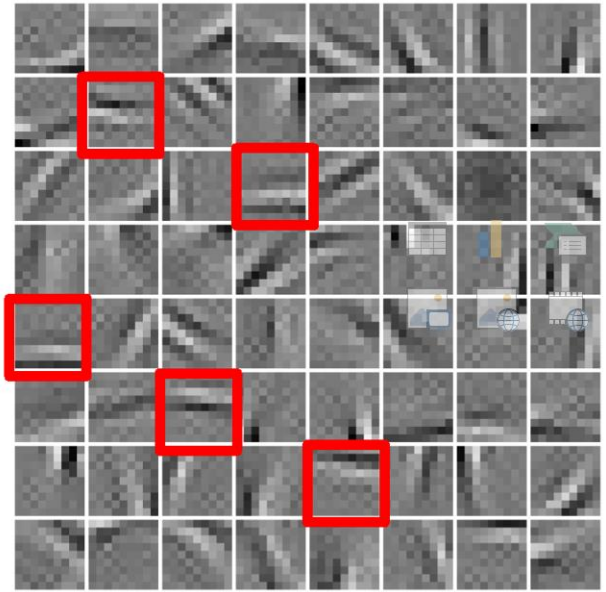
ZeNA: Zero-Aware Neural Network Accelerator

Sungjoo Yoo
Seoul National University

http://cmalab.snu.ac.kr

Redundancy in CNNs

Trained CNNs have significant amount of redundancy

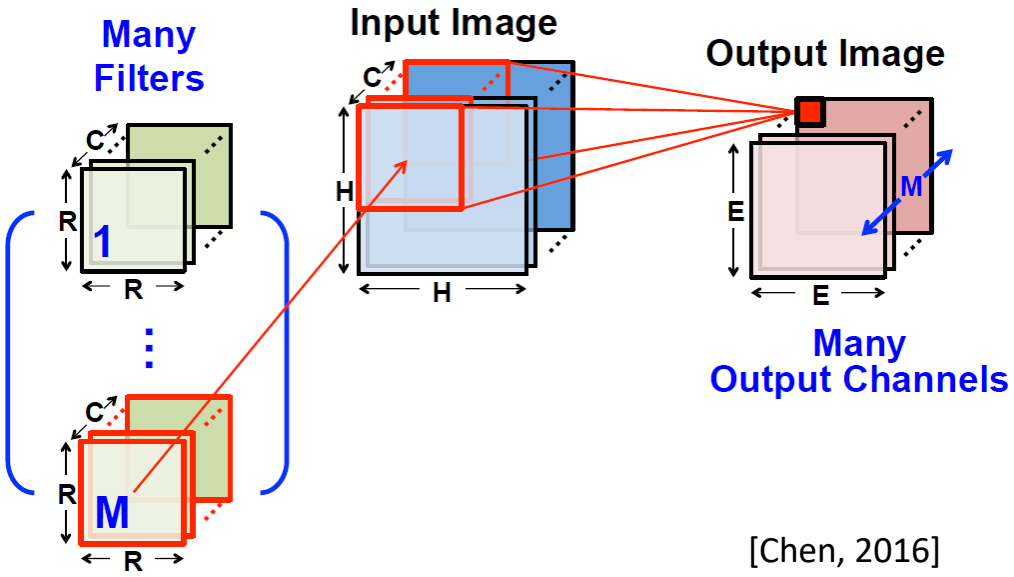


Redundant filters [LeCun]

	Pruning	ReLU
Layer	Zero Weight [%] [1]	Zero Activation [%] [2]
conv1	15.7	0
conv2	62.1	50.9
conv3	65.4	76.3
conv4	62.8	61.8
conv5	63.1	59.0

Zero values (in AlexNet)

Basic Idea: Skipping Computation with Zero Input



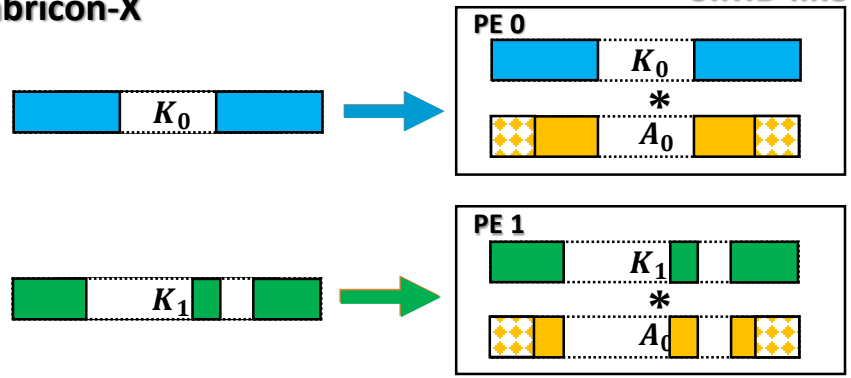
[Chen, 2016]

- In a convolutional layer, producing an output value in the three-dimensional output feature map requires $\sim 10^3$ multiplications and additions
- Basic idea: skipping zero-input multiplications

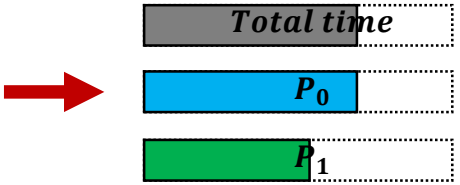
Previous Works Exploit Only One of the Two

Cambricon-X

SIMD like

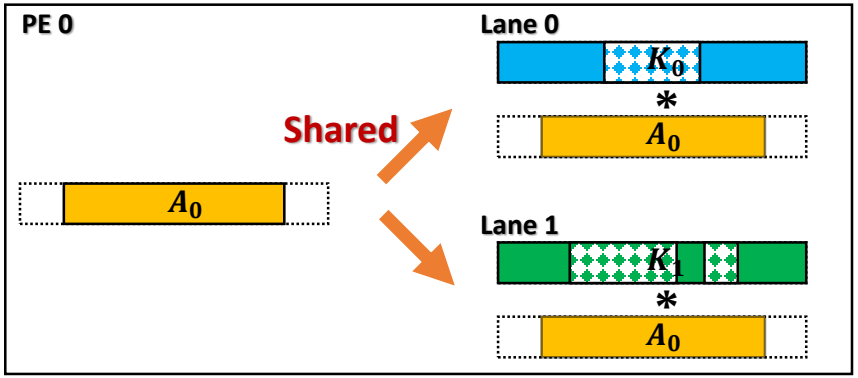


Proportional to non-zero weights

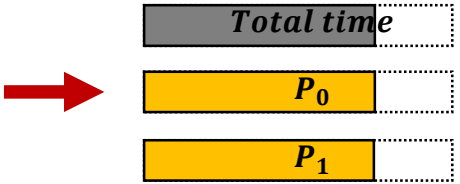


Cnvlutin

SIMD like

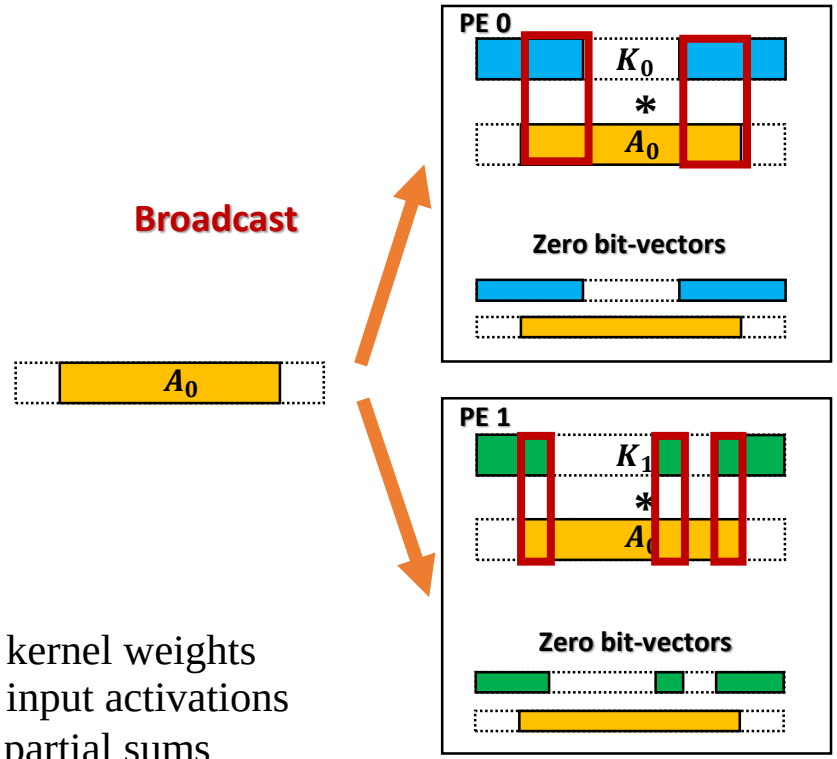


Proportional to non-zero activations



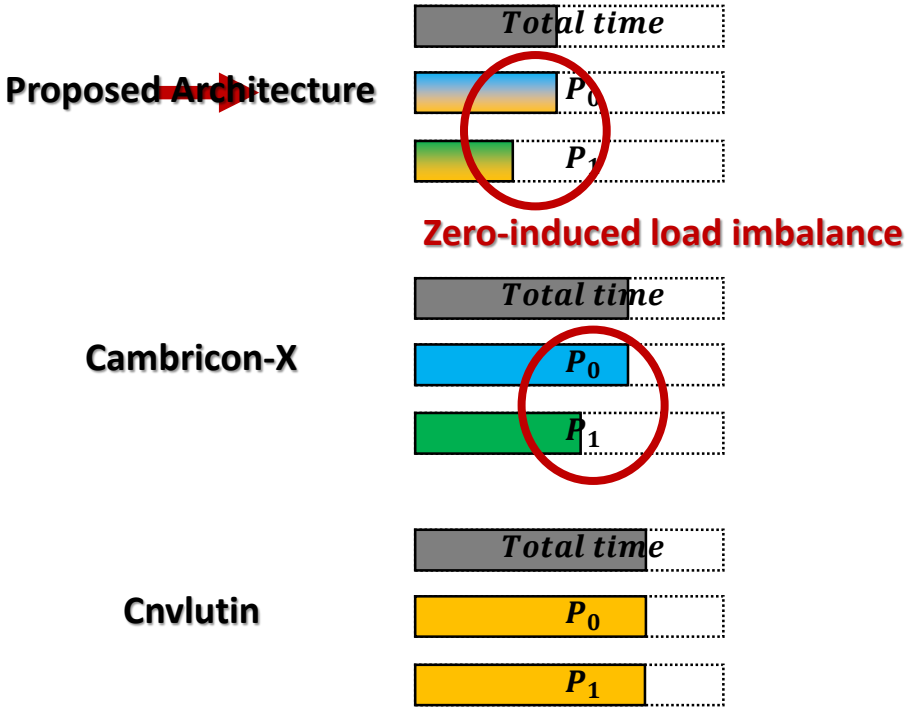
K: kernel weights
A: input activations
P: partial sums

Proposed Architecture to Exploit Both Zero Values



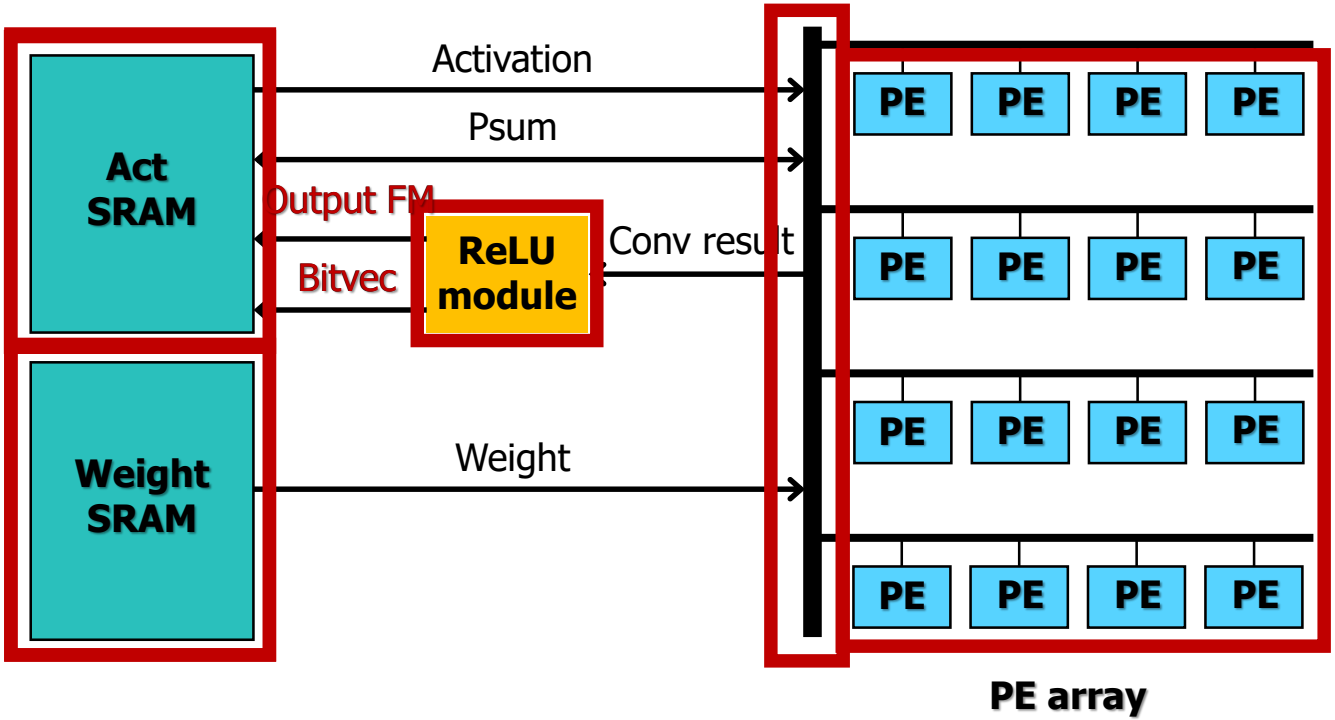
K: kernel weights
A: input activations
P: partial sums

Proportional to intersection of non-zero inputs



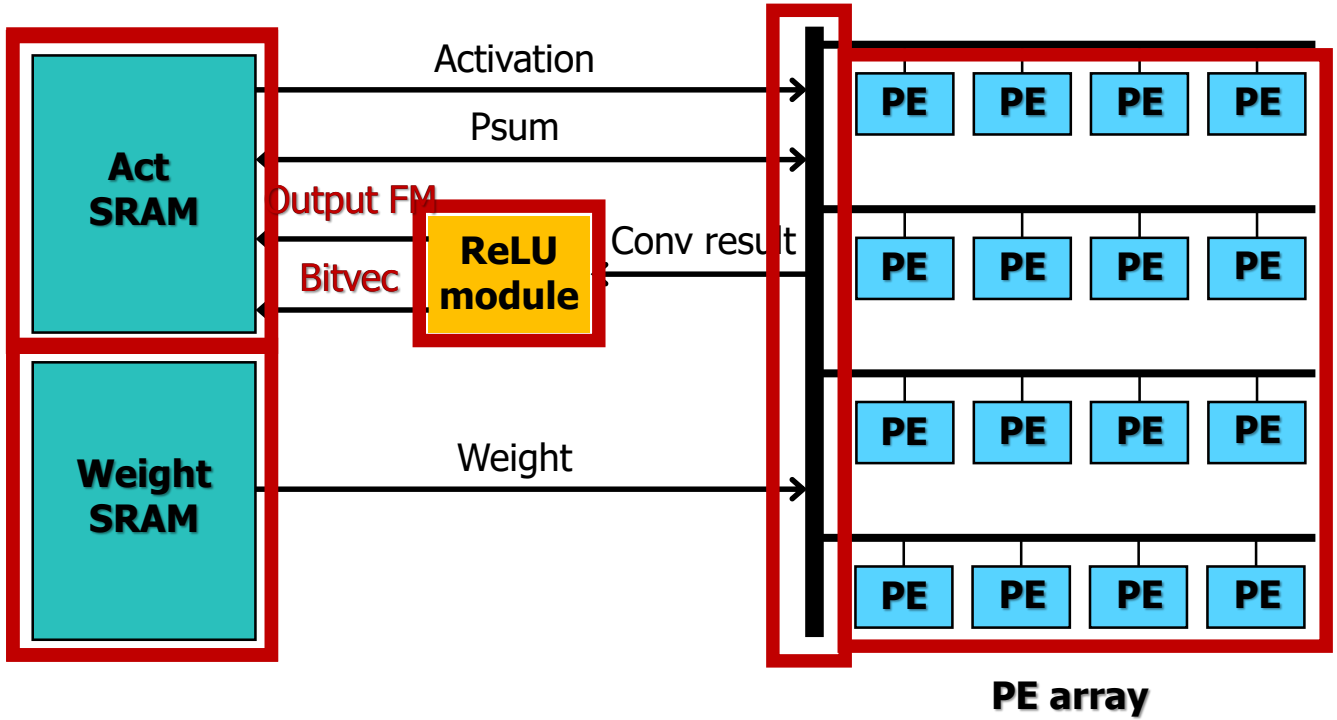
ZeNA v1.0

- **Act SRAM** stores
  - **Activations**
  - **Output partial sums**
  - **Output feature maps**
  - **Zero bit-vectors**
- **Bus** connects
  - **On-chip SRAM** and **PE array**

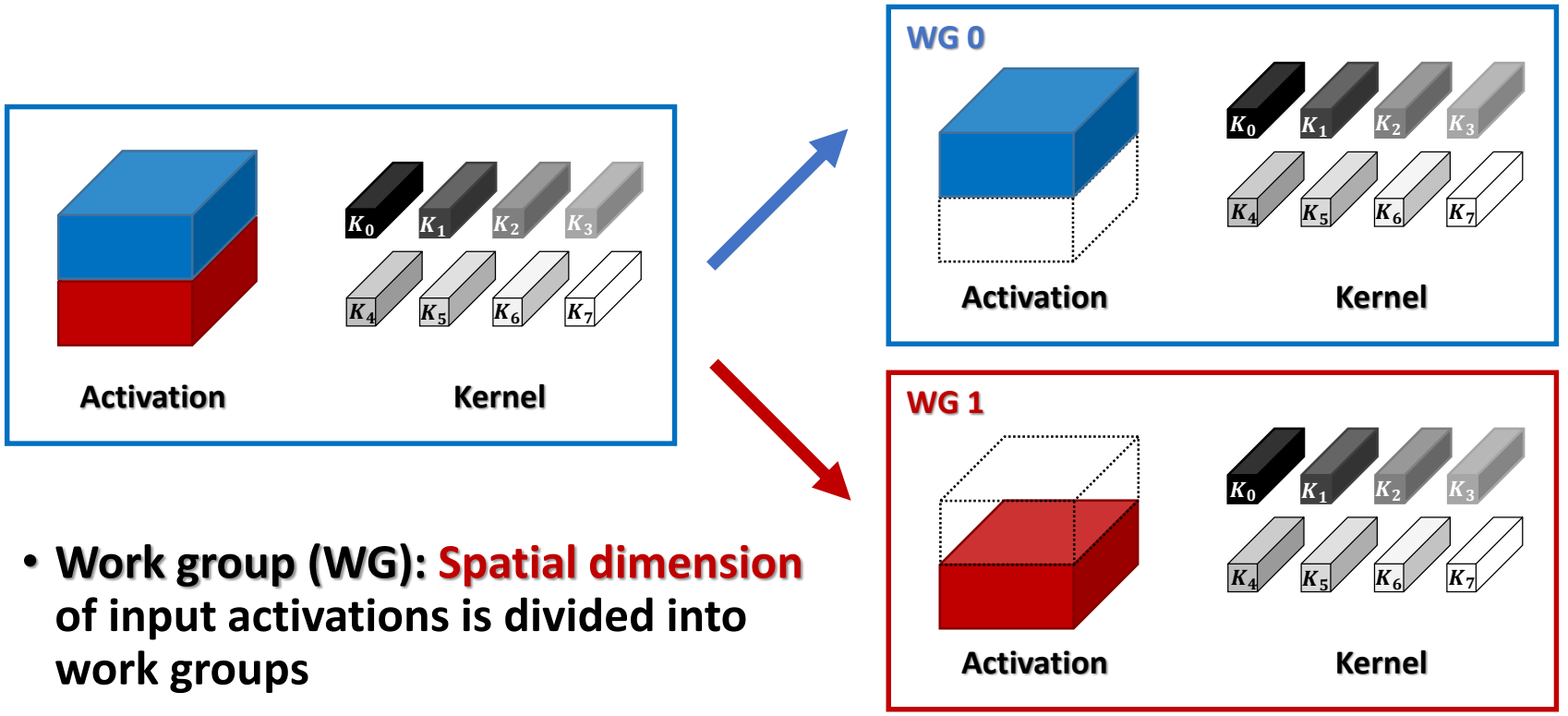


ZeNA v1.0

- **ReLU module generates**
  - **Output feature maps**
  - **Zero bit-vectors**
- **Weight SRAM stores**
  - **Kernel weights**
  - **Zero bit-vectors**

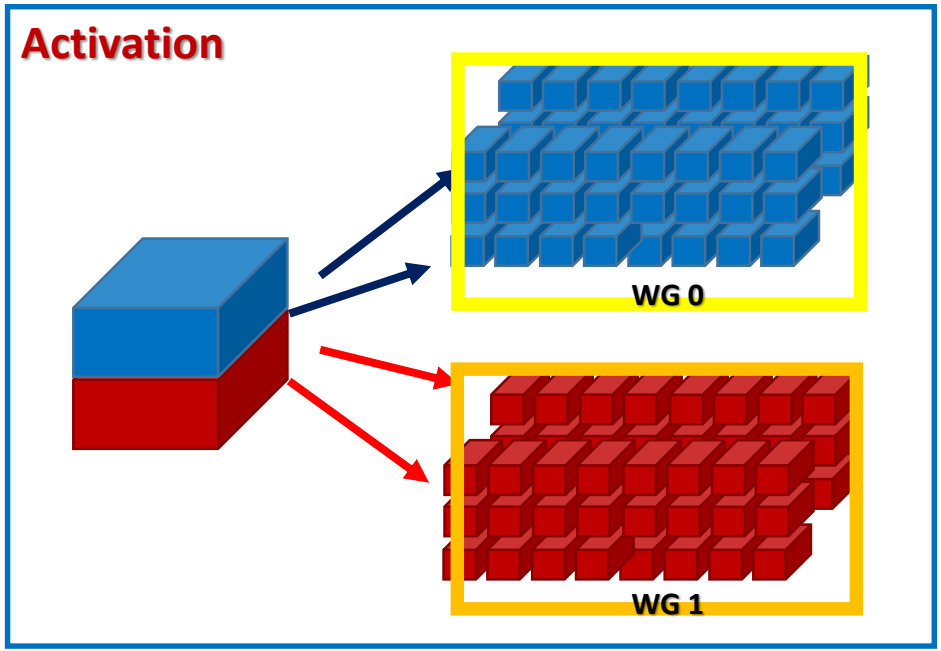


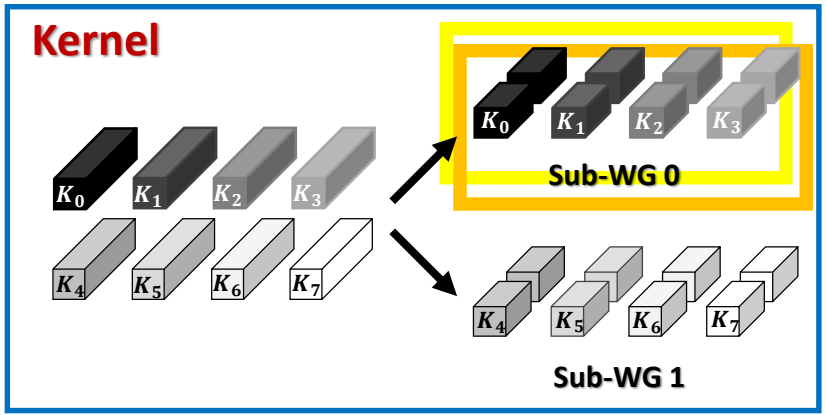
Work Group (WG)

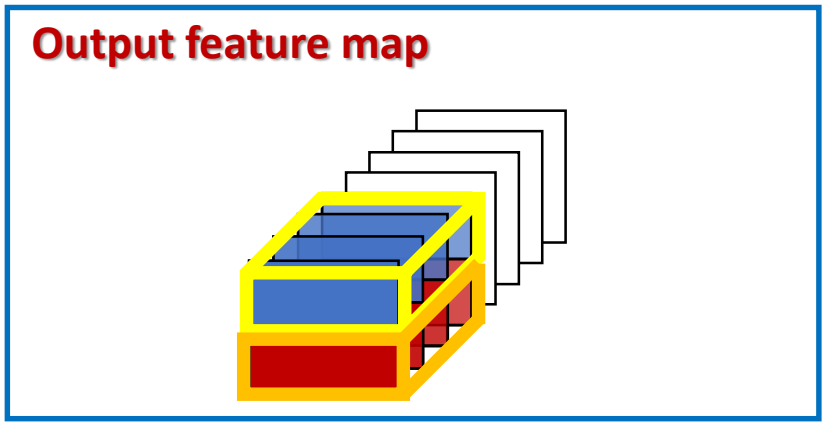


- **Work group (WG): Spatial dimension of input activations is divided into work groups**

Computation Procedure

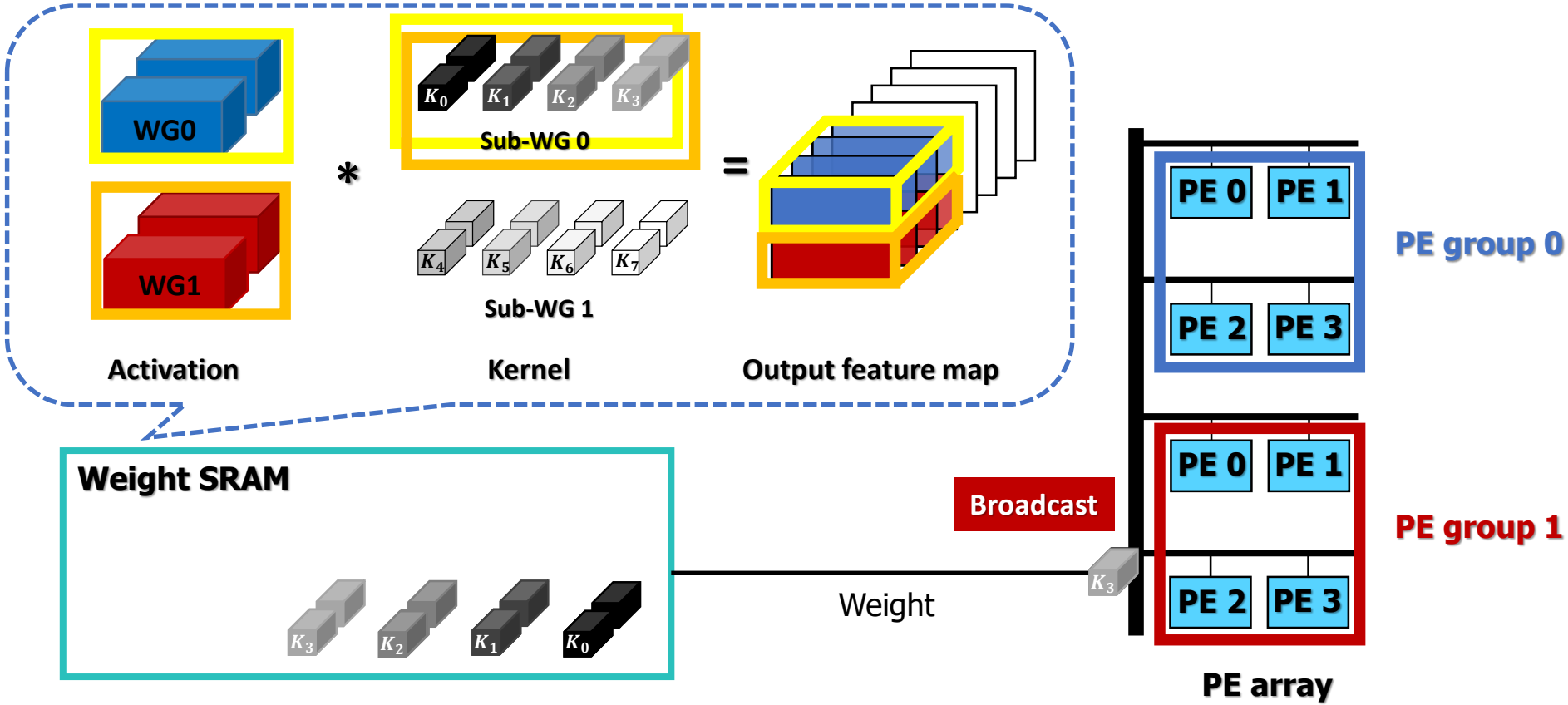




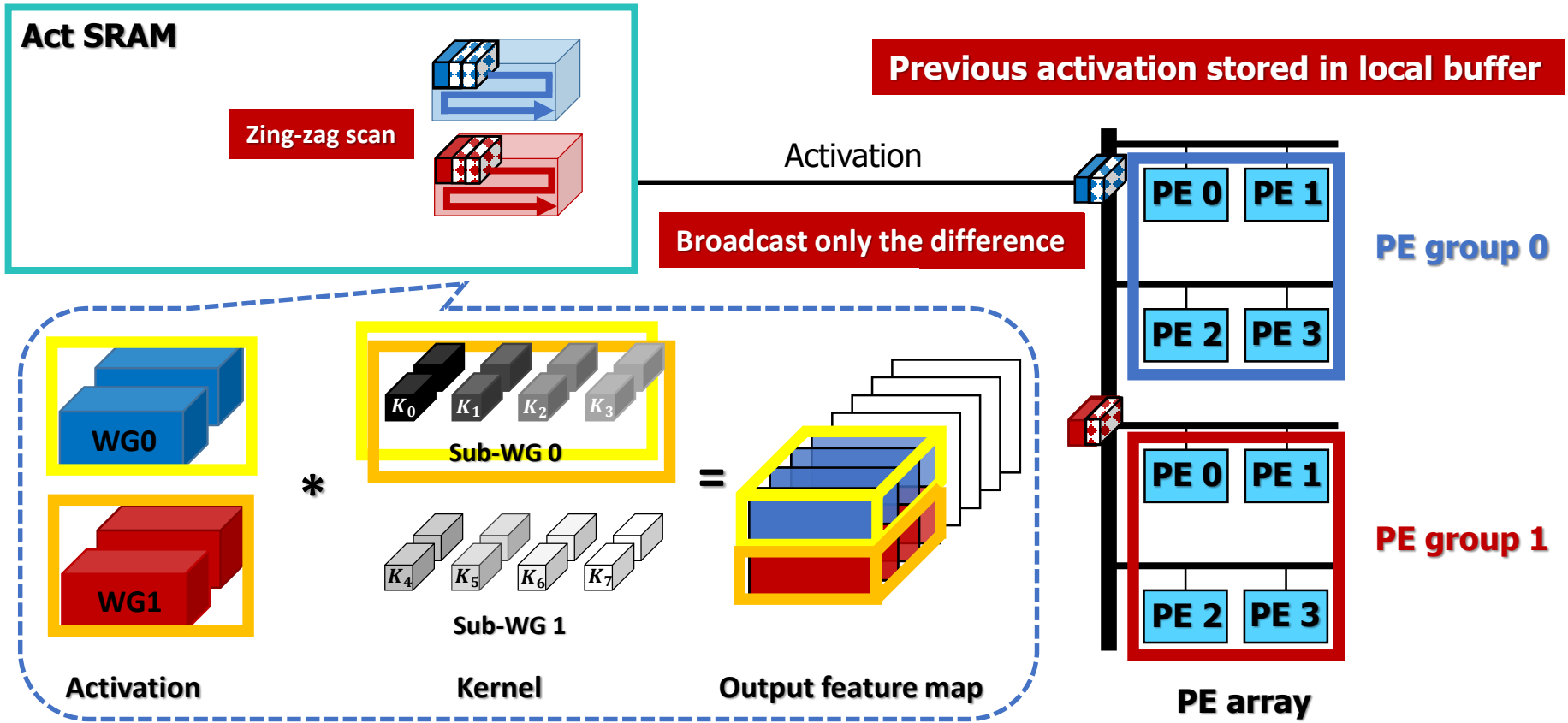


- Our accelerator performs convolution with activation and kernel tiles **iteratively** to compute output feature maps

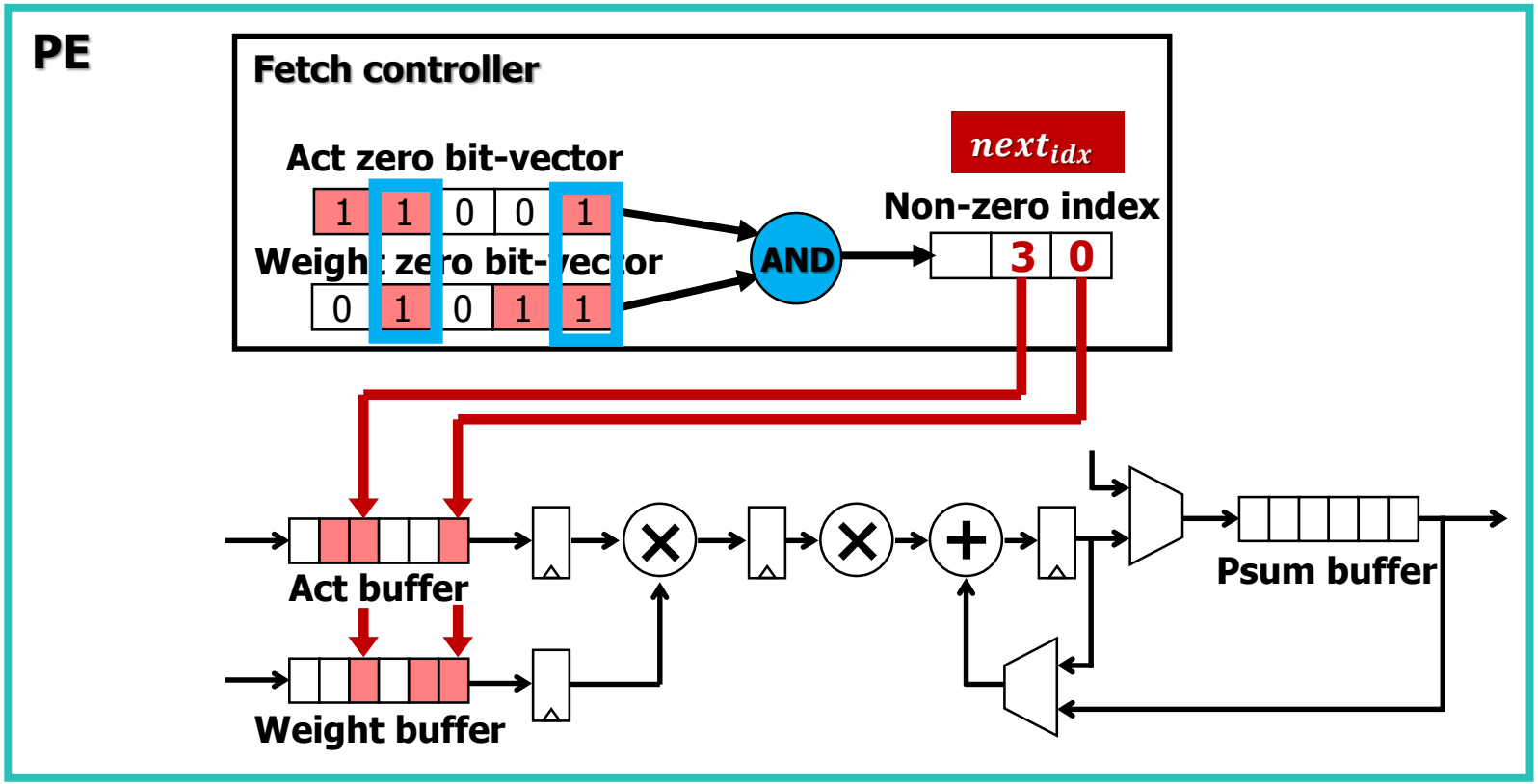
Data Flow and Computation: Kernel Broadcast



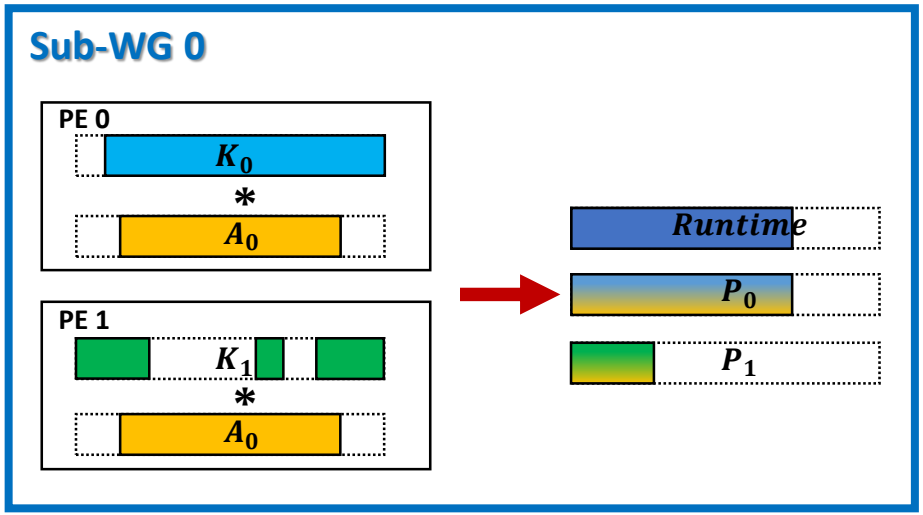
Data Flow and Computation: Activation Broadcast

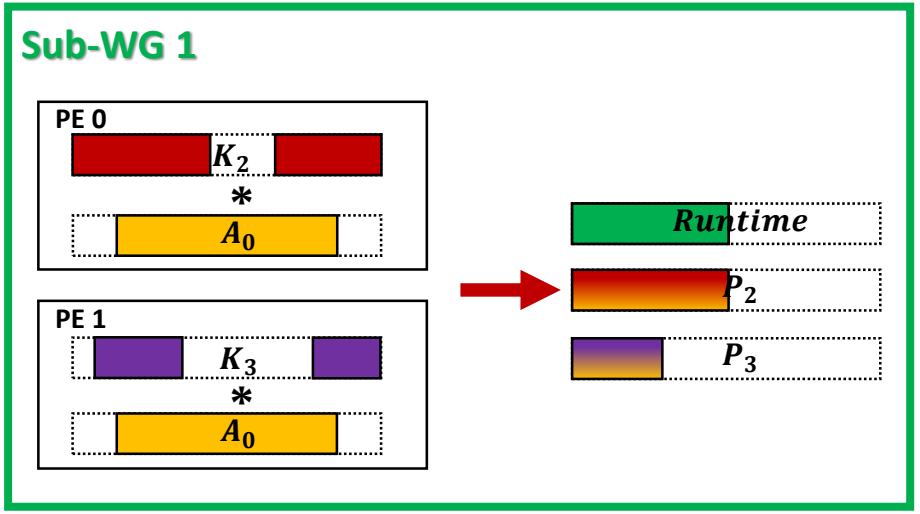


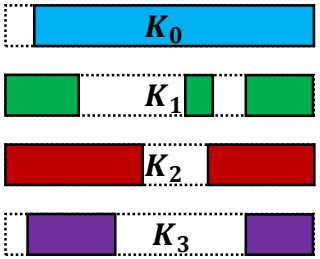
Zero-aware Processing Element (PE)



Zero-aware Kernel Allocation



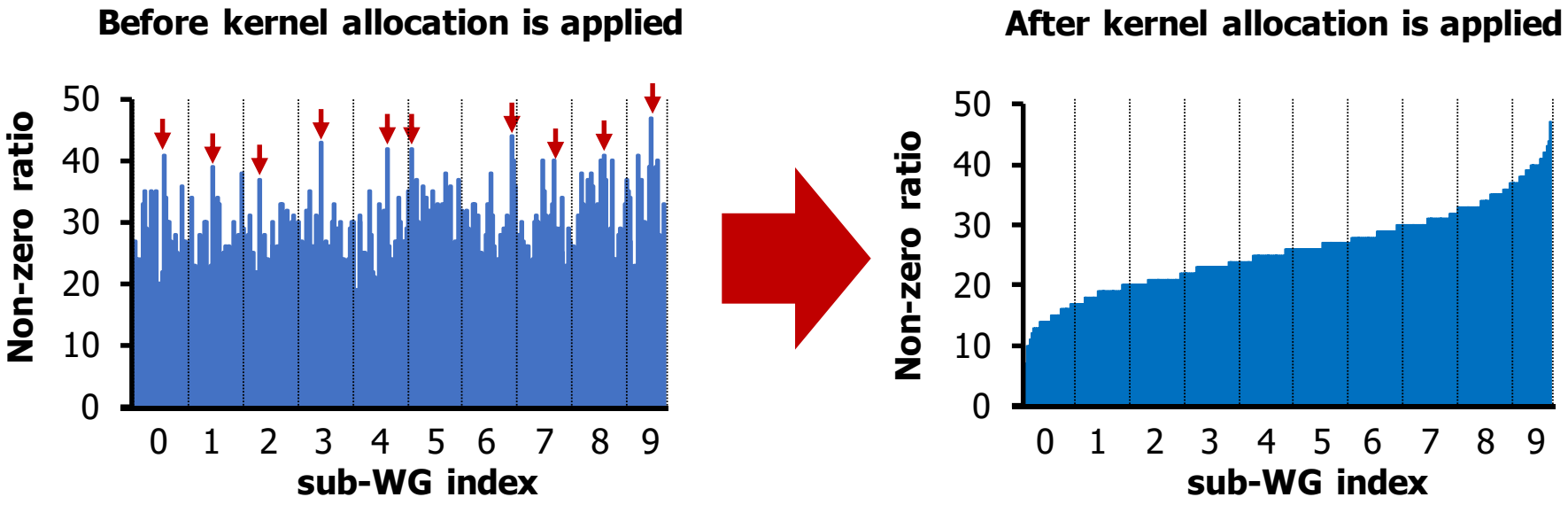




Before kernel allocation is applied

- **Typically, kernel weights are allocated to PEs based on the kernel index**

Zero-aware Kernel Allocation

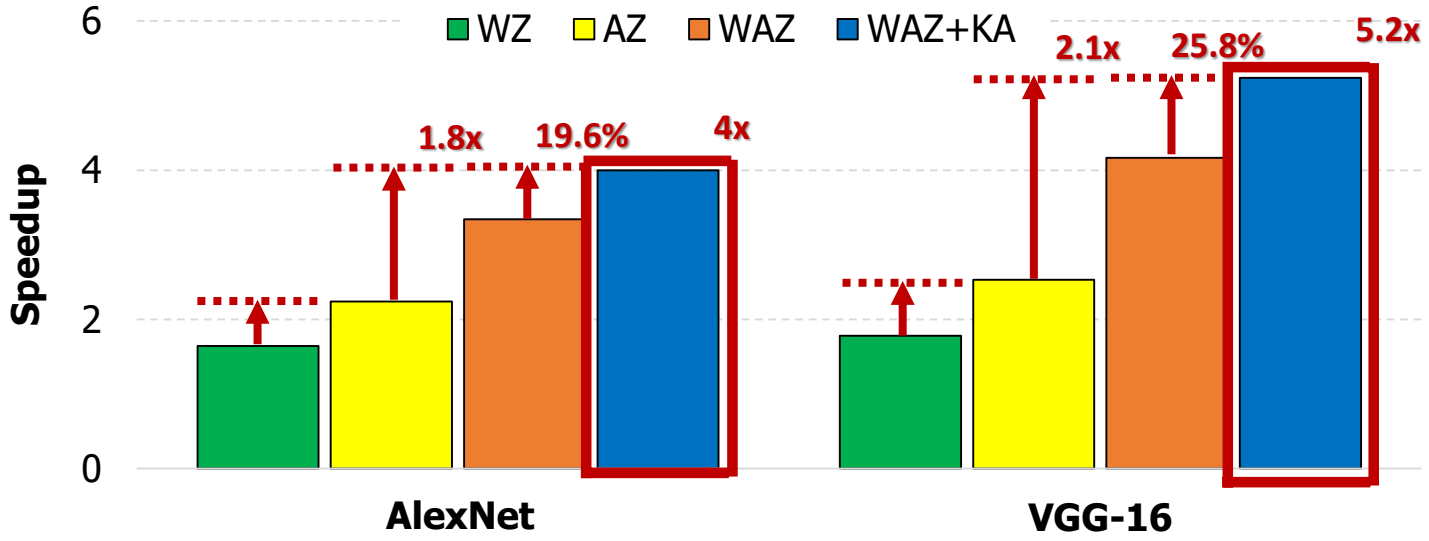


- Non-zero weight ratio of 384 kernel tiles (size of 3×3×14) of conv3 layer in the pruned AlexNet

Evaluation Methodology

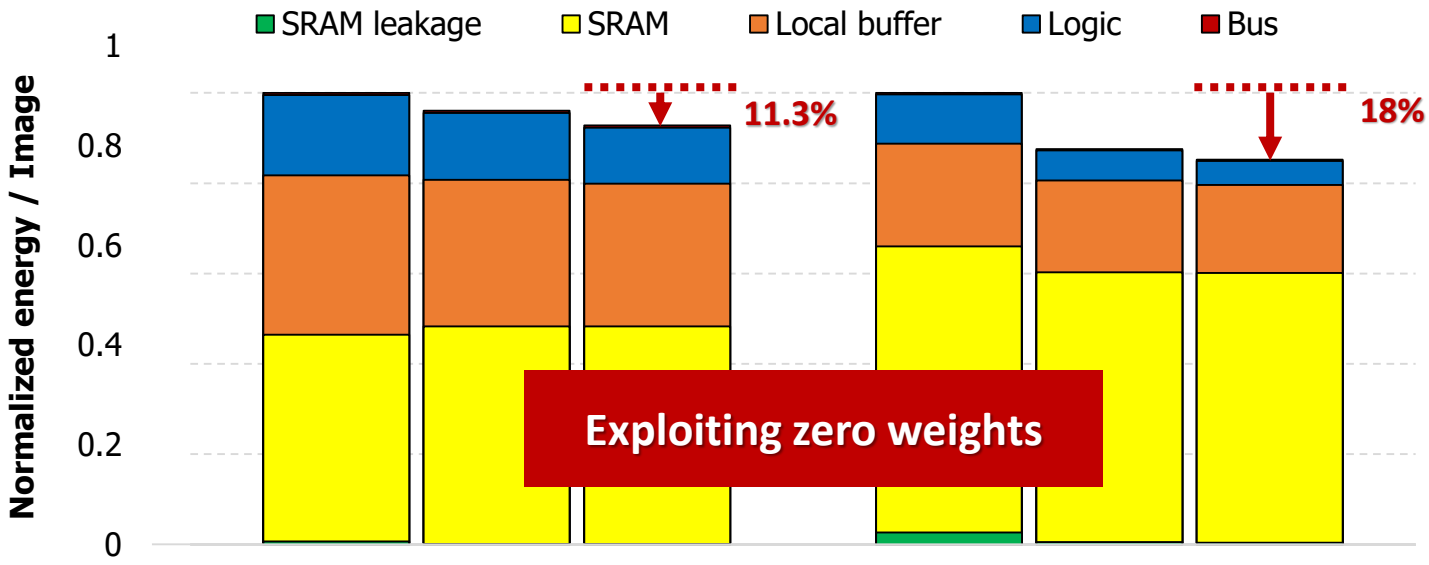
- **FixedPoint (16-bit fixed-point) and LogQuant (5-bit LogQuant) implementation**
- ***TSMC 65nm* synthesis library (200MHz)**
- **Chip layout using *Synopsys Astro***
  - **Area overhead: FixedPoint (8.5%) / LogQuant (6.9%)**
- ***Prime-Time PX* for power estimation / *CACTI v6.0* for SRAM energy and area**
- **AlexNet and VGG-16**
- **We run our accelerator with four modes:**
  - **Zero-weight-aware mode (WZ)**
  - **Zero-activation-aware mode (AZ, corresponding to Cnvlutin)**
  - **Zero-weight- and zero-activation-aware mode (WAZ)**
  - **WAZ + zero-aware kernel allocation (WAZ+KA)**

16-BIT FIXEDPOINT: Performance



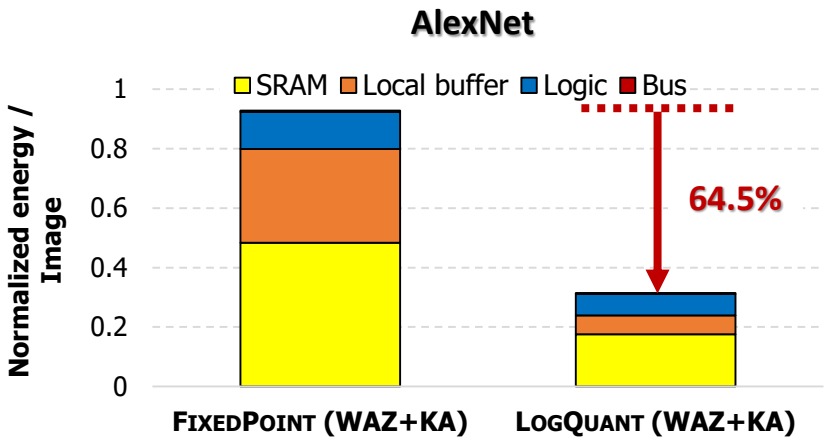
- **4x** (AlexNet) and **5.2x** (VGG-16) speed up ***w.r.t. Eyeriss***
- **1.8x** (AlexNet) and **2.1x** (VGG-16) speed up ***w.r.t. AZ (Cnvlutin)***
- **19.6%** (AlexNet) and **25.8%** (VGG-16) speed up ***w.r.t. WAZ***

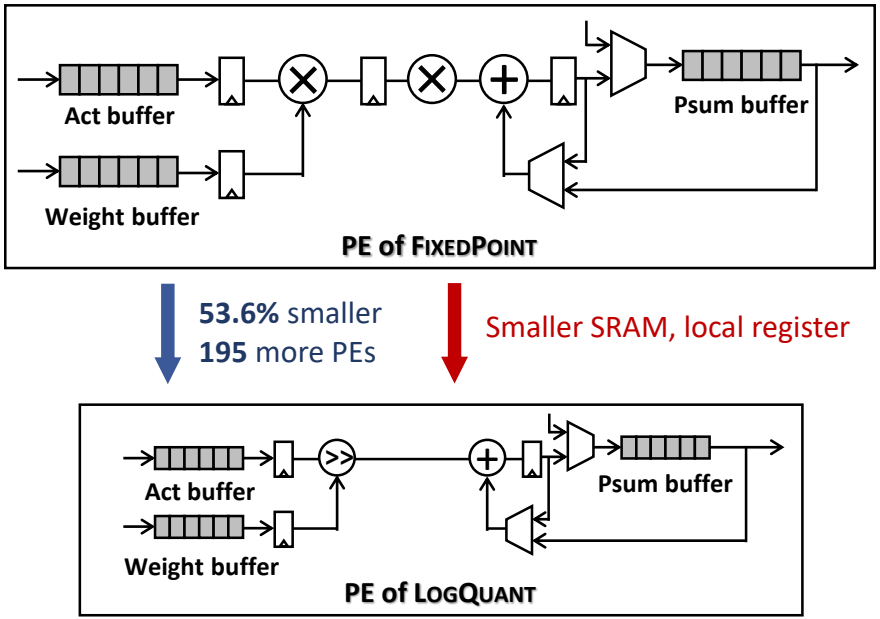
16-BIT FIXEDPOINT: Energy



- **11.3%** (AlexNet) and **18%** (VGG-16) energy reduction *w.r.t. Eyeriss*

5-BIT LOGQUANT





- **2.1x** (AlexNet) and **1.9x** (VGG-16) speed up *w.r.t. FIXEDPOINT (WAZ+KA)*
- **64.5%** (AlexNet) and **61.5%** (VGG-16) energy reduction *w.r.t. FIXEDPOINT (WAZ+KA)*

Summary

- Redundancy problem in CNNs
  - Abundant zero values in weights and activations
- ZeNA reduces runtime and energy consumption by
  - Skipping multiplications with zero input
  - Resolving zero-induced load imbalance problems
- Significant performance improvements (16-bit fixed point in ZeNA v1.1)
  - AlexNet: 4.4X speedup
  - VGG: 5.5X speedup
- Reference
  - D. Kim, et al., “A Novel Zero Weight/Activation-Aware Hardware Architecture of Convolutional Neural Network,” DATE 2017
  - D. Kim, et al., “ZeNA: Zero-Aware Neural Network Accelerator,” accepted for publication in IEEE Design & Test