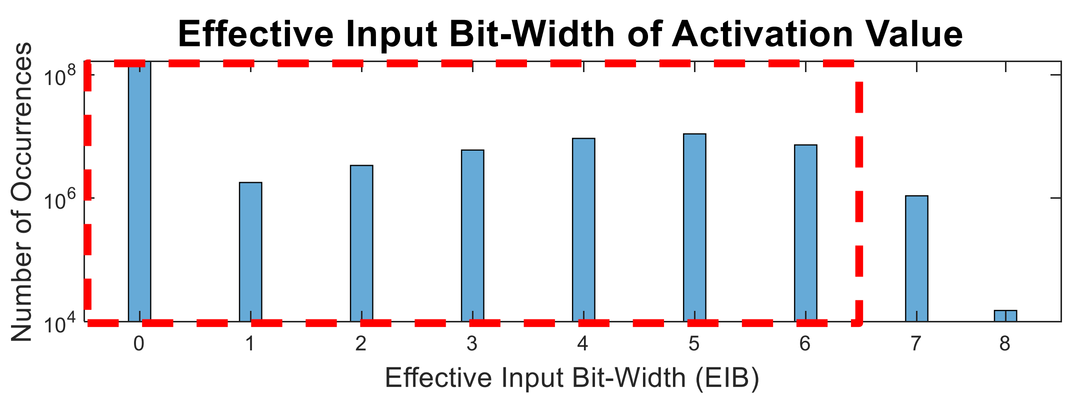
Goal: CIM quantization ***without*** data-dependent fine-tuning/training and non-uniform ADC

Motivational Experiments

- After ReLU and uint8 quantization, the **effective input bit-width (*EIB*)** for most input values are $\leq \mathbf{6}$
  - $EIB(n) = \lfloor log_2(n) \rfloor + 1, \forall\, n > 0$

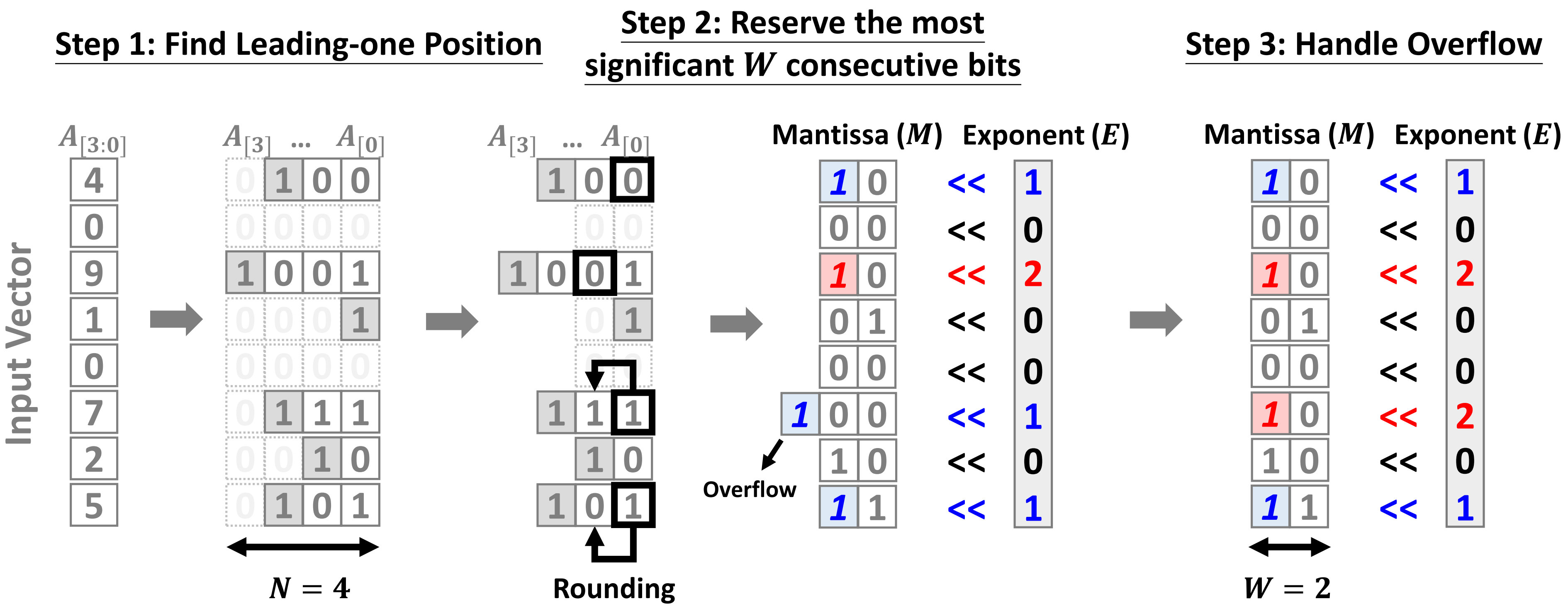






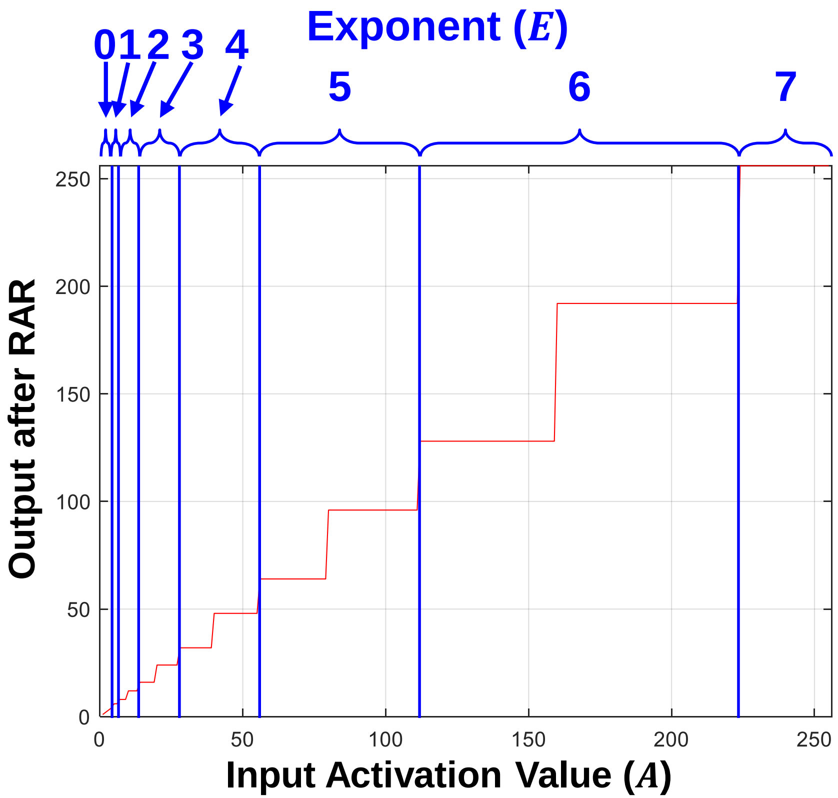
Proposed Range-aware Rounding (RAR) for Nonlinear Quantization

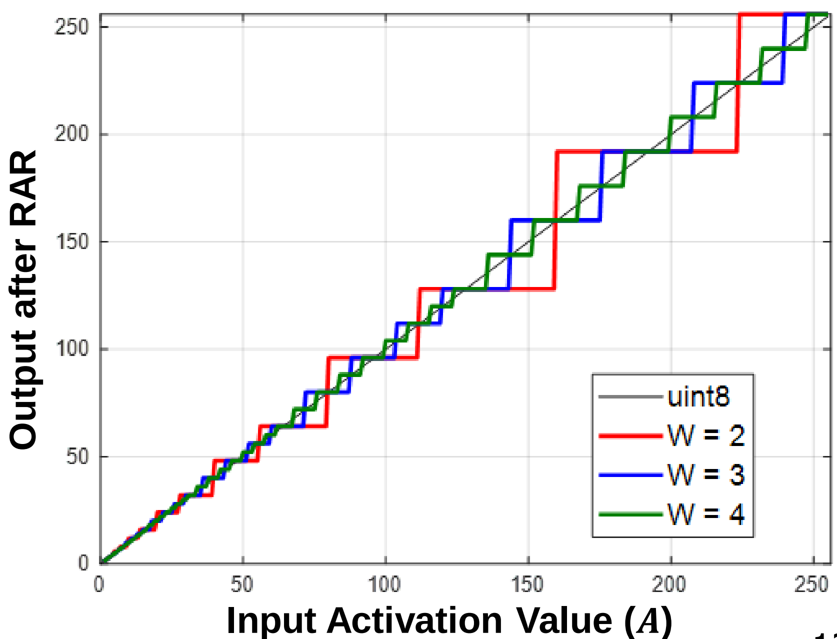
- Convert N-bit <u>fixed-point</u> data into <u>**FP-like**</u> values with short mantissa bit-width (window size, $\boldsymbol{W}$)
  - **The first toggle bit** determines the exponent value



Proposed Range-aware Rounding (RAR) for Nonlinear Quantization (Cont'd)

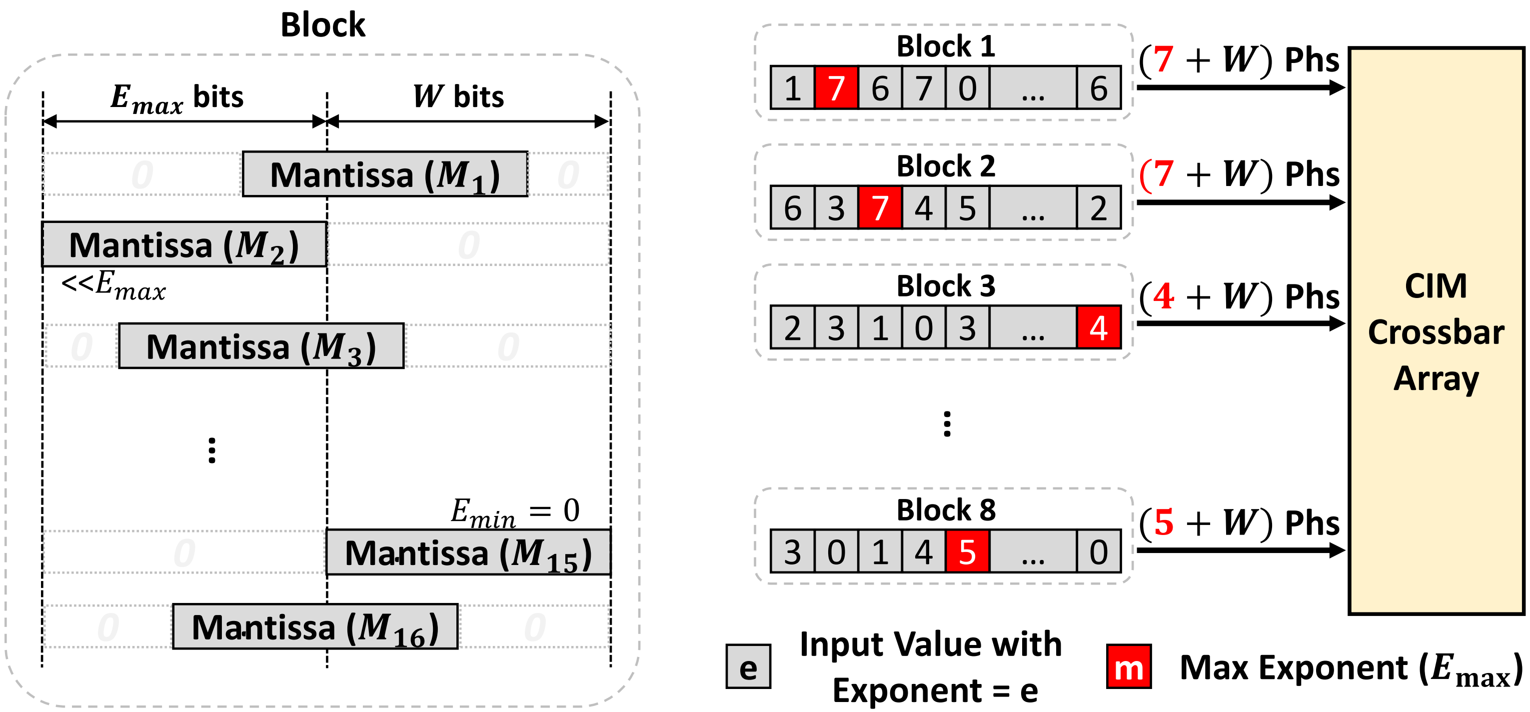
- RAR **non-linearly** divides the distribution of values according to their exponents
  - Values within the same region are **linearly** quantized





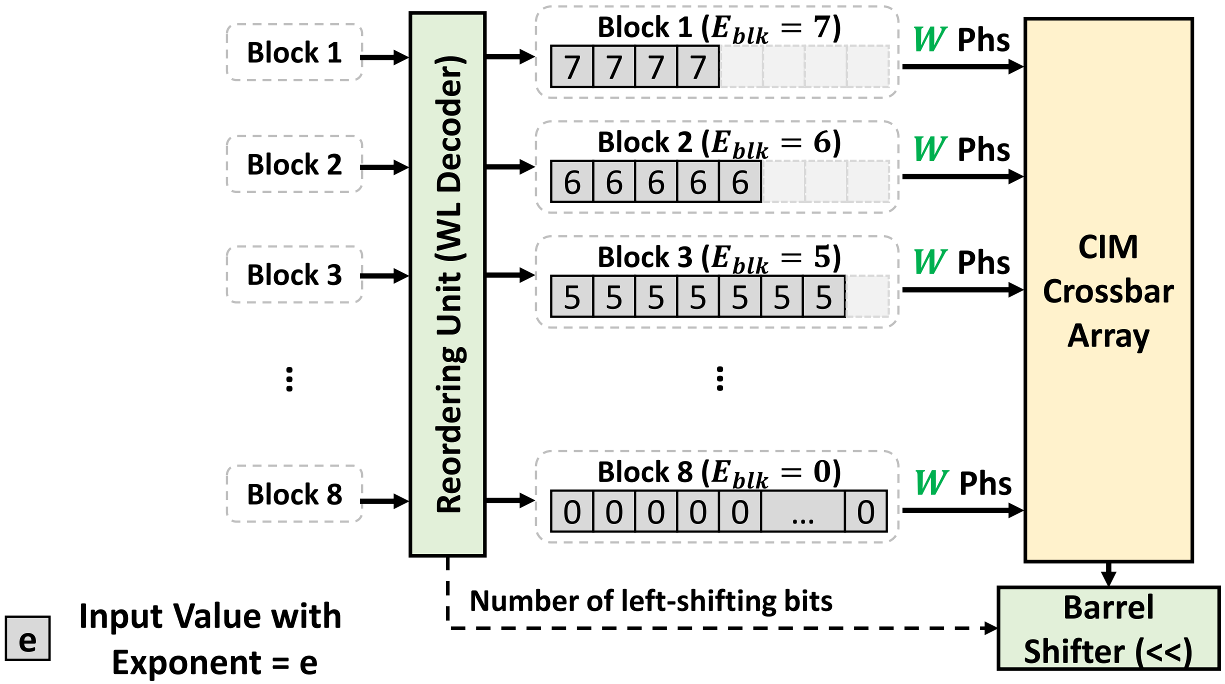
Proposed Dynamic Block-Floating-Point Arithmetic (BFP) for Integrating RAR

- BFP format offers a middle ground between the FP MAC and integer MAC
  - **Exponent pre-alignment** requires **shifting**, thus mitigating the benefit of RAR



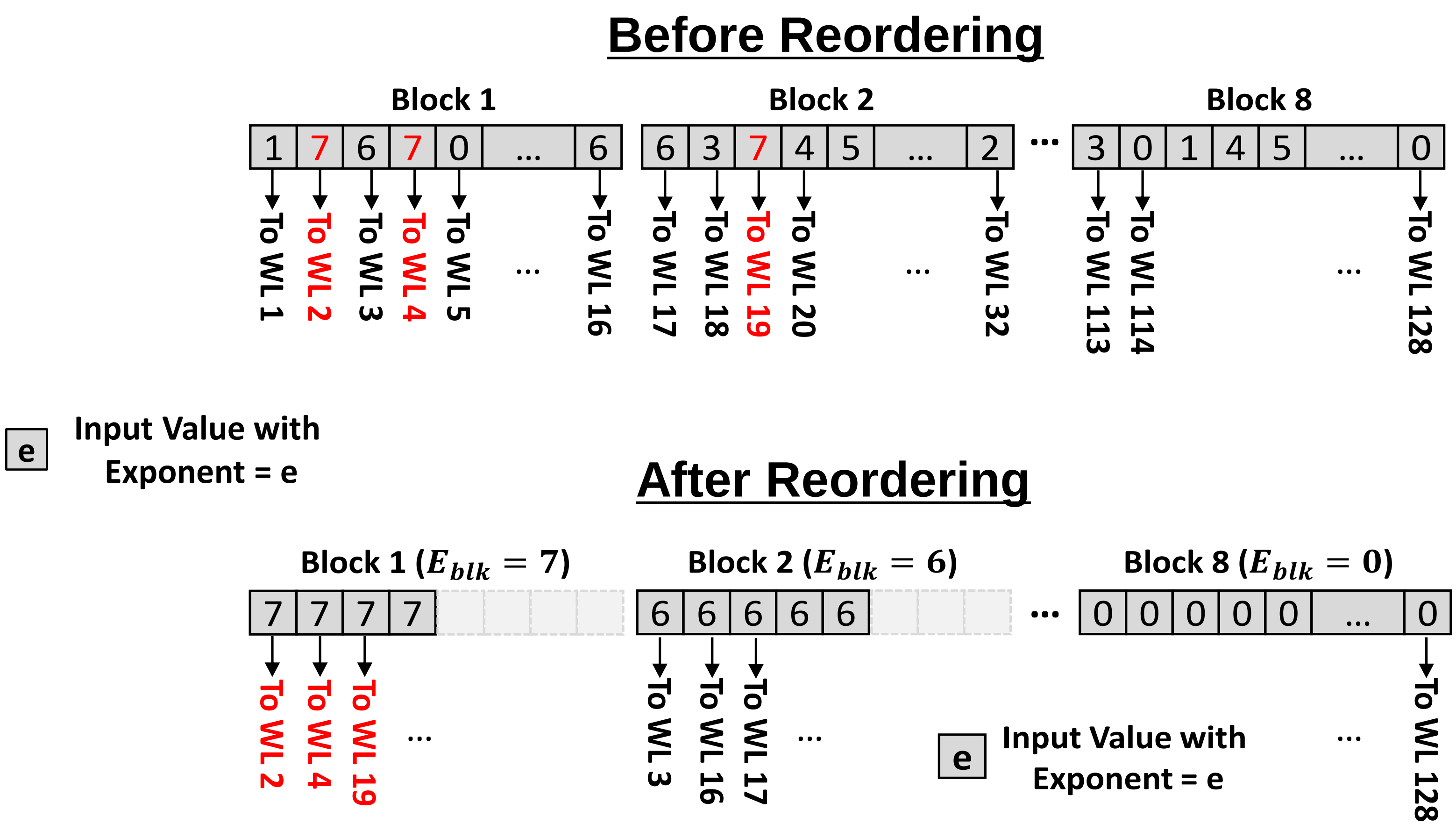
Proposed Dynamic Block-Floating-Point Arithmetic (BFP) for Integrating RAR

- Reorder inputs according to the exponents after RAR
  - Selectively turn on multiple WLs that **share the same exponent (E_{blk})** with **time-interleaved** computation



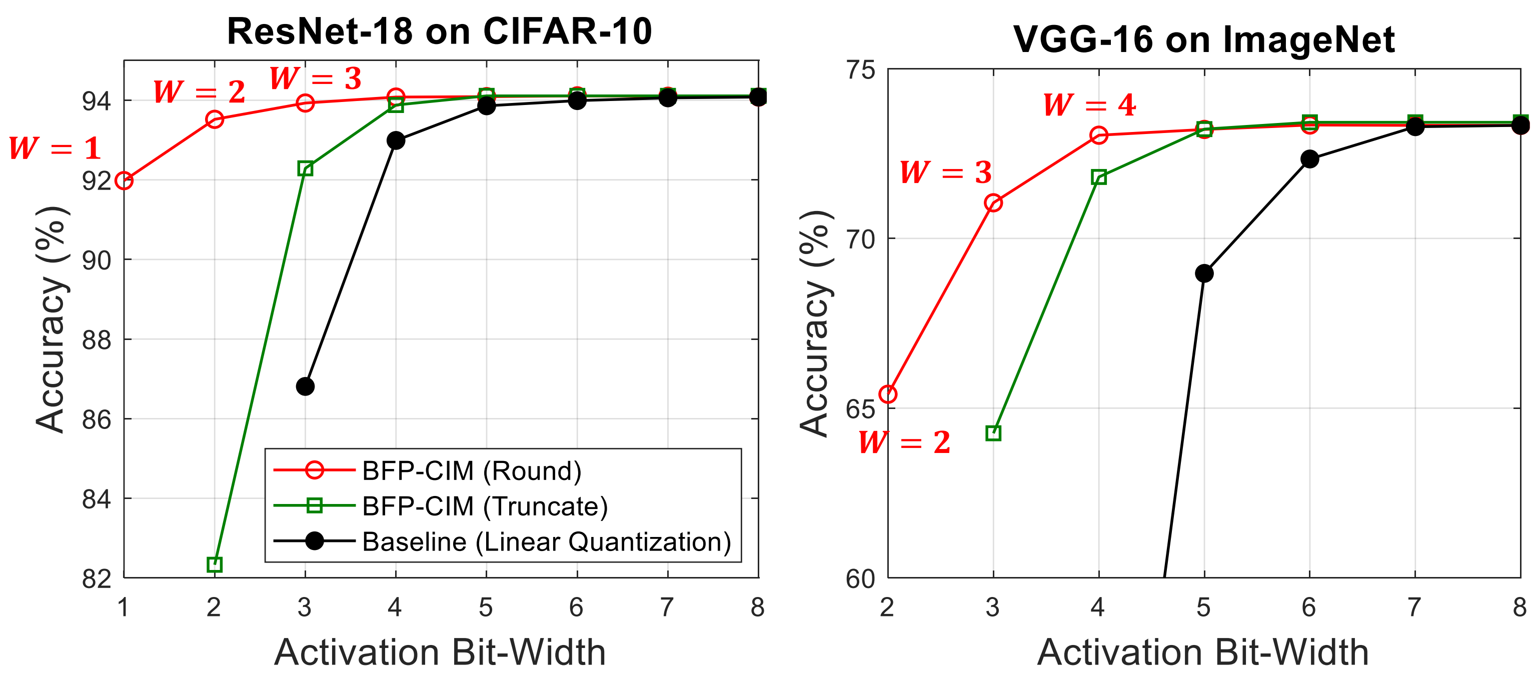
Proposed Dynamic Block-Floating-Point Arithmetic (BFP) for Integrating RAR

- **Not** explicitly shuffle the input values
  - Each value is still connected to its corresponding WL



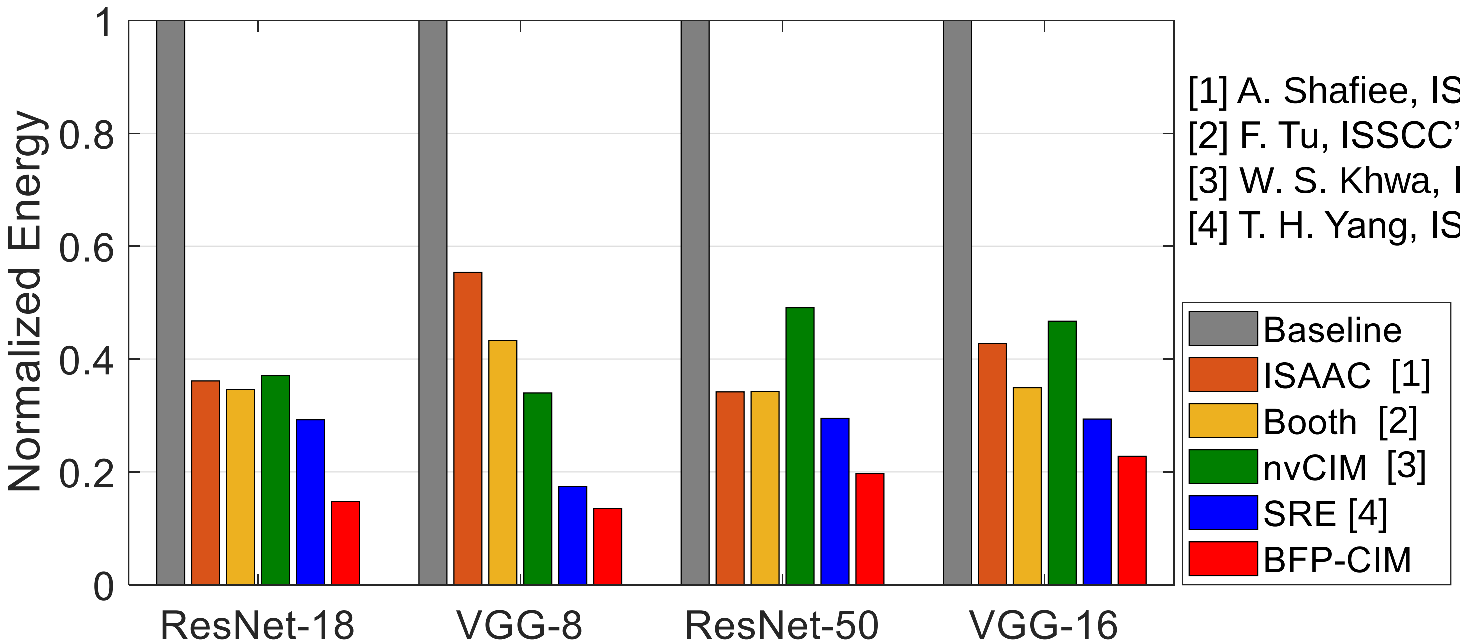
Experiment – Precision Scalability

- <u>**BFP-CIM (Round)**</u> can maintain accuracy while reducing the activation bit-width, i.e., the **window size (W)**, to three bits



Experiment – Energy Efficiency

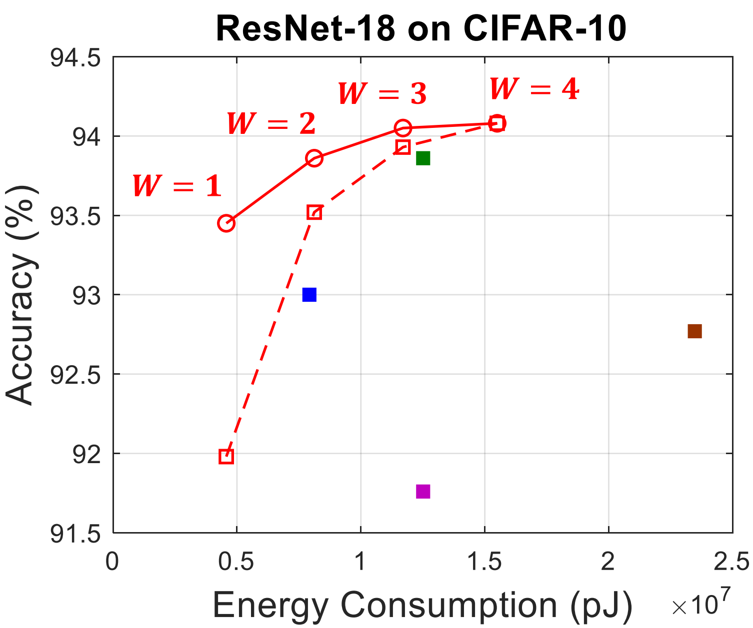
- BFP-CIM trims **<u>less contributing bit contents</u>** by RAR, leading to higher energy efficiency
  - Zero bits are also skipped

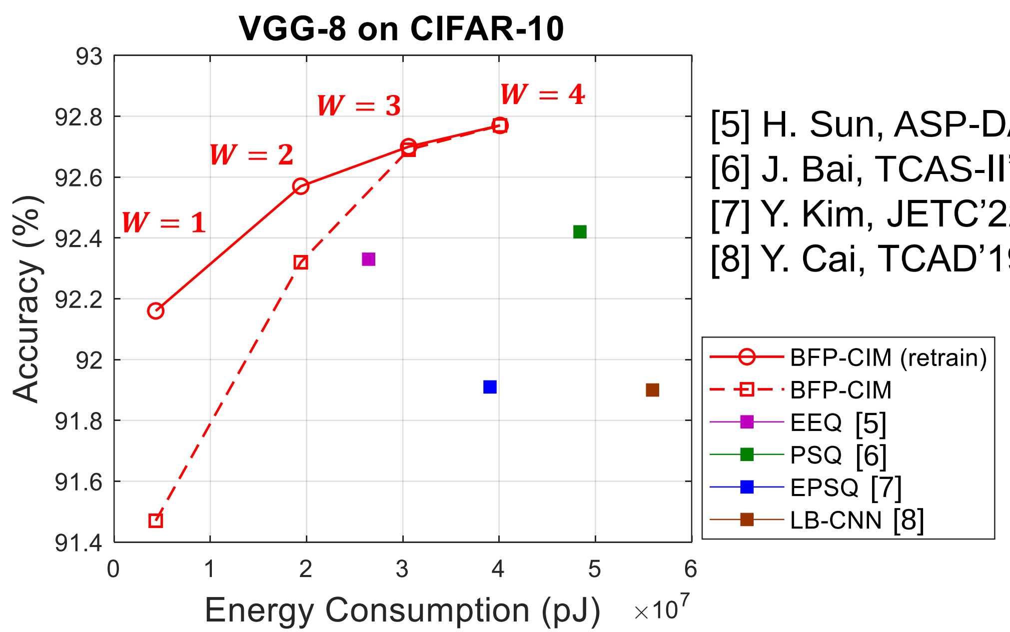


[1] A. Shafiee, ISCA'16
[2] F. Tu, ISSCC'22
[3] W. S. Khwa, ISSCC'22
[4] T. H. Yang, ISCA'19

Experiment – Energy-Accuracy Trade-off

- BFP-CIM **without retraining** still pushes the pareto front to a better trade-off than prior works
  - Window size (W) serves as a control knob





[5] H. Sun, ASP-DAC'20
[6] J. Bai, TCAS-II'23
[7] Y. Kim, JETC'22
[8] Y. Cai, TCAD'19

Summary

- We introduce Range-Aware Rounding (RAR) for dynamic quantization in CNNs
  - Eliminating data-dependent bit-width adjustments for CIM
- BFP-CIM adapts block-floating-point arithmetic to CIM, integrating RAR for processing FP-like values
  - Demonstrates up to 2.51× energy efficiency gain while maintaining accuracy levels
- More details and experiments are shown in paper