59 DESIGN AUTOMATION CONFERENCE

JULY 10 - 14, 2022

MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA

59 DESIGN AUTOMATION CONFERENCE

EBSP: Evolving Bit Sparsity Patterns for Hardware-Friendly Inference of Quantized Deep Neural Networks

Fangxin Liu (Speaker)

Wenbo Zhao, Zongwu Wang, Yongbiao Chen, Zhezhi He, Naifeng Jing, Xiaoyao Liang, and **Li Jiang***

Shanghai Jiao Tong University

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

先进计算机体系结构实验室 Advanced Computer Architecture Laboratory

Outline

- Background and motivation
- Proposal: Evolving Bit Sparsity Patterns for Hardware-Friendly Inference of Quantized Deep Neural Networks
- Design and implementation details
- Experiment results
- Conclusion

Pervasive DNN applications

DNNs are widely used:

Translation

Recommendation systems

Face recognition

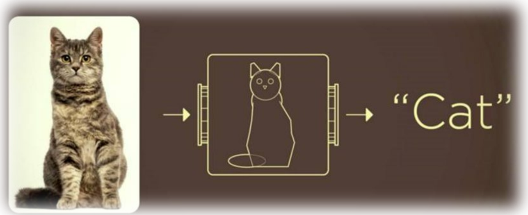
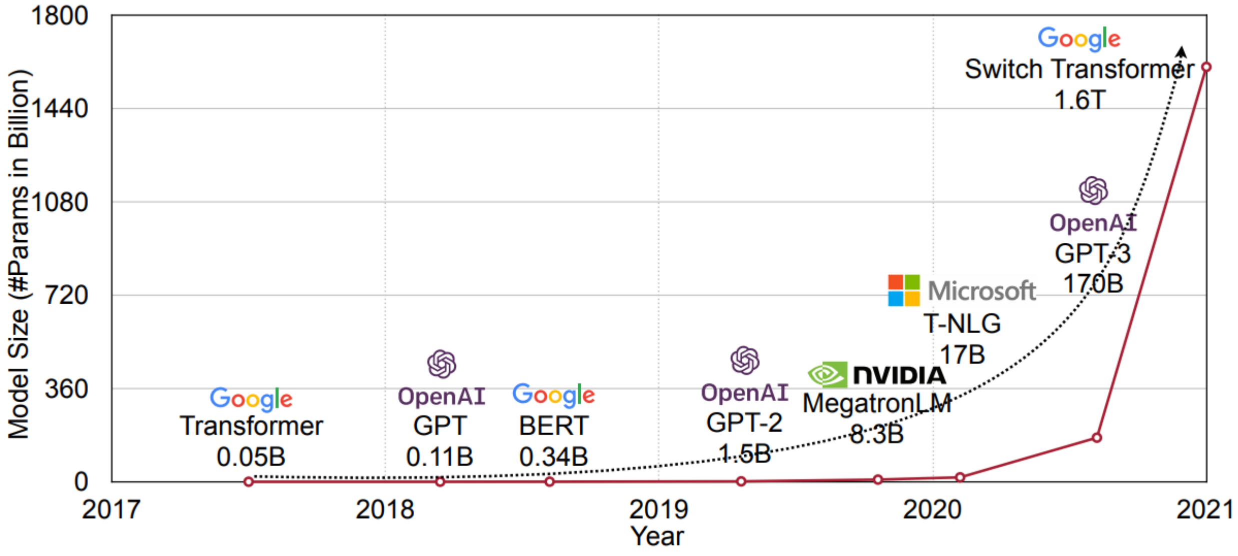


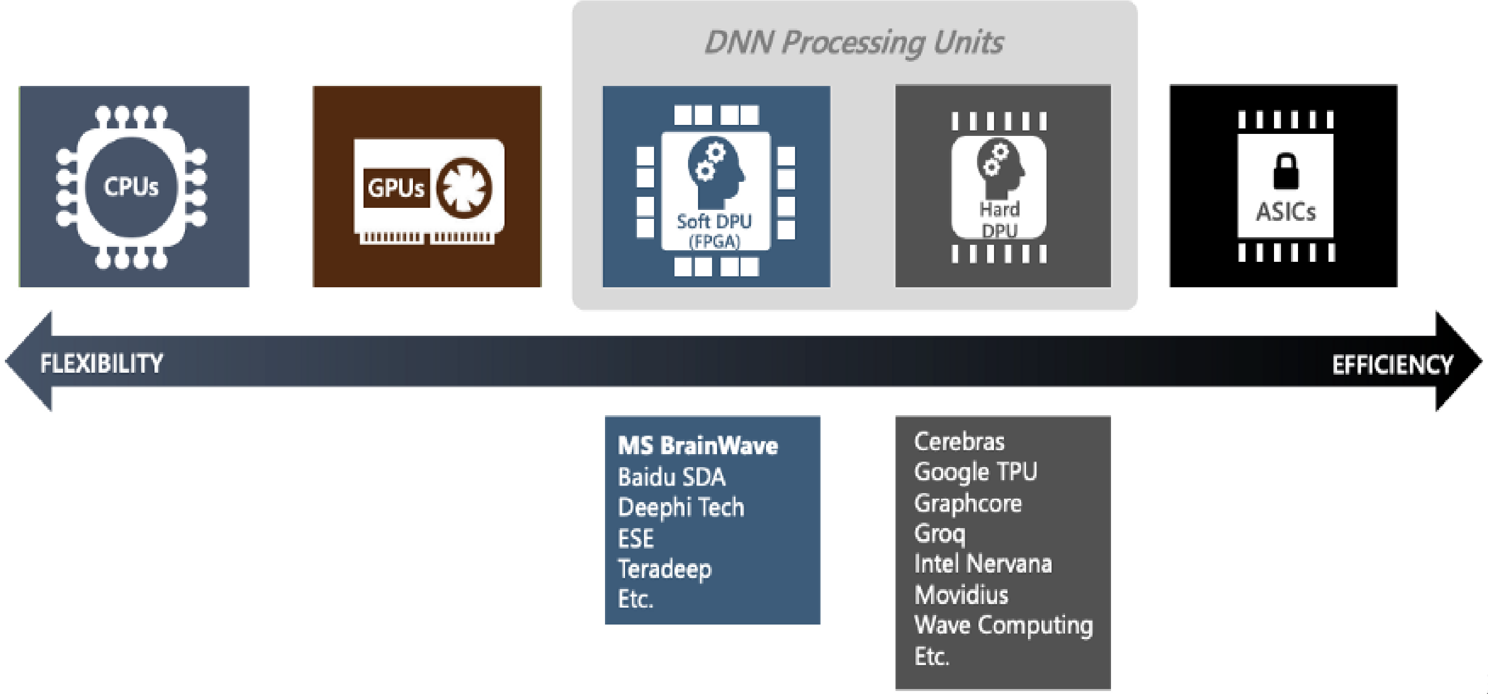
Image classification



DNN model size and computation are increasing exponentially

DNN Acceleration

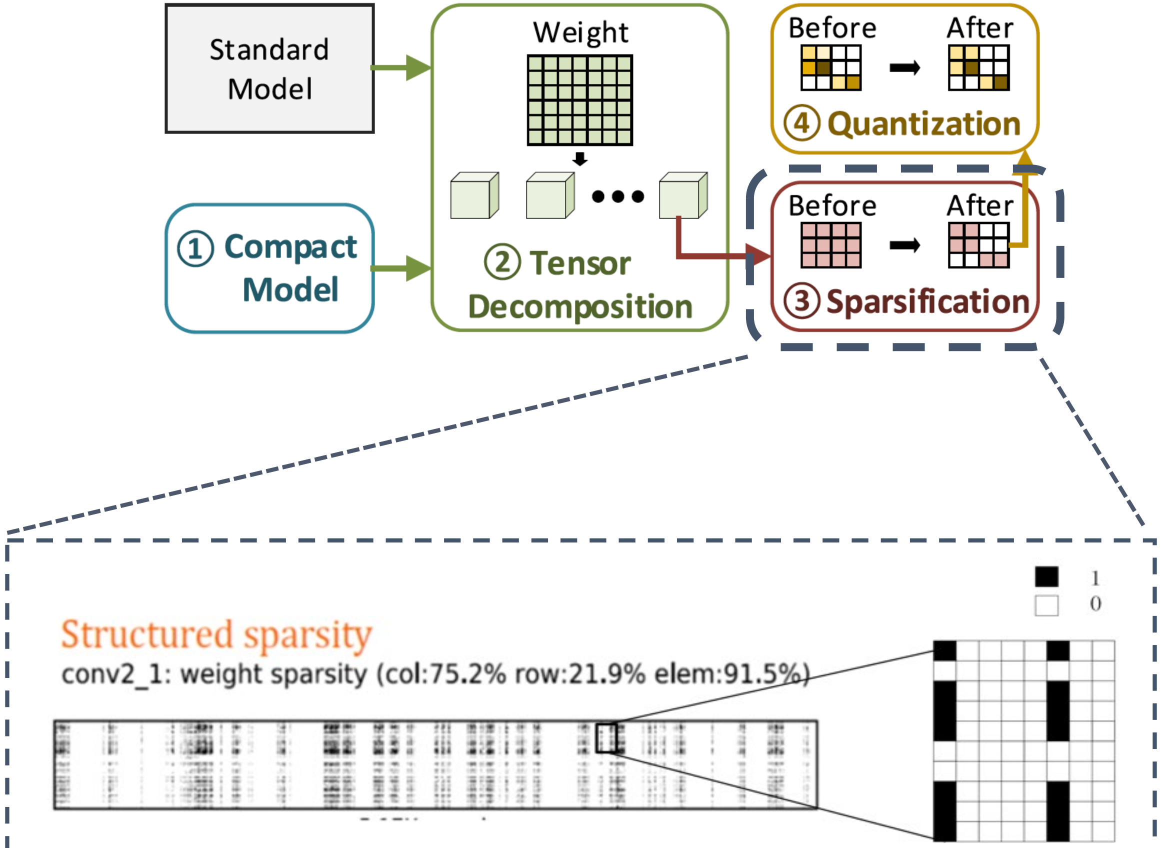
Nowadays, accelerators are gaining a lot of traction, as more and more DNNs become targets for accelerations.



Processing Alternatives for DNNs [Source: Microsoft]

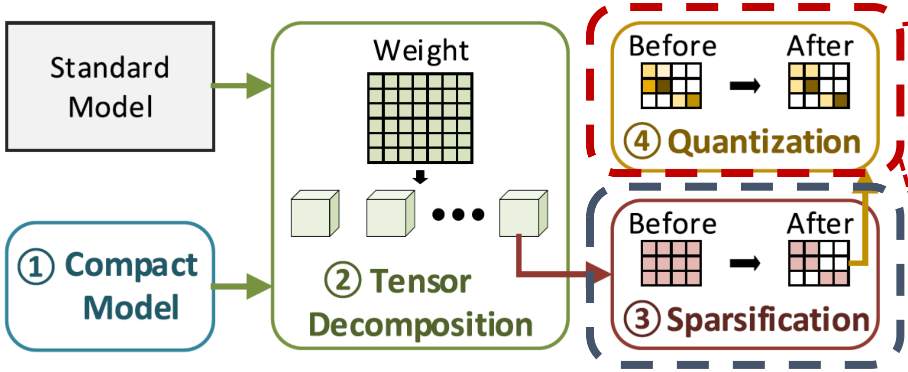
59 DESIGN AUTOMATION CONFERENCE

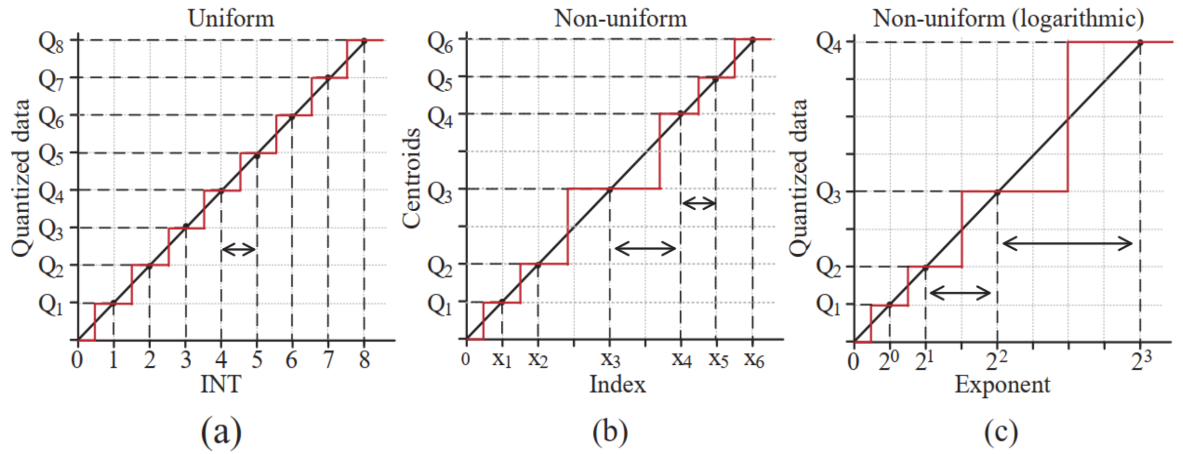
DNN Acceleration



DNN model contains high redundancy, remove unimportant weights

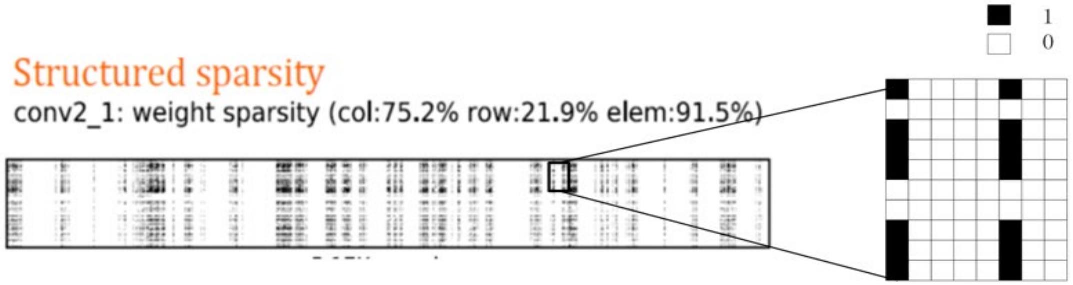
DNN Acceleration





Quantization:

- Using lower precision to represent operands.
- Using lower precision math.



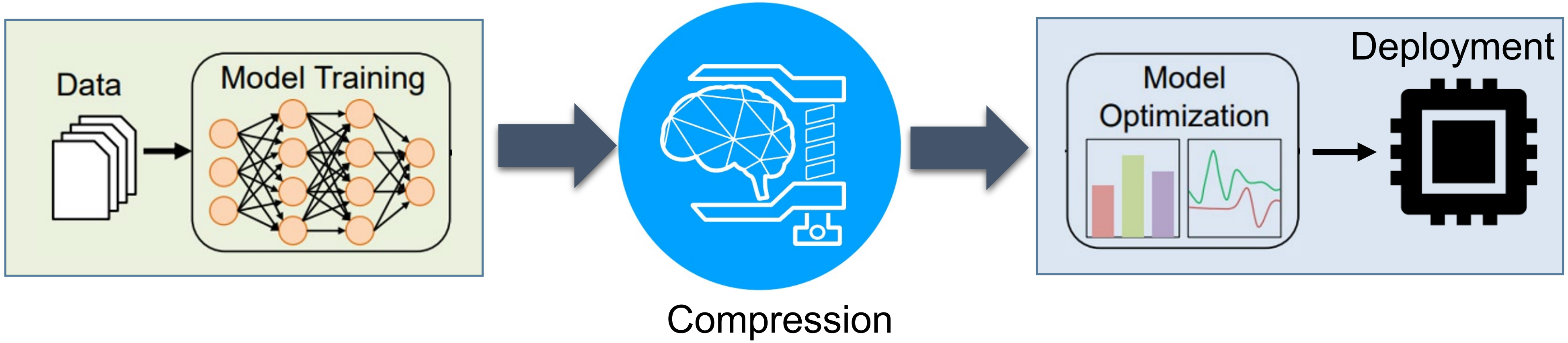
DNN model contains high redundancy, remove unimportant weights

	FP32	FP16	Int8
MobileNet v1	1	1.91	2.49
MobileNet v2	1	1.50	1.90
ResNet50 (v1.5)	1	2.07	3.52
VGG-16	1	2.63	2.71
VGG-19	1	2.88	3.09
Inception v3	1	2.38	3.95
Inception v4	1	2.99	4.42
ResNext101	1	2.49	3.55

Inference with low precision on GPU [Source: NVIDIA]

DESIGN AUTOMATION CONFERENCE 59

Challenges in Compression Techniques



- 1) Quantization methods focus on improving the compression rate of ultra low-precision DNN models, resulting in significant accuracy losses.
- 2) Sparsification methods need additional indexing overhead for addressing non-zero elements and irregular access/execution patterns.
- 3) Sparsification or ultra low-precision quantization methods always introduce ancillary overheads, which is implementation-unfriendly.

59 DESIGN AUTOMATION CONFERENCE

Overview of Our EBSP Algorithm

We revisit the quantization process from a new angle of bit-level sparsity

Quantization

The reduction of the precision of an operand can be taken as forcing one or more bits among the operand to be zero (lower significant bit is more likely to be zero)

Pruning bits

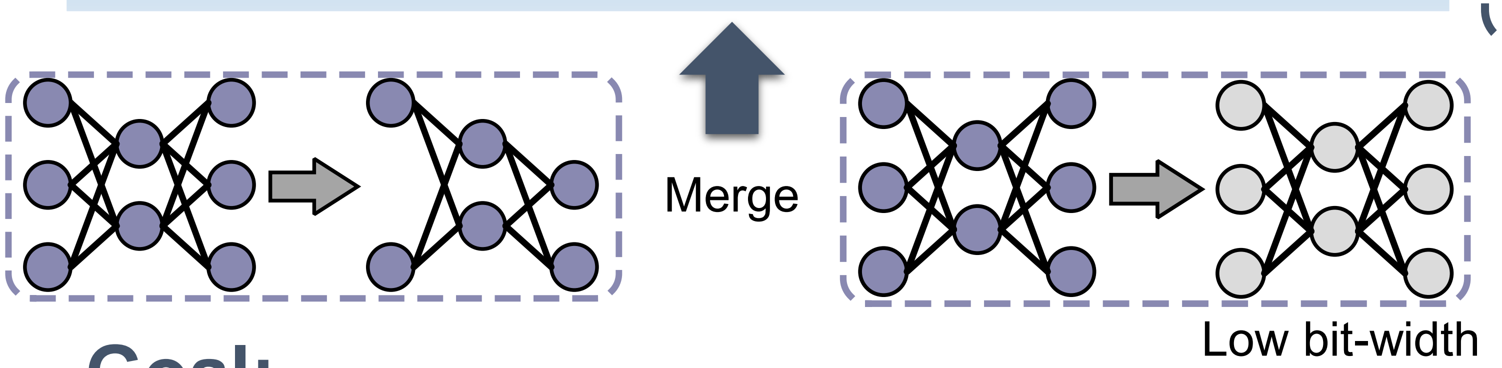
Quantization can be viewed as increasing bit-level sparsity among the operand

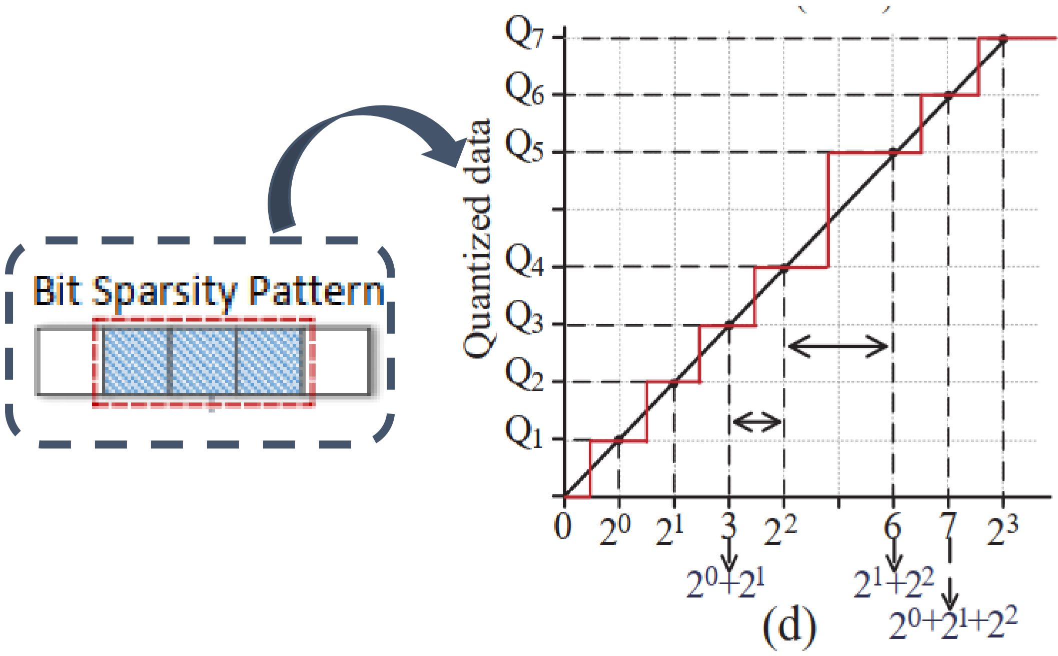
59 DESIGN AUTOMATION CONFERENCE

Overview of Our EBSP Algorithm

Coupling Quantization with Hardware

- The proposed quantization scheme incorporating the **bit sparsity pattern** can be considered as a **variant** of the non-uniform quantization.

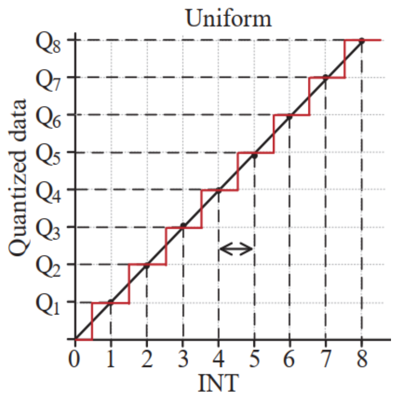




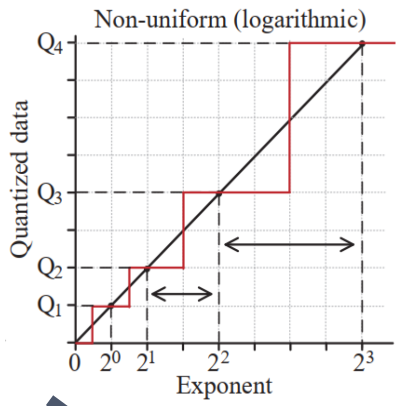
Goal:

- Eliminate multiplication operations in the (quantized) DNN.
- Address the non-negligible accuracy loss of quantization with low bit-width.

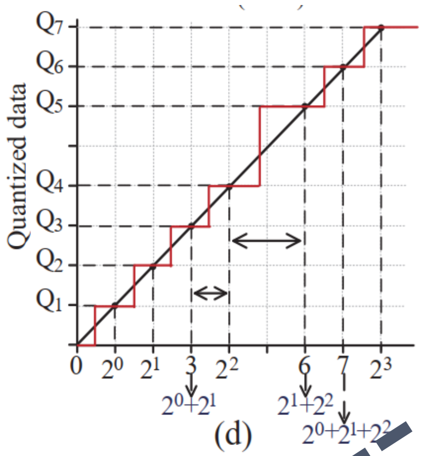
Overview of Our EBSP Algorithm



(a)



(c)



(d)

Complex Multiplication

High Accuracy

Simple Multiplication (Shift)

Low Accuracy

High Accuracy

Introduce only hardware that assists in combining LUT entries to realize multiplications.

- LUTs can be reconfigured to support different bit-width
- the reduced operand precision enables fewer LUT entries

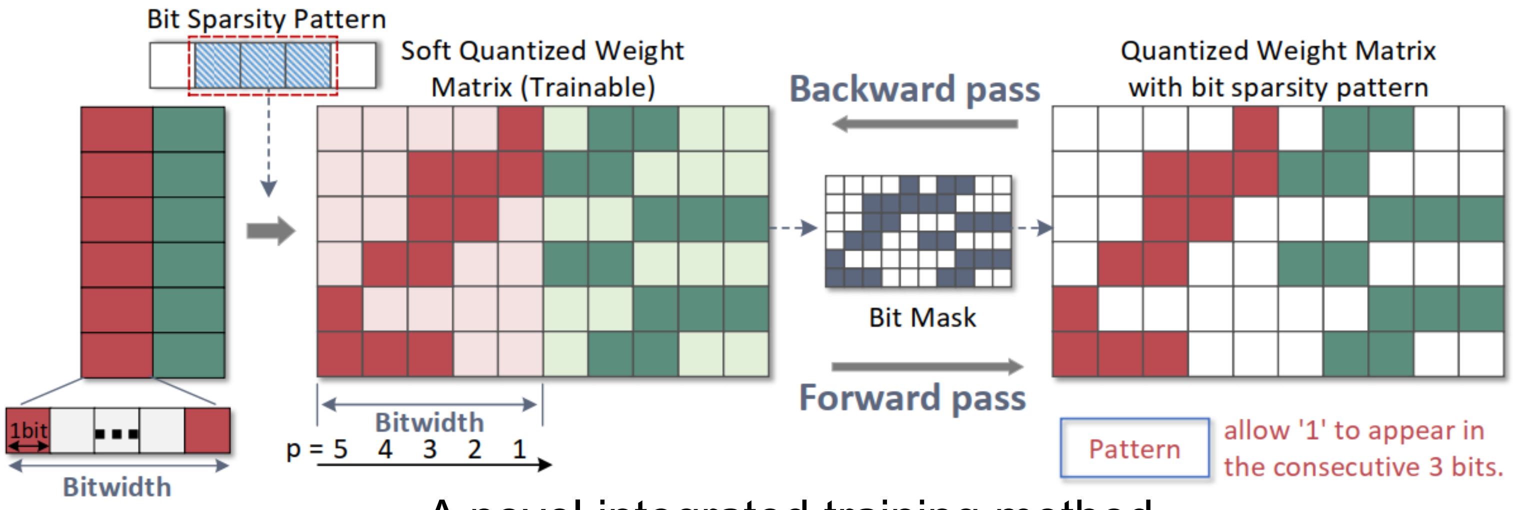
Overview of Our EBSP Algorithm

Problem in the LUT-based Scheme:

- An excessive number of entries to cover all the possible combinations of weights and activations.
  - To compute a multiplication with INT8 quantization in one cycle, 65,536 ($2^8 \times 2^8$ combinations) entries are needed in the LUT.

We divide the training process into three phases sequentially:

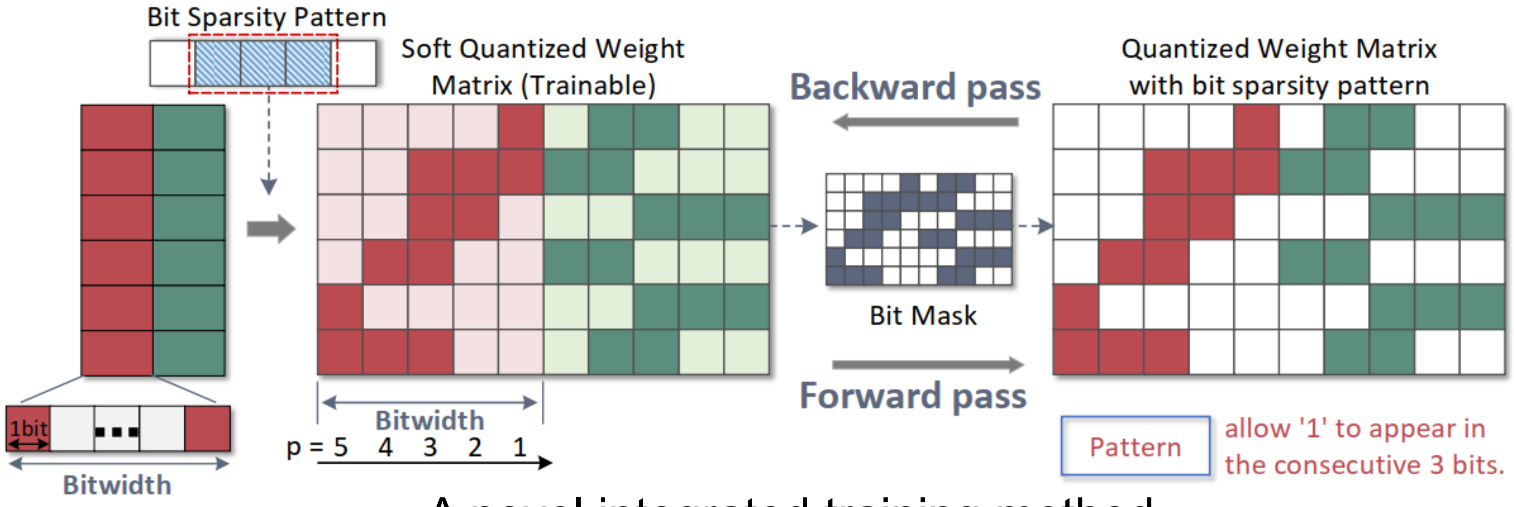
- **Masking**
- **Forward passing**
- **Backward passing**



A novel integrated training method

DESIGN AUTOMATION CONFERENCE 59

Overview of Our EBSP Algorithm



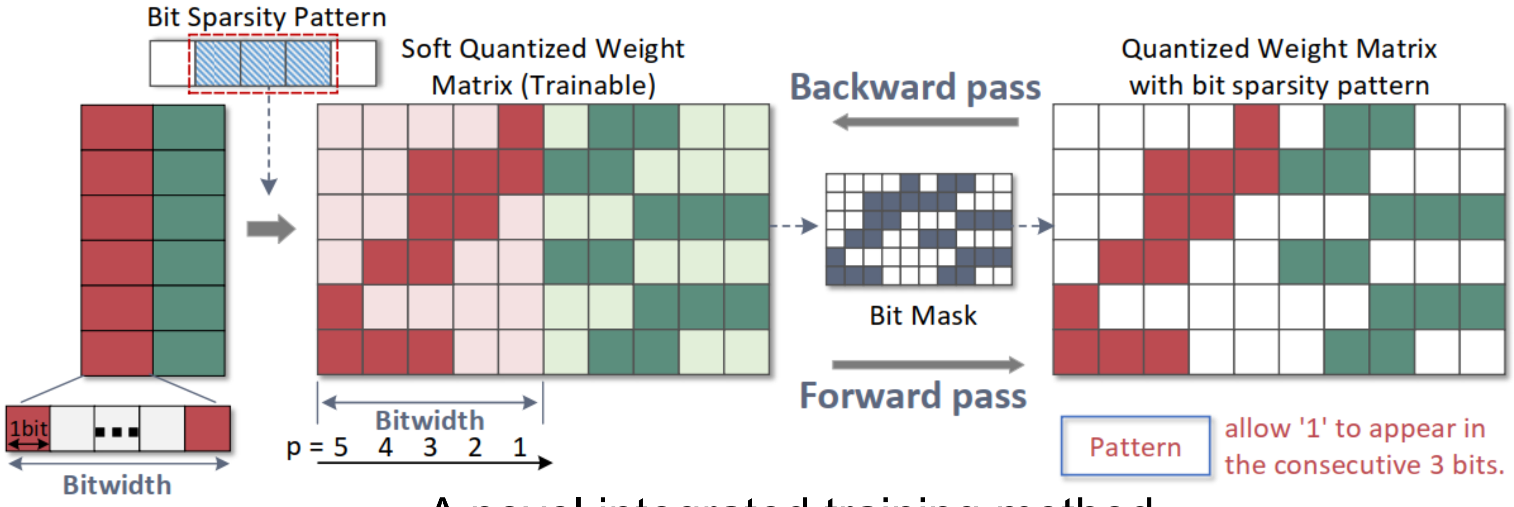
A novel integrated training method

bit sparsity constraint is that a maximum of 3 consecutive '1's exists in the bits within the weights at a given bit-width.

Masking

- the weight matrix is quantized to the target bit-width and imposes bit sparsity constrain in the bits within the target bit-width.

Overview of Our EBSP Algorithm

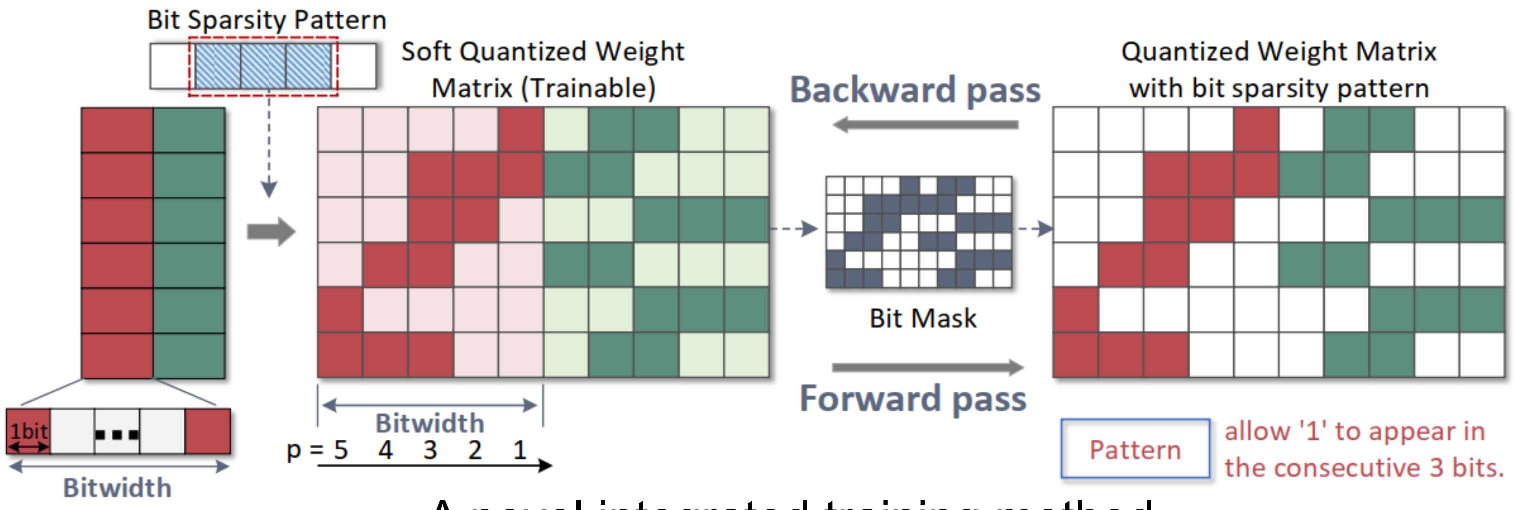


A novel integrated training method

- Masking
- **Forward passing**
  - the original weight matrices are quantized prior to passing through masking

The layer computations are carried out with the bit sparsity pattern of the quantized weight matrices.

Overview of Our EBSP Algorithm



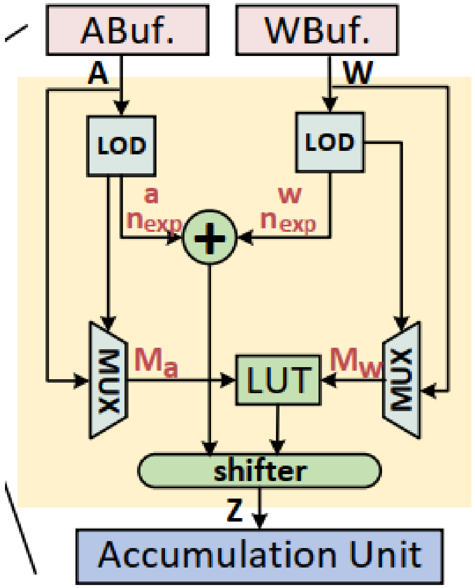
A novel integrated training method

ADMM for Weight Quantization:
higher compression ratio and lower accuracy degradation

- Masking
- Forward passing
- **Backward passing**
  - add a normalization term to the loss to decay the weights toward the quantized one.

Overview of Our EBSP Architecture

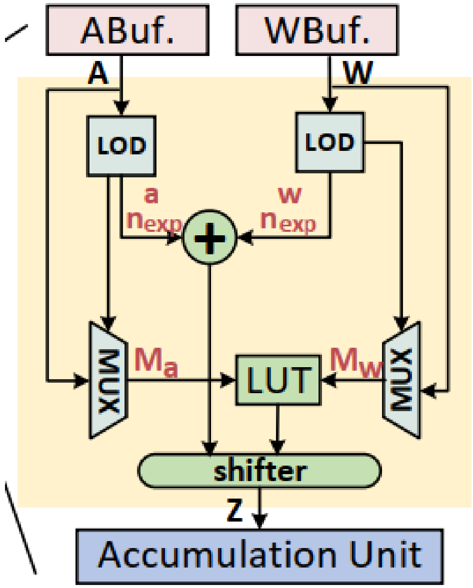
- PE is designed with two components:
  - steering logic ✔
  - arithmetic logic



The steering logic is composed of leading one detector (LOD) that dynamically locate the most significant `1' bit and a multiplexer that extracts significant digits to send to the LUT.

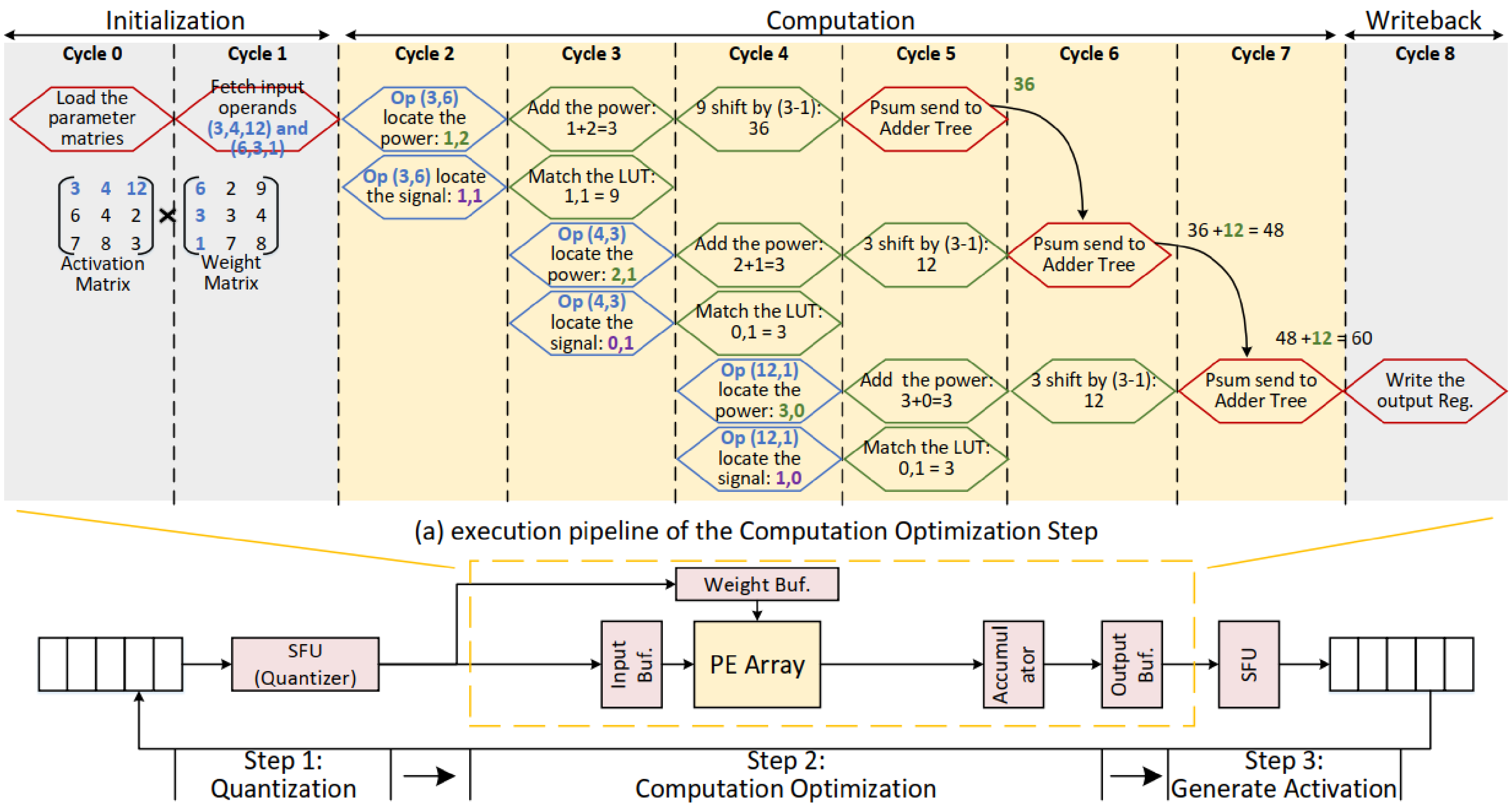
Overview of Our EBSP Architecture

- PE is designed with two components:
  - steering logic
  - arithmetic logic ✔



The arithmetic logic is composed of a LUT with few entries, adder, and shifter (e.g., barrel shifter) that implements the multiplication.

Overview of Our EBSP Architecture



(a) execution pipeline of the Computation Optimization Step

(b) Illustration of the EBSP that runs each modules.

Experiment Settings

Dataset

- CIFAR-10
- ImageNet

Network

- AlexNet
- VGGNet
- ResNet
- MobileNet

Modeling architecture

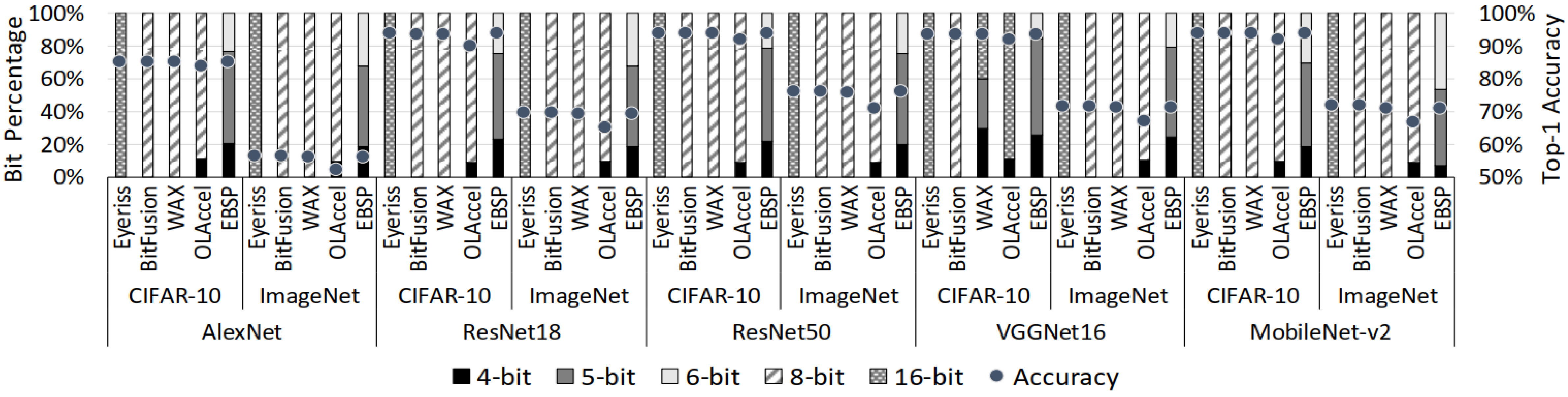
- Simulation
- CACTI

	Eyeriss [4]	BitFusion [24]	WAX [10]	OLAccel [20]	EBSP
Bit-width	16-bit	4-bit	8-bit	4&16-bit	6-bit (3)†
Data Format	Integer	Integer	Fixed-point	Integer	Integer
# PEs	224	3168	102	2499	4818
Area (mm2)	0.32	0.32	0.32	0.32	0.32

† this denotes the length of bit sparsity pattern, which determines LUT entries.

DESIGN AUTOMATION CONFERENCE 59

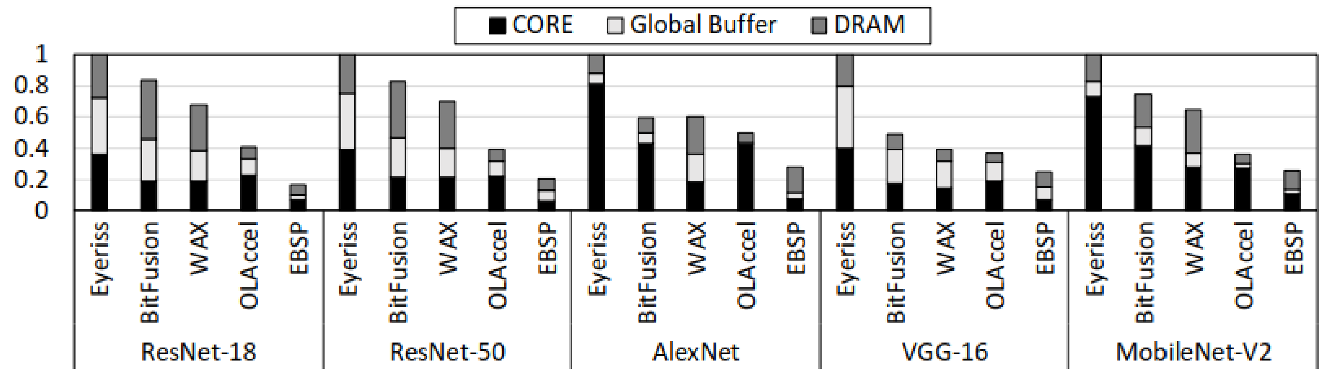
Experiment Results — Accuracy



- CIFAR-10: EBSP shows nearly no accuracy loss compared to Eyesiss (full INT16), BitFusion (full INT8) and WAX (full fixed-point 8-bit), and a 2.2% accuracy improvement over the OLAccel.
- ImageNet: EBSP shows a 0.31% accuracy loss compared to Eyesiss, WAX and BitFusion and a significant 4.32% accuracy improvement over OLAccel.

59 DESIGN AUTOMATION CONFERENCE

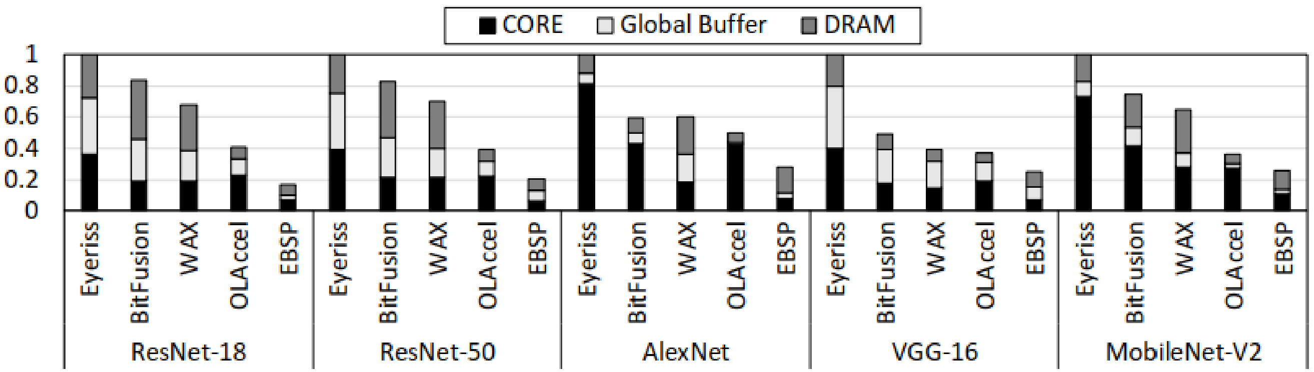
Experiment Results — Energy & Performance



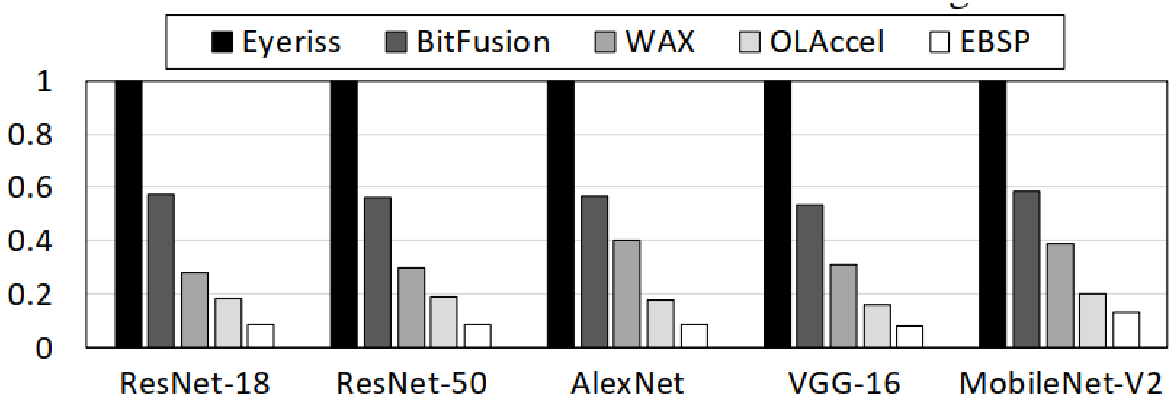
- Taking ResNet-50 as an example, compared to Eyeriss, BitFusion, WAX and OLAccel, EBSP consumes 87.3%, 79.7%, 75.2% and 58.9% less energy, respectively.

59 DESIGN AUTOMATION CONFERENCE

Experiment Results — Energy & Performance



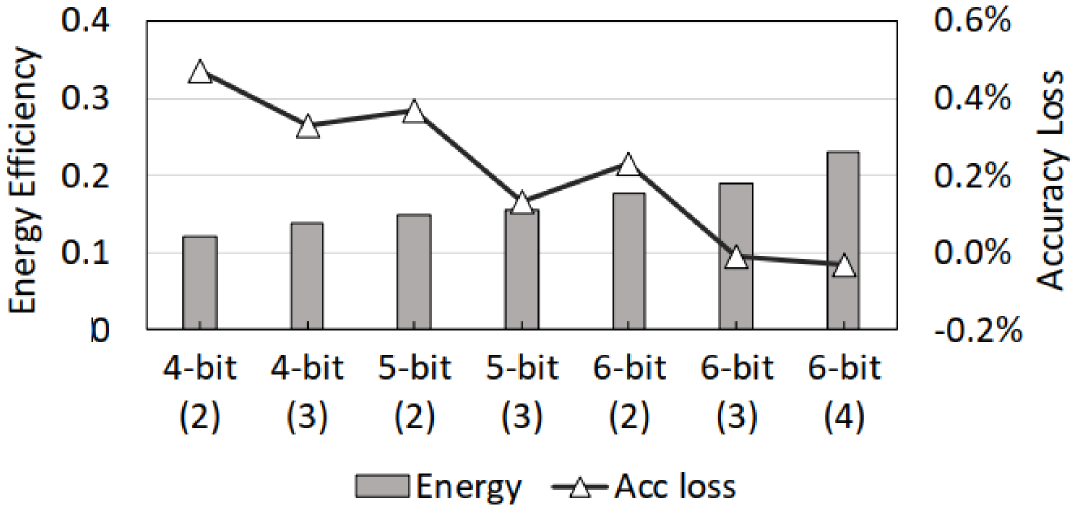
- Compared to Eyeriss, EBSP achieves nearly 93% acceleration improvement.



- Taking ResNet-50 as an example, compared to Eyeriss, BitFusion, WAX and OLAccel, EBSP consumes 87.3%, 79.7%, 75.2% and 58.9% less energy, respectively.

59 DESIGN AUTOMATION CONFERENCE

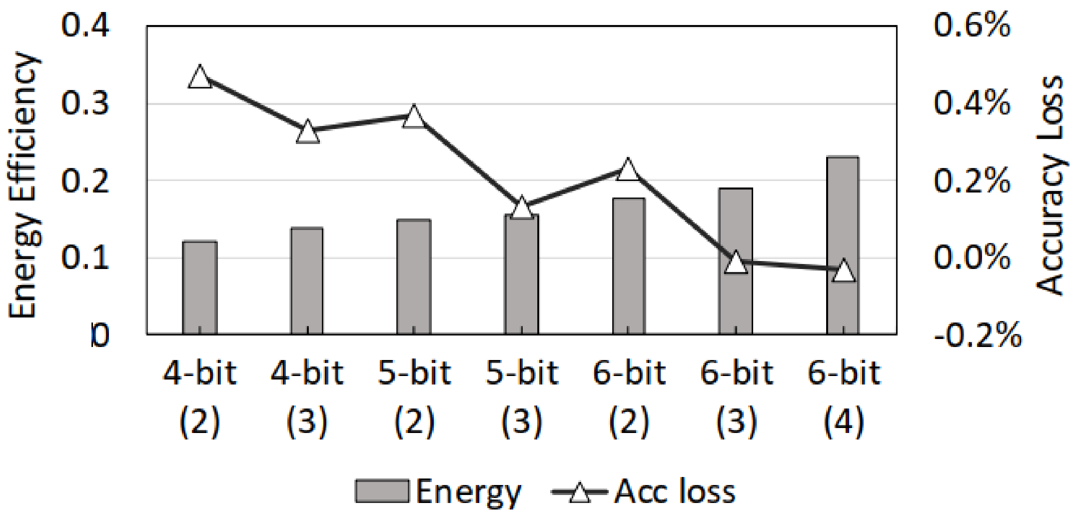
Experiment Results — Pattern Length



- EBSP has the capability of tuning the length of bit sparsity pattern to sustain the same accuracy levels as Eyesiss, while gaining notable energy efficiency.

DESIGN AUTOMATION CONFERENCE 59

Experiment Results — Pattern Length



- EBSP has the capability of tuning the length of bit sparsity pattern to sustain the same accuracy levels as Eyesiss, while gaining notable energy efficiency.
- Quantized DNNs with bitwidth of 5-bit and pattern length of 3, EBSP achieves an optimal point (0.13% accuracy loss with 97.3% energy reduction over Eyeriss) on ImageNet

59 DESIGN AUTOMATION CONFERENCE

Conclusion

- novel hardware-friendly quantization algorithm
  - form bit sparsity patterns in quantization-aware training
  - reap the full advantages of sparsity and quantization
- An efficient execute-search dual-engine PIM-based architecture
  - Non-Multiplication Engine
  - Execution Flow
  - Minimum Required Modifications
- Keep high accuracy while gaining large performance improvement

DESIGN AUTOMATION CONFERENCE 59

Thank you !

EBSP: Evolving Bit Sparsity Patterns for Hardware-Friendly Inference of Quantized Deep Neural Networks

Fangxin Liu (Speaker)
Wenbo Zhao, Zongwu Wang, Yongbiao Chen, Zhezhi He, Naifeng Jing, Xiaoyao Liang, and Li Jiang*
Shanghai Jiao Tong University

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

59 DESIGN AUTOMATION CONFERENCE