2nd Qualcomm Taiwan University Research Workshop 第二屆高通台灣研發合作成果發表會 Qualcomm

Energy-Efficient AI Accelerator for On-Device Training

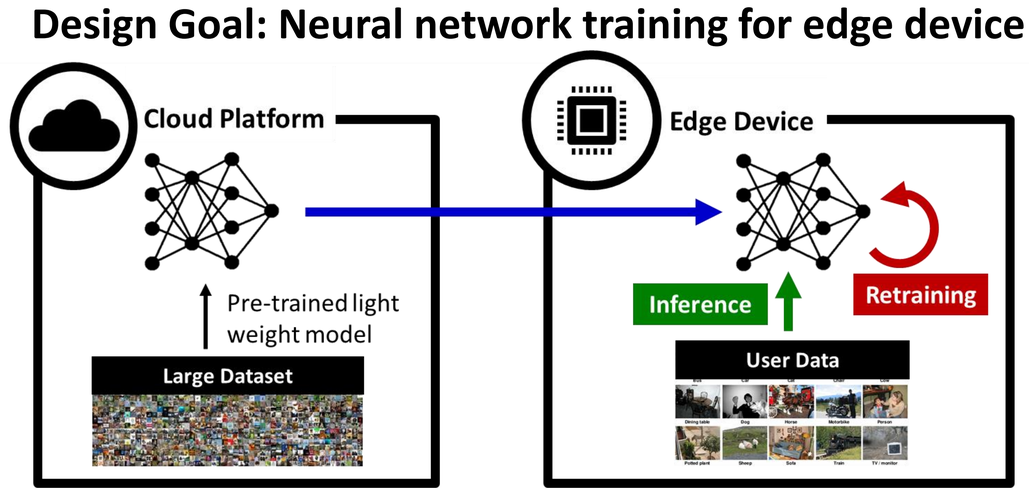
Zih-Sing Fu and Chia-Hsiang Yang
National Taiwan University

Research Overview

❖ **Motivation**
Energy-efficient AI accelerator for on-device training

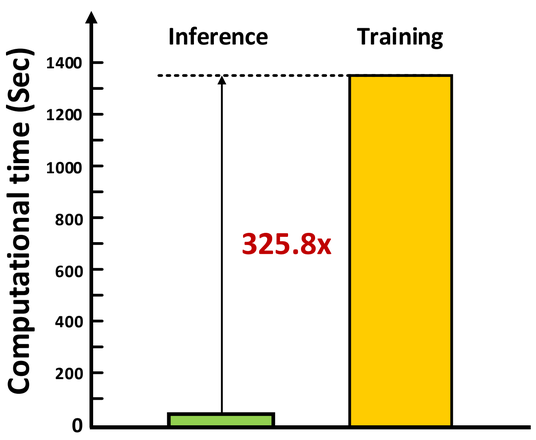
❖ **Challenges & Analyzed Techniques**
- ❑ Training requires large computation
  → Exploiting sparsity in data to skip zero MAC
- ❑ Training requires large data dynamic range
  → Selecting appropriate data type for training
- ❑ Training requires large external memory access (EMA) [1]
  → Reducing EMA with compression

Design Goal: Neural network training for edge device

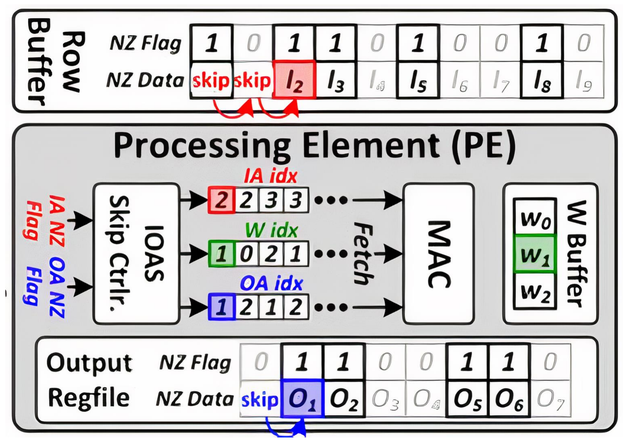


Sparsity Exploitation

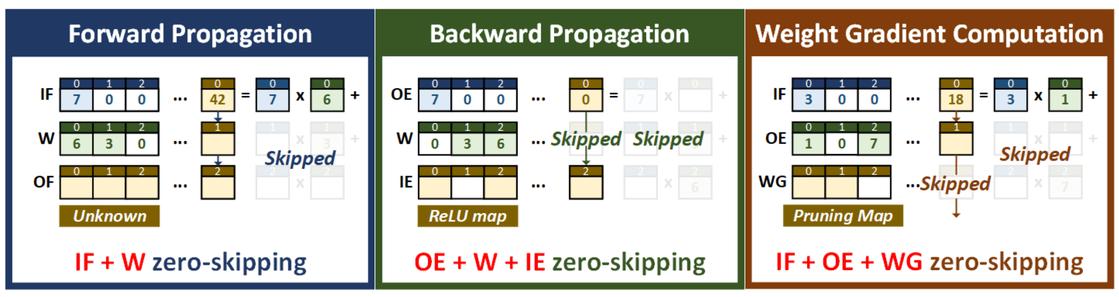
- ❖ Training complexity >> inference complexity
- ❖ Utilize sparsity-aware architecture to reduce computation
  - ❖ [2,3] exploits sparsity of activations and errors
  - ❖ [4] exploits sparsity of activations and weights
- ❖ Sparsity appears in all stages in neural network training
  - ❖ Activations and errors: ReLU
  - ❖ Weights and gradients: Pruning



Model: VGG16, Dataset: CIFAR10



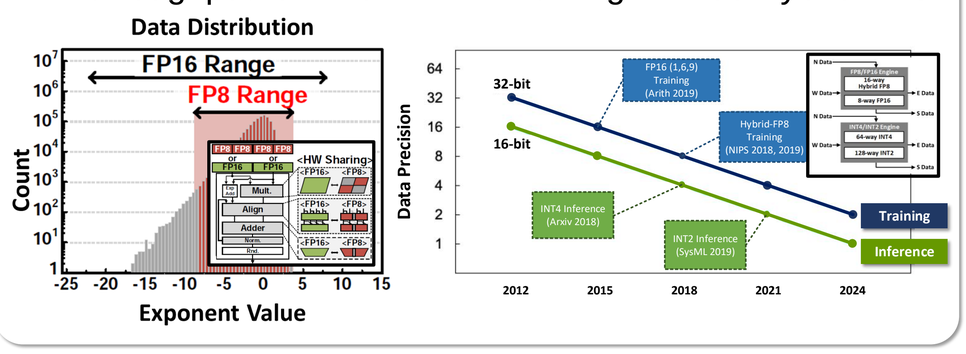
Sparsity-aware architecture example [3]



Sparsity that can be exploited in neural network training

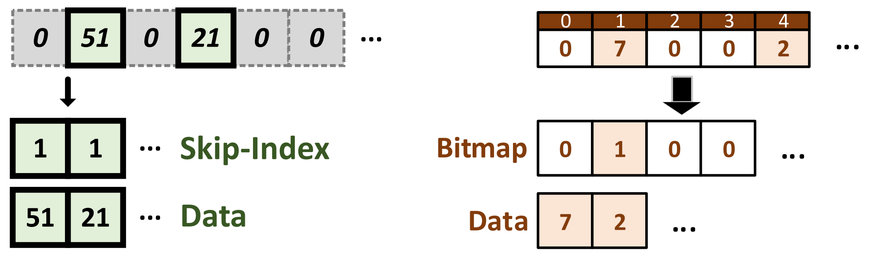
Data Type for Training

- ❖ Mixed-precision MAC [2]
- ❖ Data → FP8/FP16 tensors
- ❖ x2 throughput for PF8
- ❖ Low precision at inference [5]
- ❖ Inference: INT4 / INT2
- ❖ Training: FP16 / Hybrid-FP8

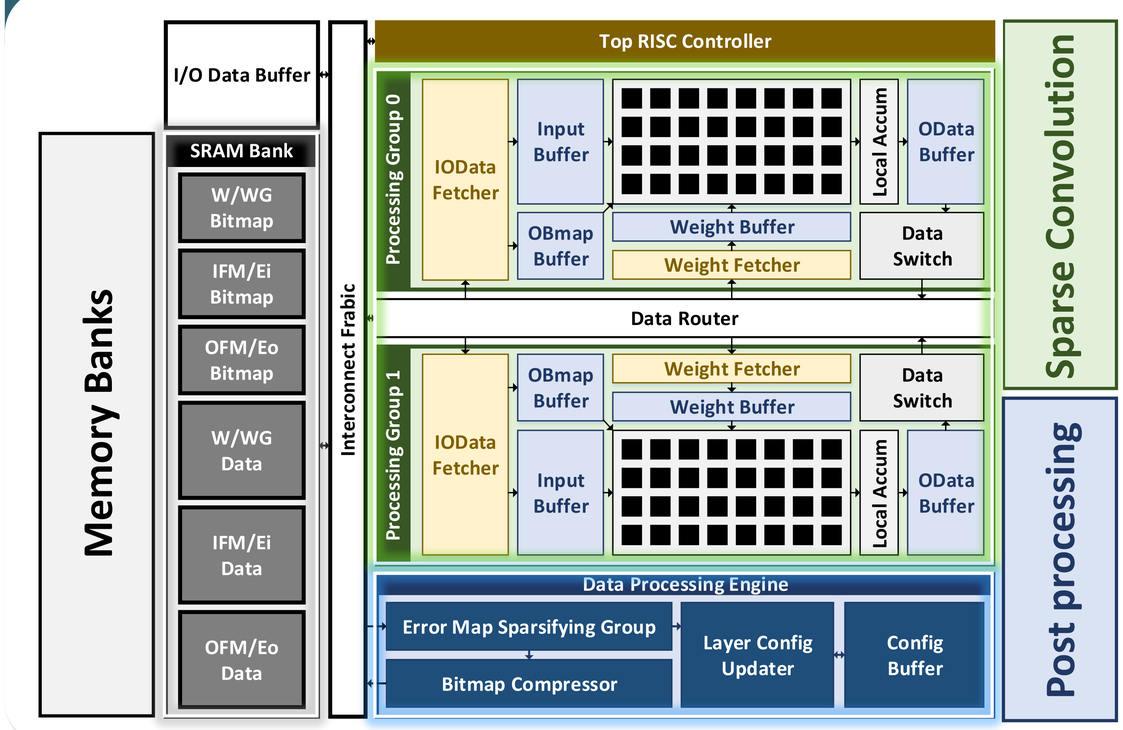


Data Compression

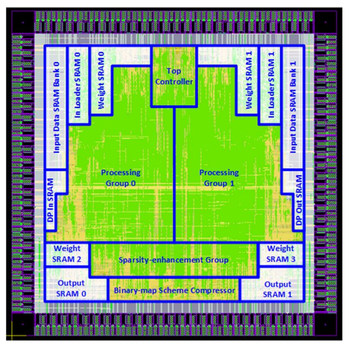
- ❖ Run-length encoding
- ❖ CR related to distribution
- ❖ Skip-idx width hard to decide
- ❖ Binary-mask scheme
- ❖ CR indep. of distribution
- ❖ 65% EMA↓ at training



System Architecture



Chip Layout & Performance



Technology	40nm CMOS
Chip Size	2.78 x 2.78 mm²
Core Size	2.20 x 2.20 mm²
Gate Count	7113K
On-chip SRAM (KB)	160
Max. Frequency	200MHz
Peak Performance (TOPS)	20.48 (8 bit)
Area Efficiency (GOPS/mm²)	4266.7 (8 bit)

		ISSCC'19 [2]	ISSCC'20 [3]	ISSCC'21 [5]	This Work
Supported Operations	Sparsity Exploitation	Yes	Yes	No	**Yes**
	Sparsity Inducing	No	No	No	**Yes**
	Training Precision*	16 bit	16 bit	16 bit	**8 bit**
	Compression	No	No	No	**Yes**
Technology		65nm	65nm	7nm	**40nm**
Supply Voltage (V)		0.78 - 1.1	0.7 - 1.1	0.55 - 0.75	**0.9**
Area (mm²)		16.0	32.4	19.6	**4.8**
On-chip SRAM (KB)		372	676	8000	**160**
Max Frequency (MHz)		200	200	1600	**200**
Peak Performance* (TOPS)		> 0.3	14.03	25.6	**20.48**
Area Efficiency (GOPS/mm²)		24.0	433.0	1306.1	**4266.7**

* Max precision required for training

[1] T.-J. Yang *et al.*, "Designing energy-efficient convolutional neural networks using energy-aware pruning," *CVPR*, June 2017.
[2] J. Lee *et al.*, "LNPU: A 25.3 TFLOPS/W sparse deep-neural-network learning processor with fine-grained mixed precision of FP8-FP16," *ISSCC*, Feb. 2019.
[3] S. Kang *et al.*, "7.4 GANPU: A 135TFLOPS/W multi-DNN training processor for GANs with speculative dual-sparsity exploitation," *ISSCC*, Feb. 2020.
[4] Y. Yu *et al.*, "SPRING: A Sparsity-Aware Reduced-Precision Monolithic 3D CNN Accelerator Architecture for Training and Inference," *IEEE Transactions on Emerging Topics in Computing*, 2020.
[5] A. Agrawal *et al.*, "A 7nm 4-Core AI Chip with 25.6TFLOPS Hybrid FP8 Training, 102.4TOPS INT4 Inference and Workload-Aware Throttling," *ISSCC*, Feb. 2021.