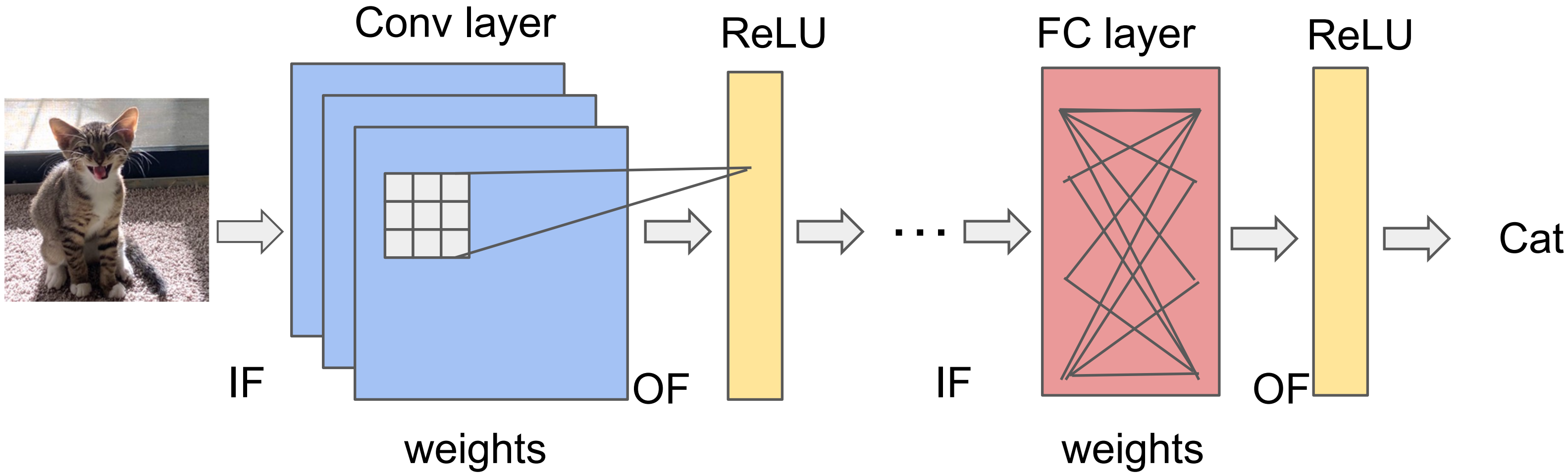
BitSET: Bit-Serial Early Termination for Computation Reduction in Convolutional Neural Networks

Yunjie Pan, Jiecao Yu, Andrew Lukefahr, Reetuparna Das, Scott Mahlke

University of Michigan
panyj@umich.edu

CASES’23 Sept 20, 2023

Conv and FC Dominates the CNN Workloads



>80% of the runtime is Convolution (Conv) and Fully-connected (FC) layers

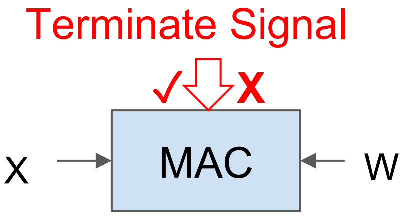
The main operation is **MAC (Multiply-Accumulate)**

Research Challenge

How to reduce **the number of MAC operations** in CNN inference?

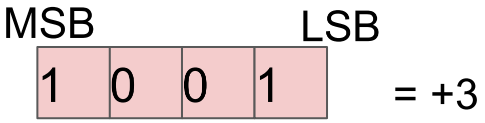
Solution: BitSET, Software-Hardware Co-design

Algorithm



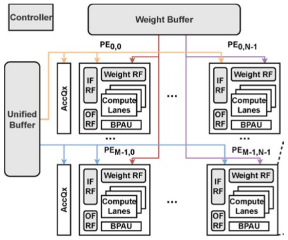
Early termination

Encoding



Terminate even earlier

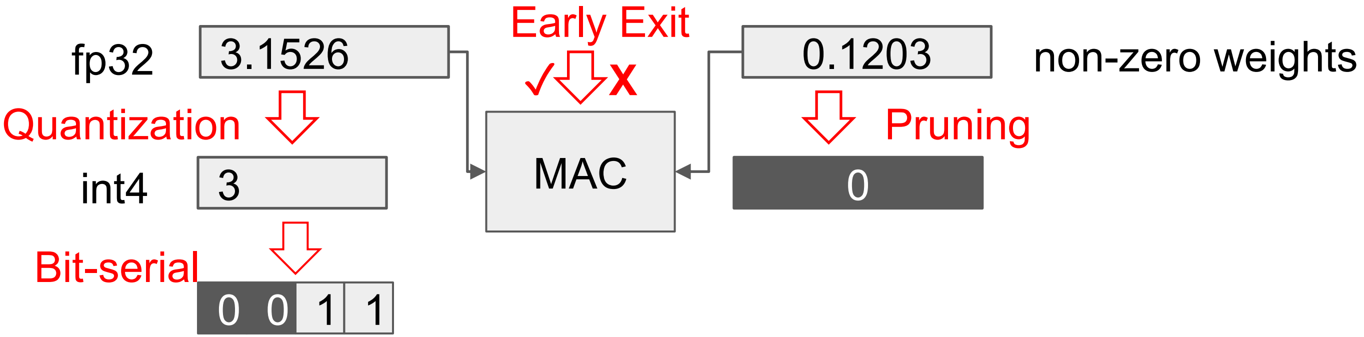
Accelerator



Customized hardware

Conventional Methods to Reduce # MAC Operations

- **Quantization**: reduce precision
- **Pruning**: Make redundant weight values to be zero
- **Bit-serial computation:** bit-level sparsity
- **Early Exit**: Trade off accuracy with efficiency



Can We **aggressively** Skip **Bit-level** Computation?

- **Quantization**: reduce precision
- **Pruning**: Make redundant weight values to be zero
- **Bit-serial computation:** bit-level sparsity
- **Early Exit**: Trade off accuracy with efficiency

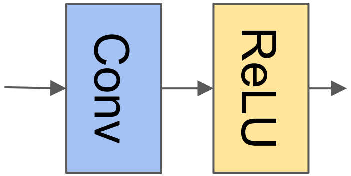
How?

Runtime information using characteristic of CNN model structure

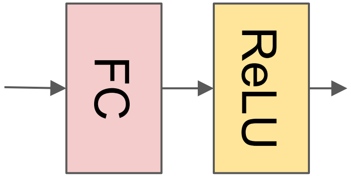
Why bit-level?

Finer-granularity compared to value-level (stop at any bit)

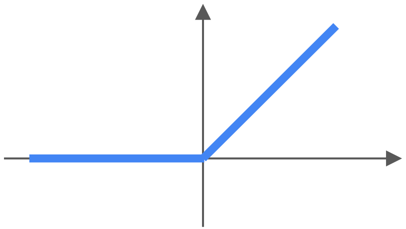
Characteristic of the CNN Model Structure



Conv + ReLU



FC + ReLU

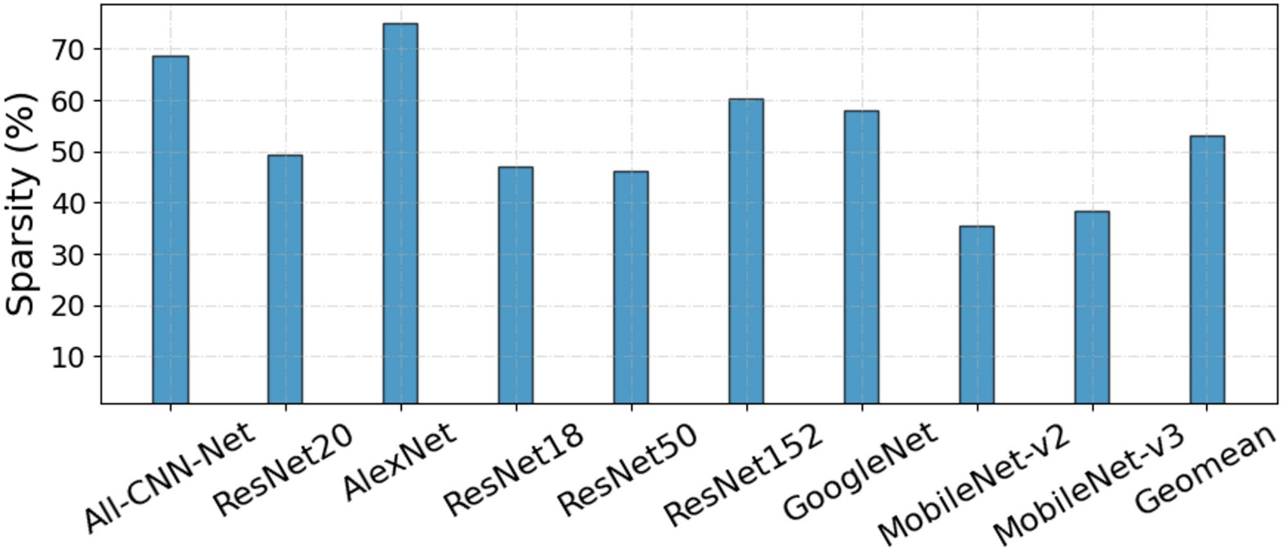


ReLU $y = \min(x, 0)$

ReLU clamps negative values to zero

Don’t care how “negative” the value is

Opportunities to Skip Computation

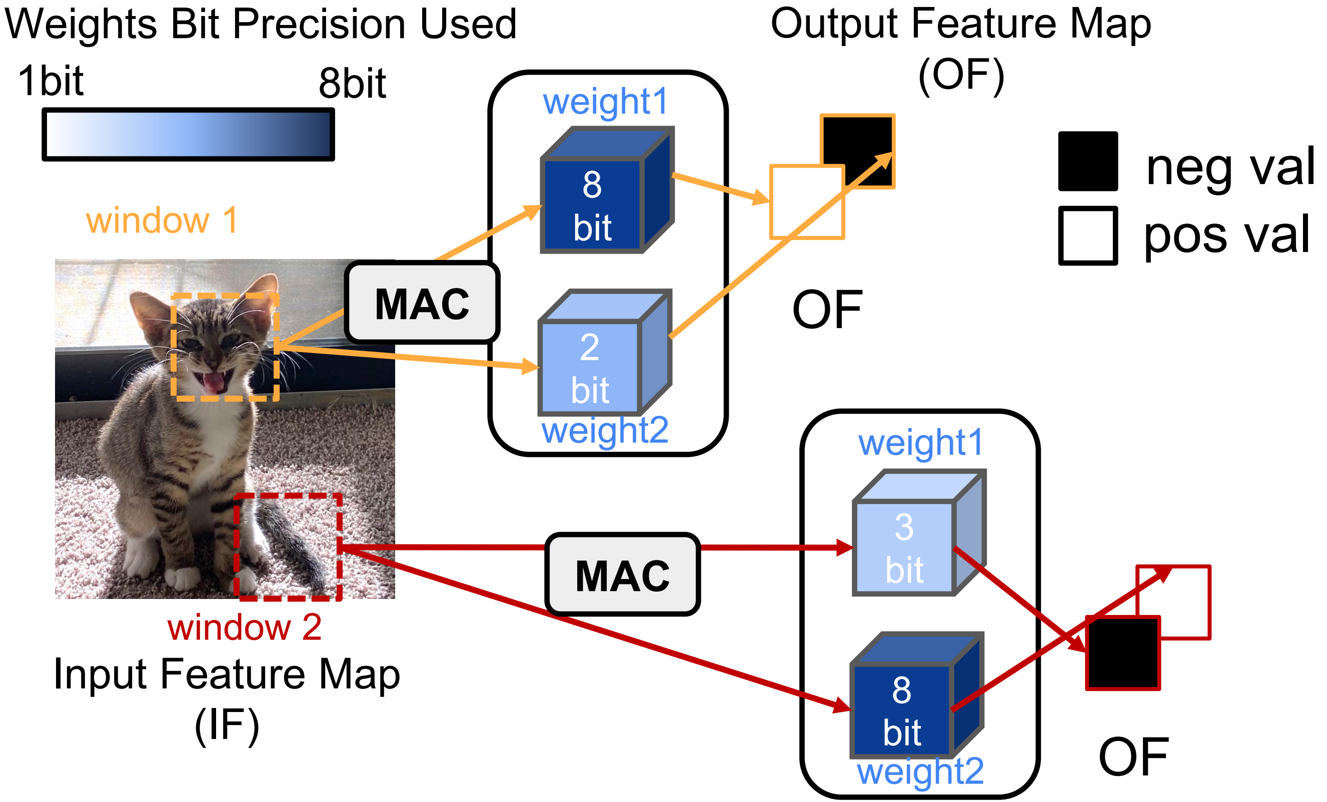


>50% of **Conv/Fc** outputs are negative, resulting in great **sparsity** after ReLU

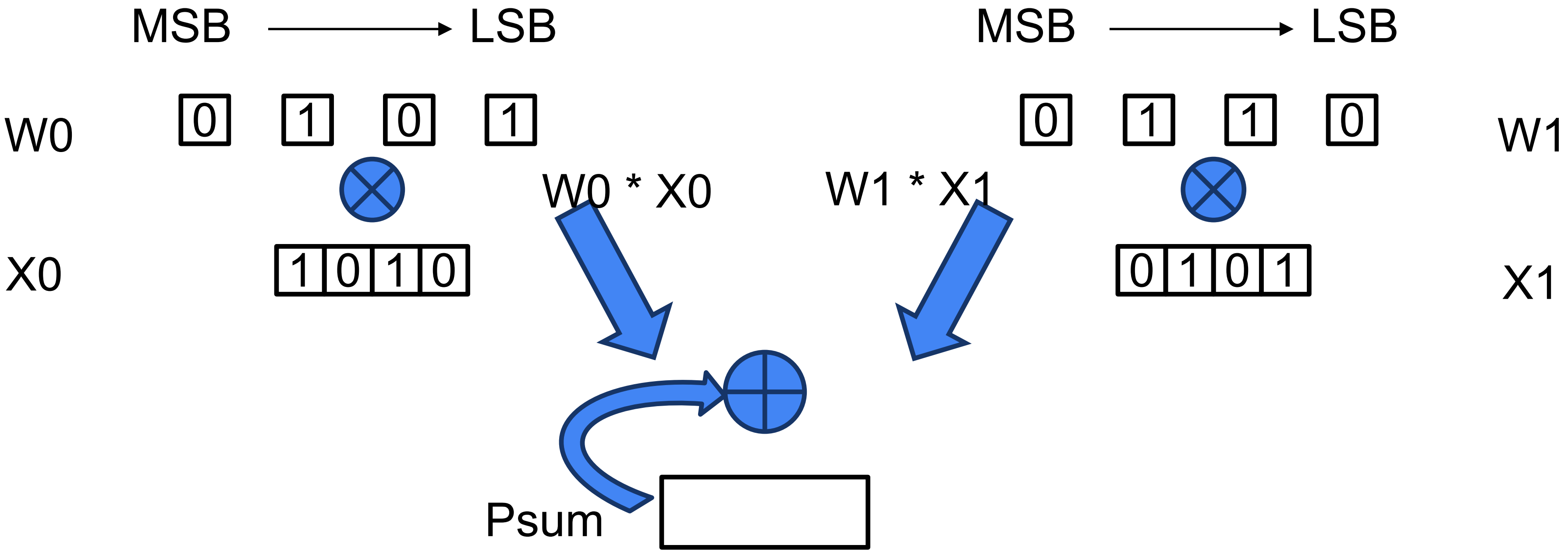
Skip “Negative” Computation at Bit-level

- Early Exit with **high-order bits of weights** for **negative** outputs
  - Predict and skip
  - Use as few bits of weights as possible, do as little computation as possible
- Use **all bits of weights** for **positive** outputs
  - Exact values
  - Act as normal MAC

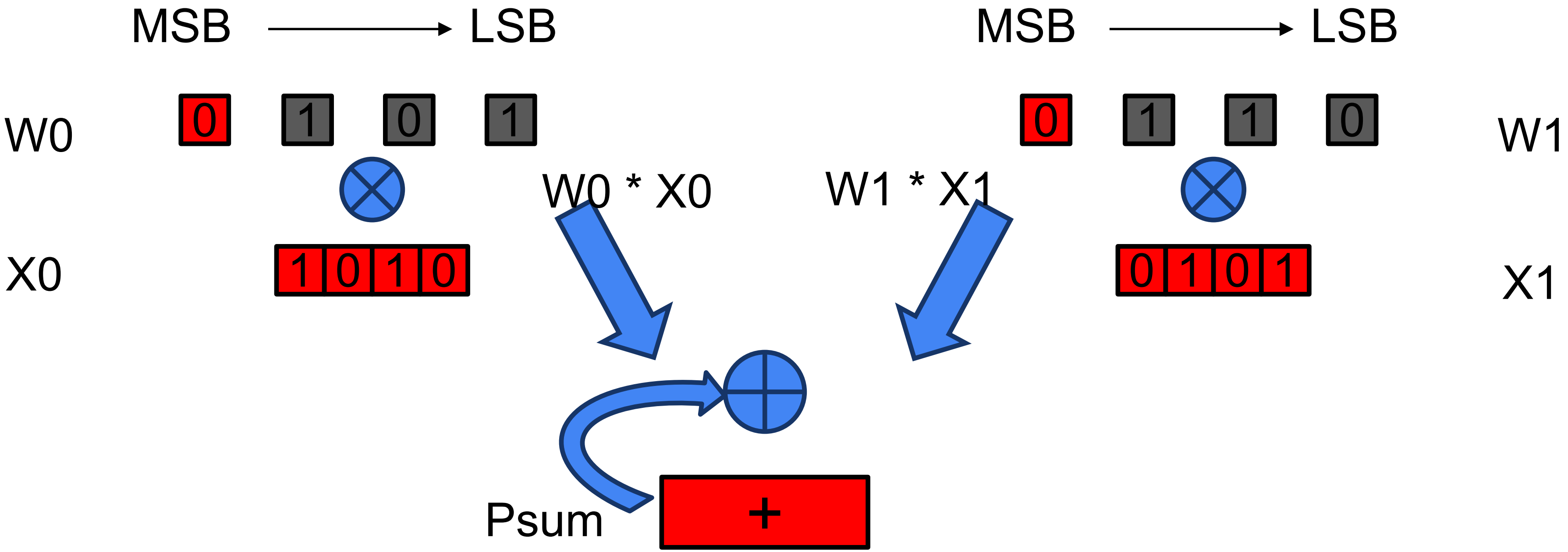
Early Termination: **Fewer Bits** of Weights For Neg Outputs



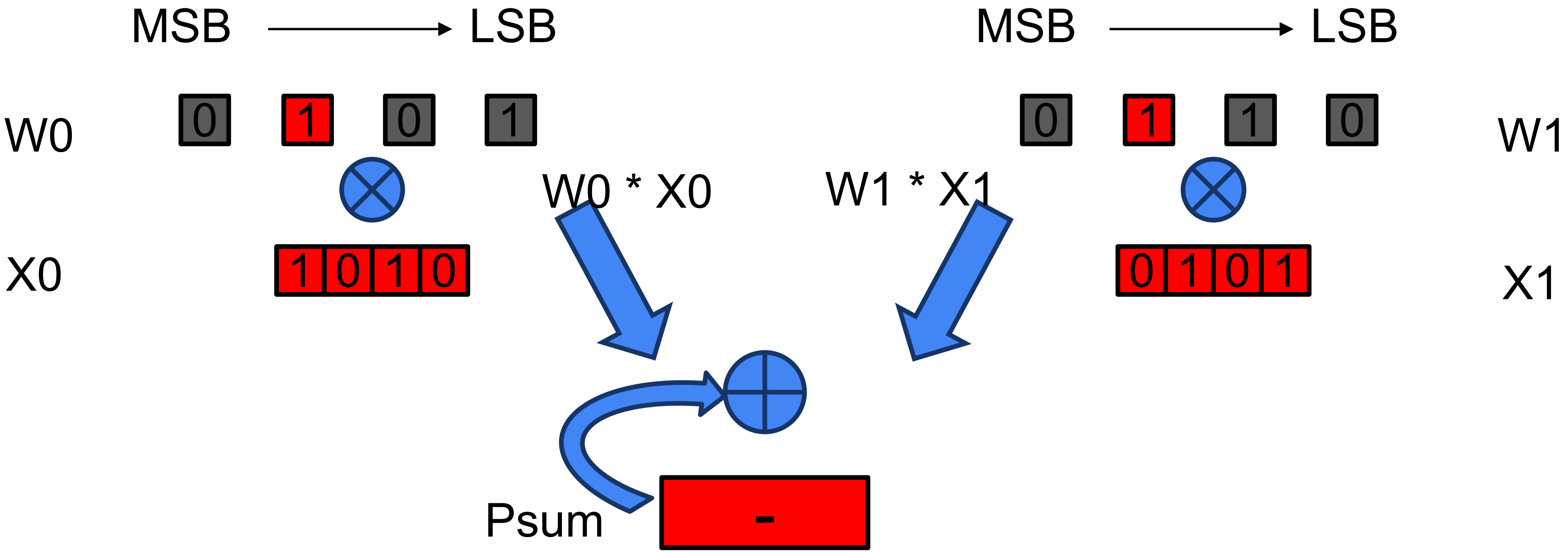
Normal Bit-serial MAC



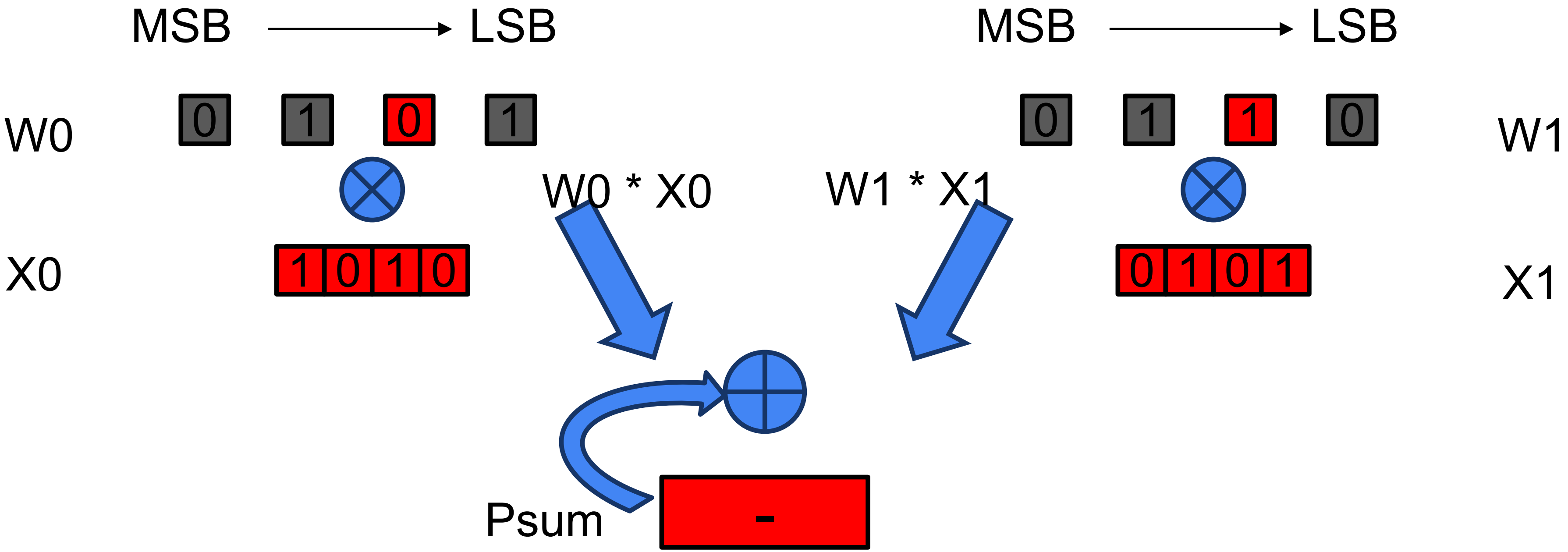
Normal Bit-serial MAC (Step 1 / 4)



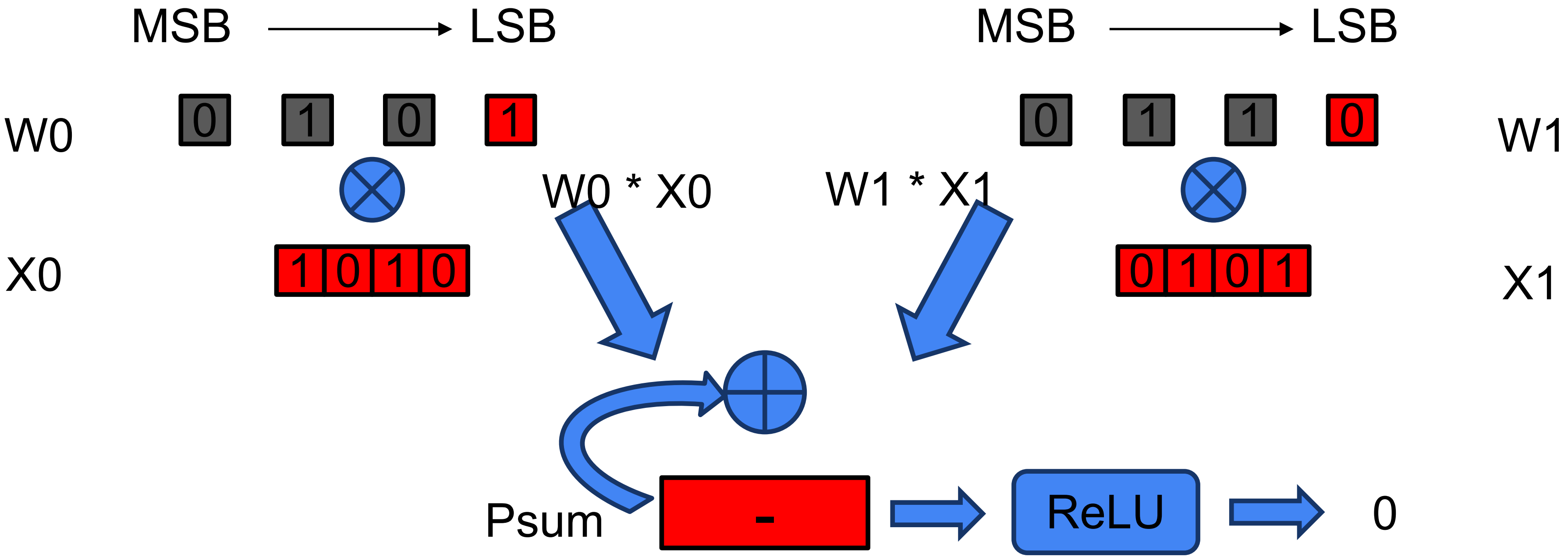
Normal Bit-serial MAC (Step 2 / 4)



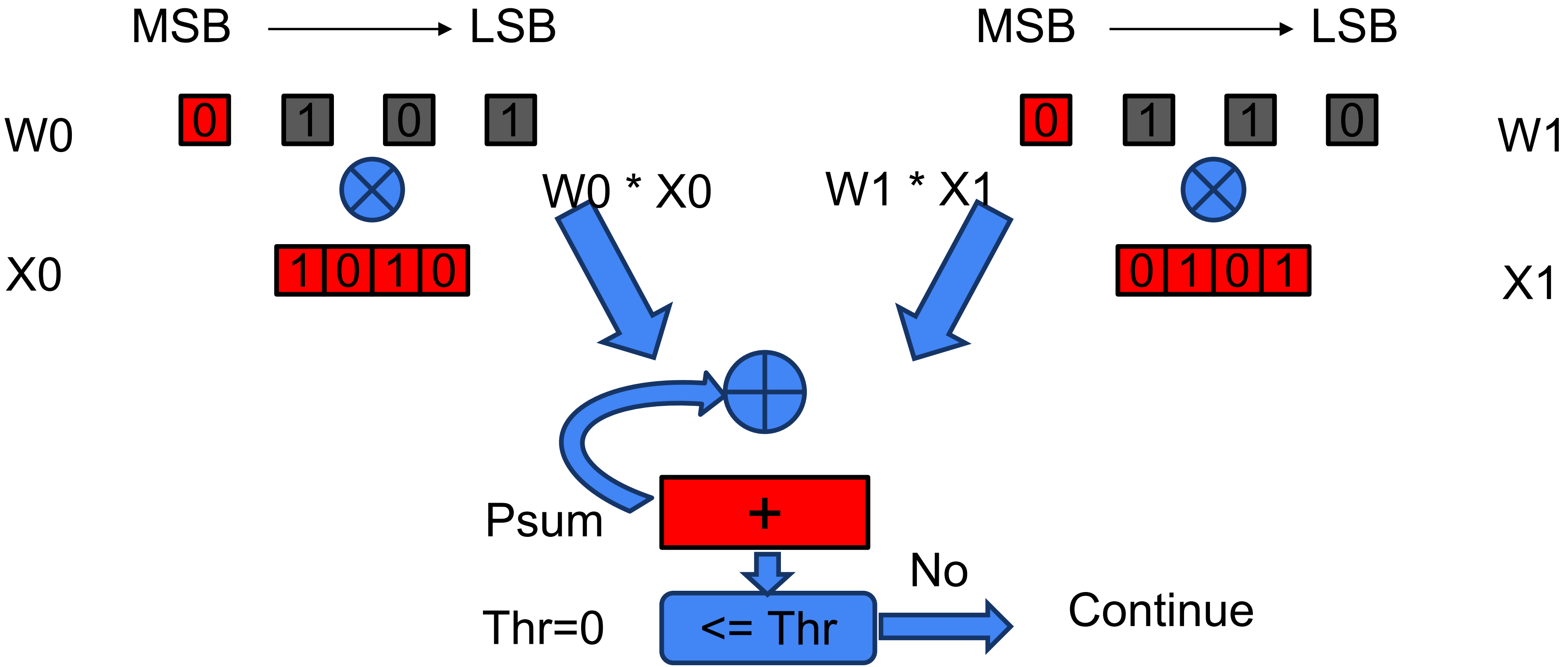
Normal Bit-serial MAC (Step 3 / 4)



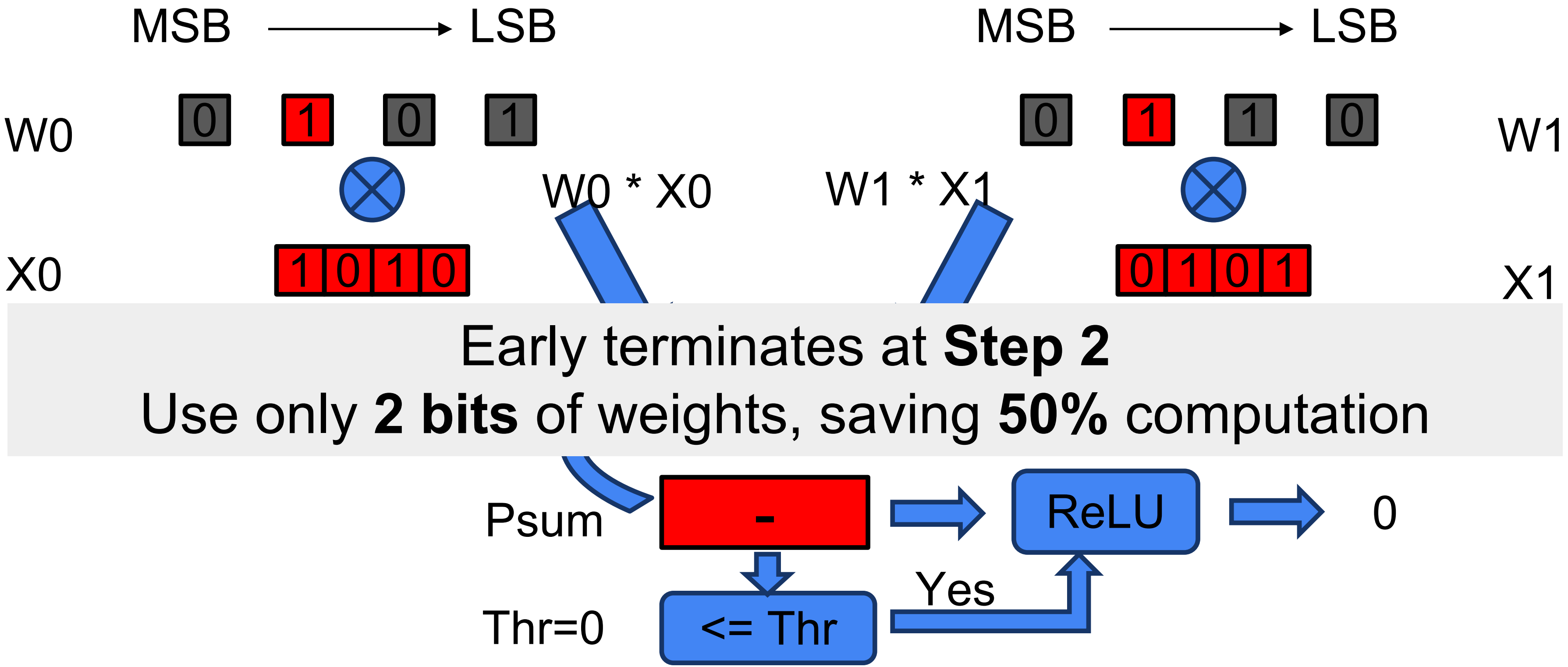
Normal Bit-serial MAC (Step 4 / 4)



Bit-serial MAC With **Early Termination** (Step 1)



Bit-serial MAC With **Early Termination** (Step 2)



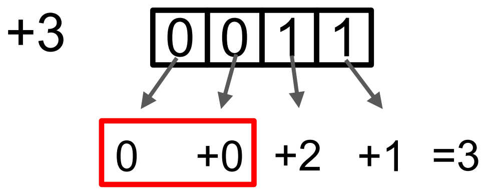
Early terminates at **Step 2**

Use only **2 bits** of weights, saving **50%** computation

Encoding For Weights

Existing Encodings

2’s complement



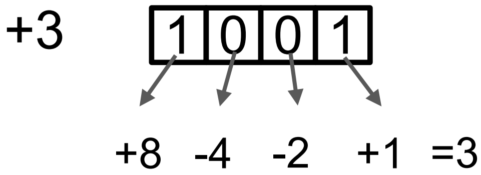
Needs at least first **6 bit**

1’s complement

Needs at least first **4 bit**

for 8-bit weights

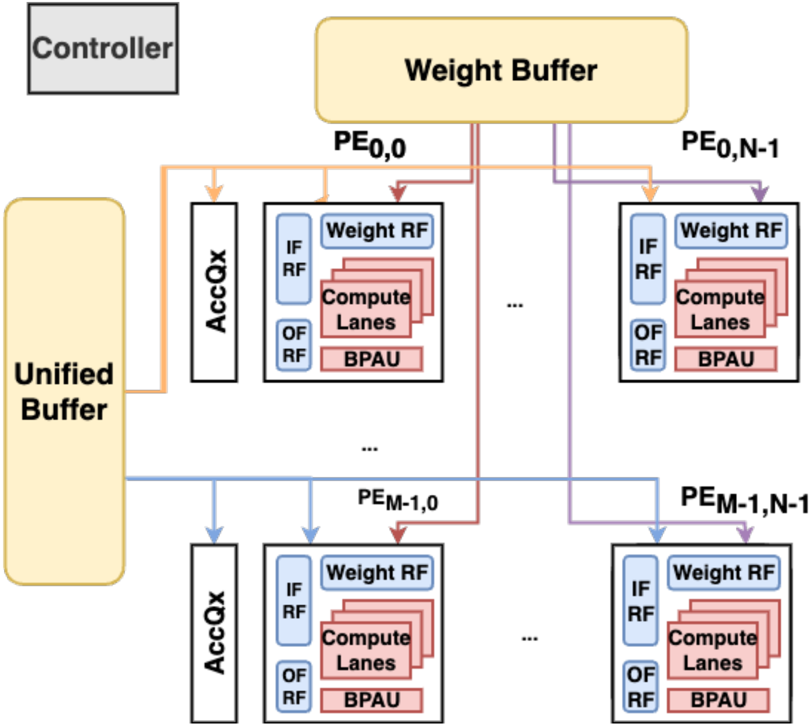
BitSET Encoding



Needs the **1st** bit only

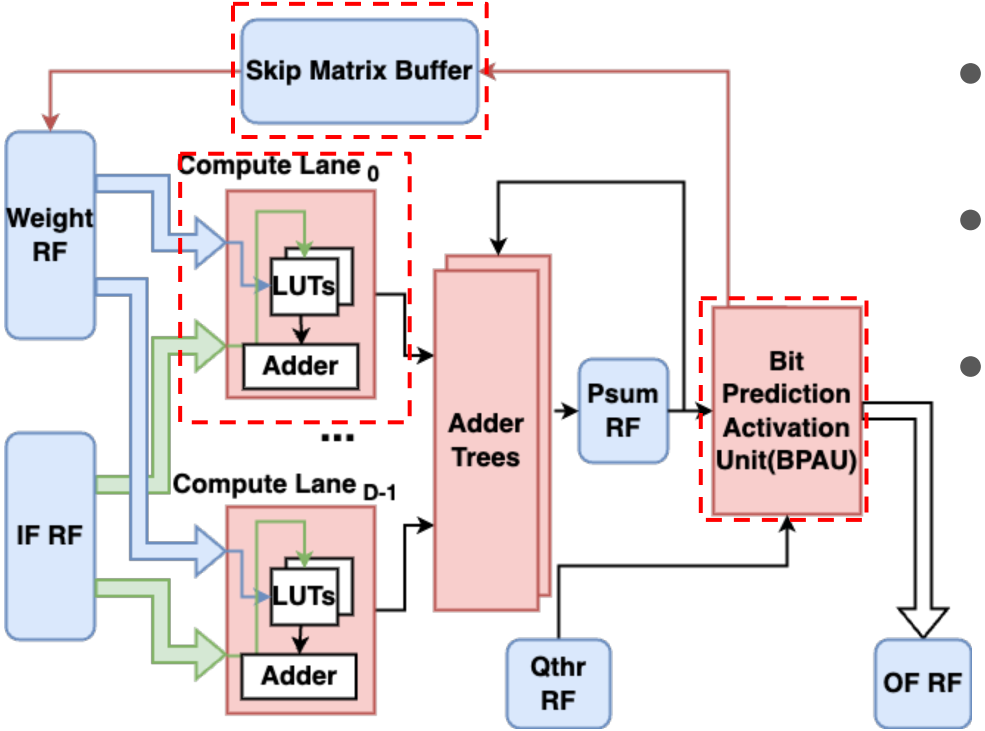
for 8-bit weights

Architecture Overview



- Architecture:
  - 2D MXN PE array
  - Unified Buffer (IF/ OF)
  - Weight Buffer
- Dataflow: Output Stationary (OS)
- Reduce **workload imbalance**: Double buffering

PE Microarchitecture



- **Compute Lane** uses LUT for bit-serial MAC operation
- **BPAU** compares Psum with Thr and send terminate signal
- **Skip Matrix Buffer** store the information of whether to skip the corresponding weight

Experiment Setup

Workloads

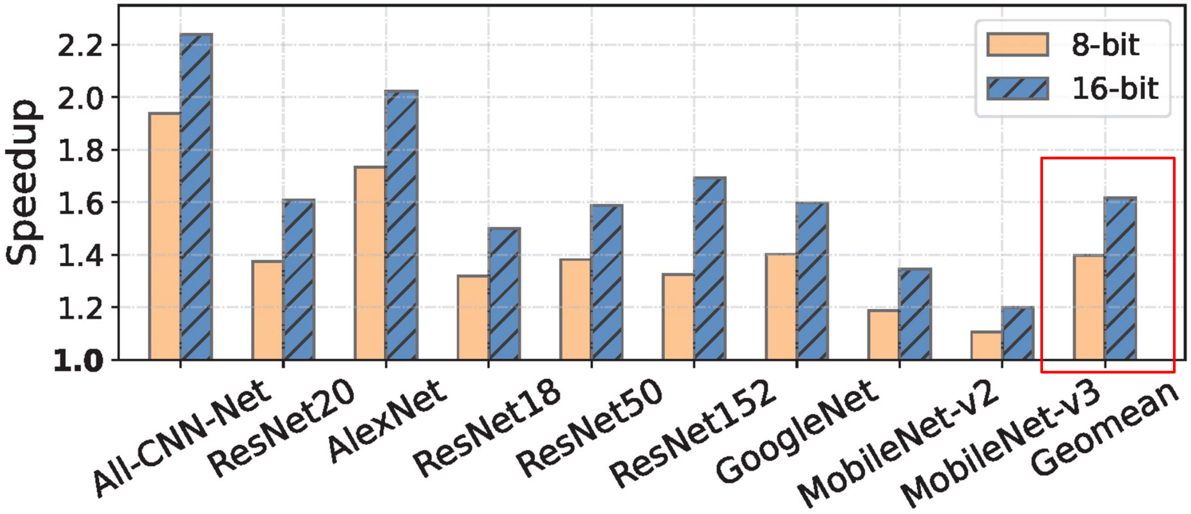
Datasets	CNN models
CIFAR-10	All-CNN-Net, ResNet20
ImageNet	AlexNet, ResNet18/50/152, GoogleNet, MobileNet-v2/v3

Hardware Implementation

- Implemented in SystemVerilog
- Synopsys Design Compiler with 45nm Nangate Open-cell Library
- Cycle-level accurate simulator to model latency
- Baseline design: UNPU[1], A bit-serial CNN accelerator
- Area overhead of BitSET is 2.3% over UNPU baseline

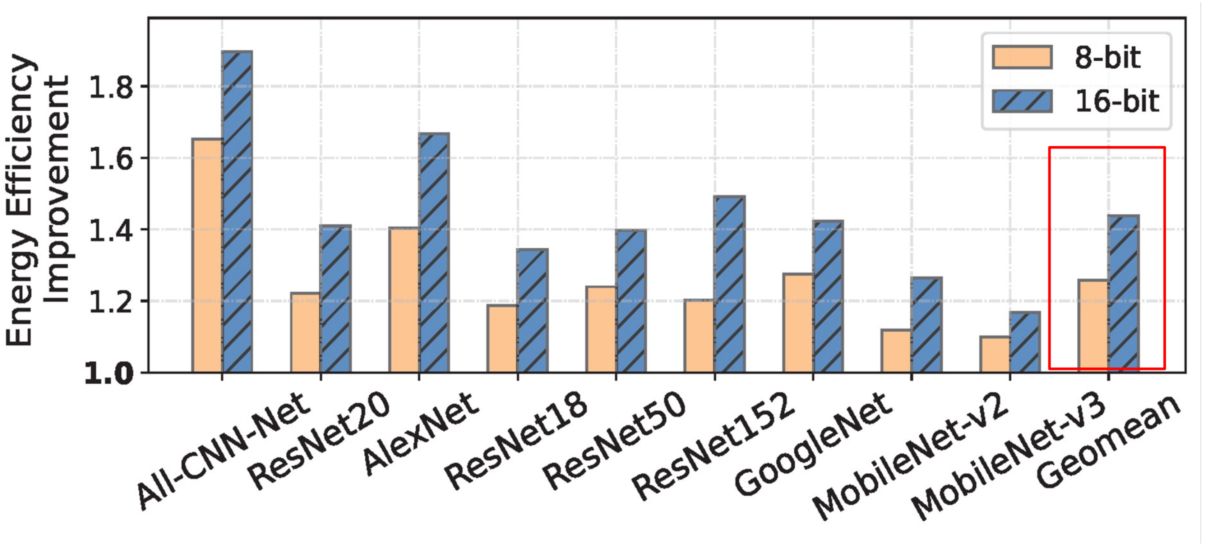
[1]Lee, Jinmook, et al. "UNPU: An energy-efficient deep neural network accelerator with fully variable weight bit precision." IEEE Journal of Solid-State Circuits 54.1 (2018): 173-185.

Speedup Over UNPU (precision = 8 bit or 16 bit)



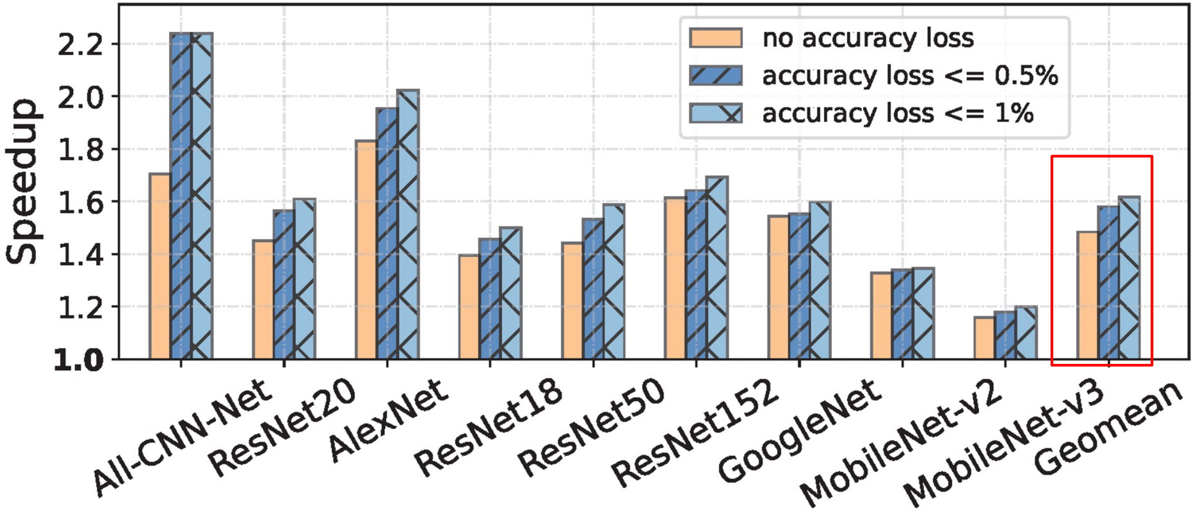
On average, 1.6x speedup over UNPU due to 52% bit-level MAC operation reduction

Energy Efficiency Improvement Over UNPU



On average, 1.4x energy efficiency improvement over UNPU

Speedup With Different **Accuracy Loss Constraints**



As accuracy loss tolerance is relaxed more, the speedup increases

Conclusion

- BitSET leverages the runtime information to **predictively terminate bit-level computation early** in CNNs.
- BitSET is a **hardware-software co-design**, which includes an algorithm, an encoding and an accelerator.
- **1.6x** speedup and **1.4x** energy efficiency improvement when allowing 1% accuracy loss

Q & A ?