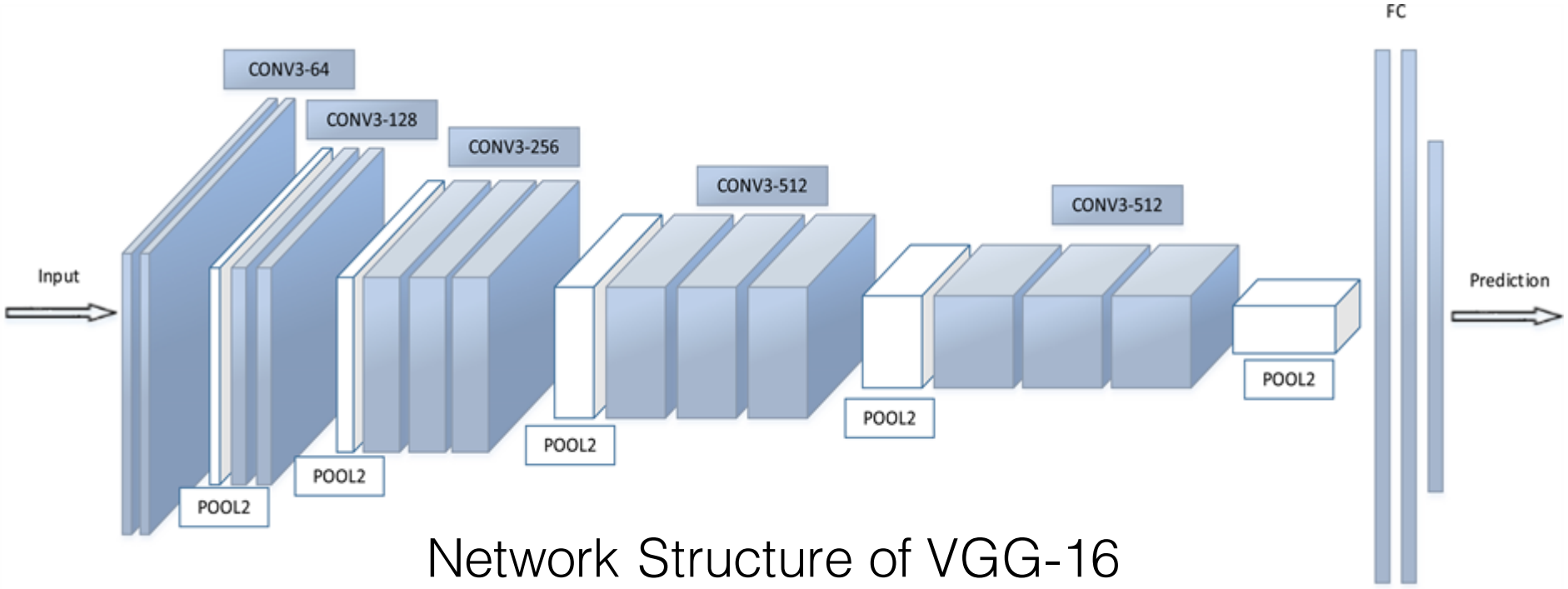
Supporting Compressed-Sparse Activations and Weights on SIMD-like Accelerator for Sparse Convolutional Neural Networks

Chien-Yu Lin and Bo-Cheng Lai

Institute of Electronics Engineering
National Chiao Tung University
國立交通大學

Convolutional Neural Network

- CNN now dominants visual recognition applications
  - Face recognition, object detection, autonomous vehicles…
- Major components: deep convolutional layers

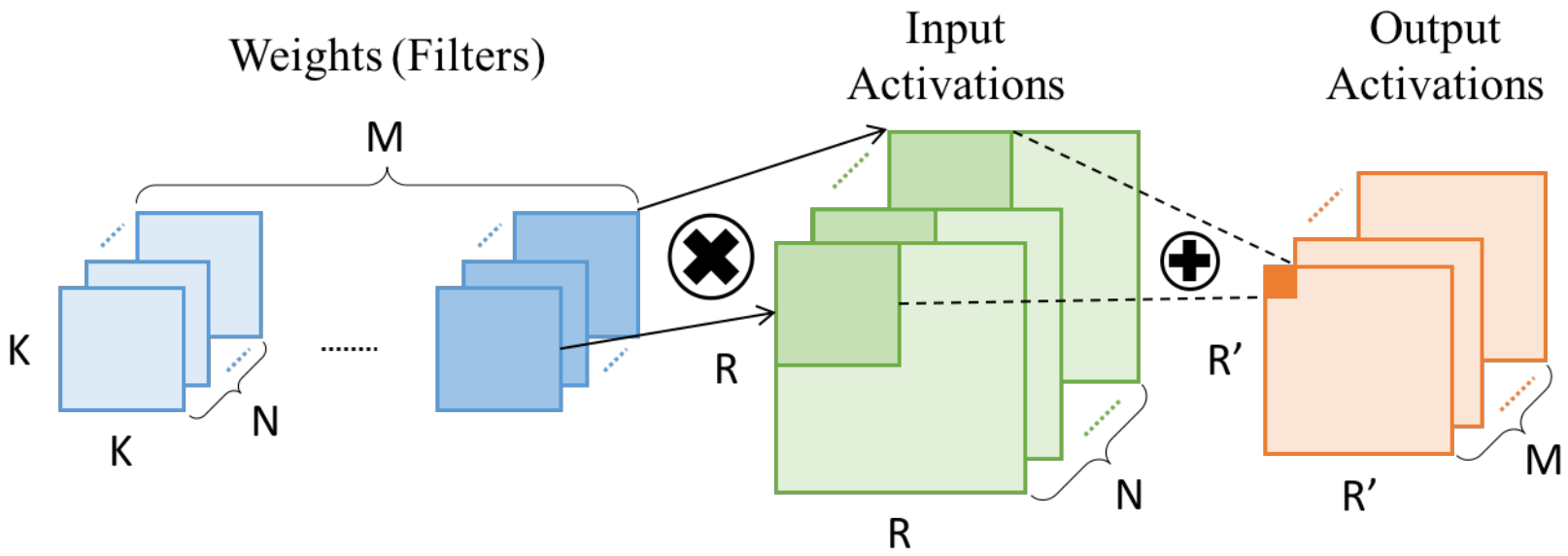


Network Structure of VGG-16

k. Simonyan et al., Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR 2015
http://file.scirp.org/Html/4-7800353_65406.htm

Convolutional Layer

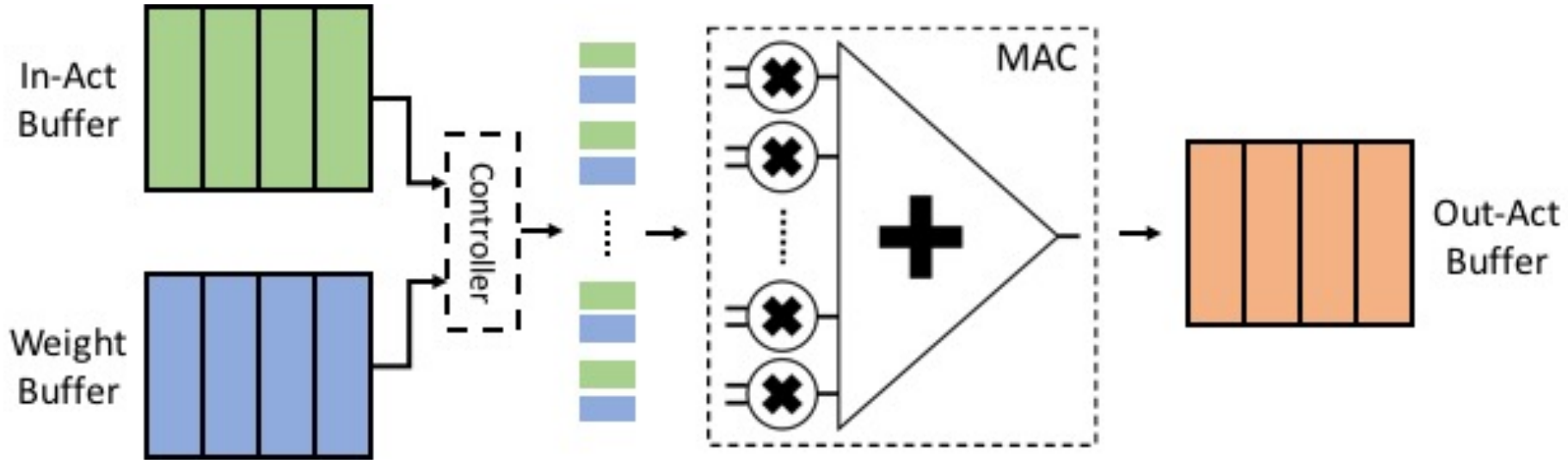
- A lot of parallel multiplications and additions



Computations of a conv layer

CNN Acceleration with SIMD

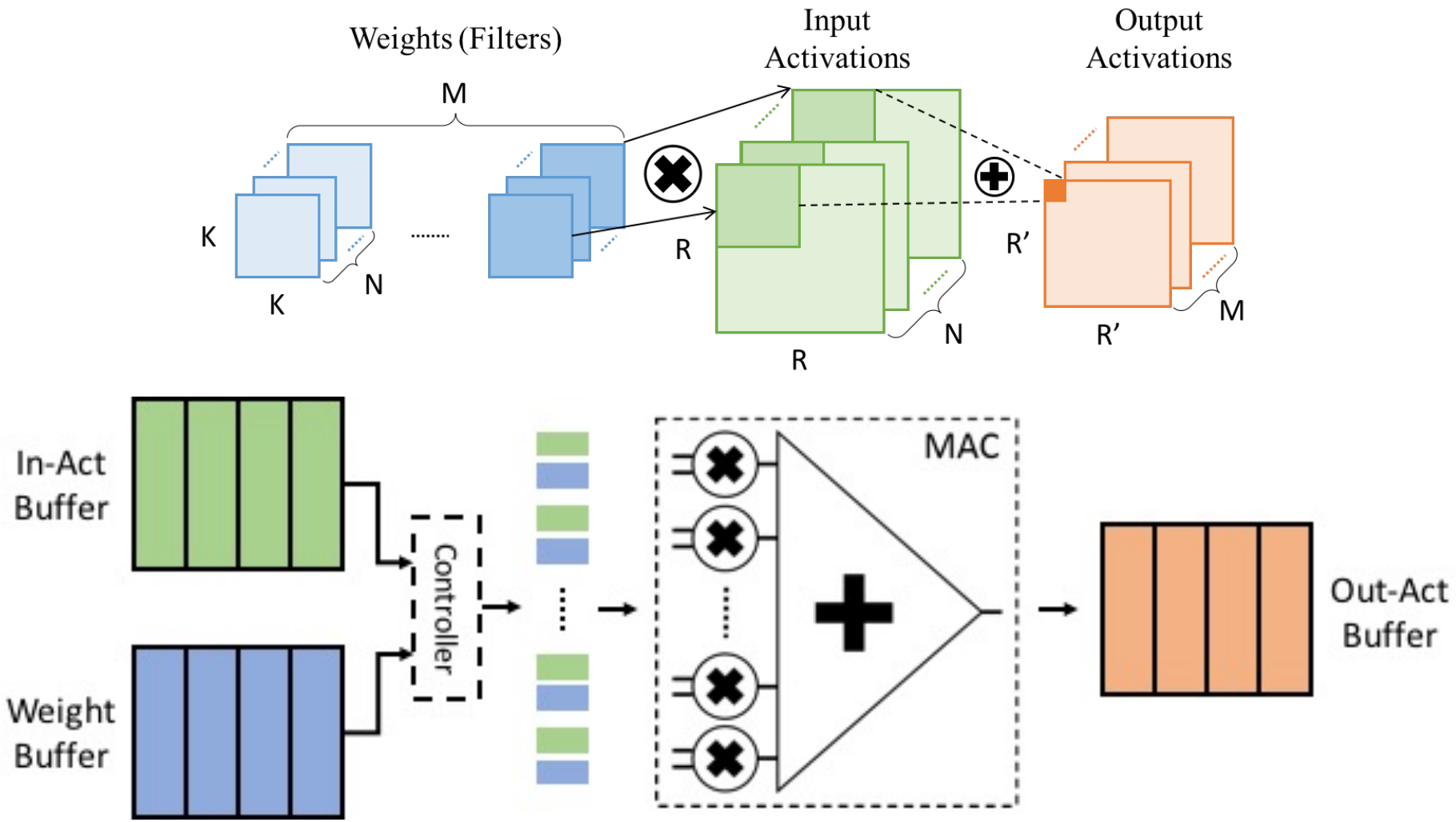
- MAC unit can efficiently perform CNN and thus, adopted by many CNN accelerators [Google TPU, DianNao, Zhang 2015, Cambricon-X]



SIMD-like Architecture

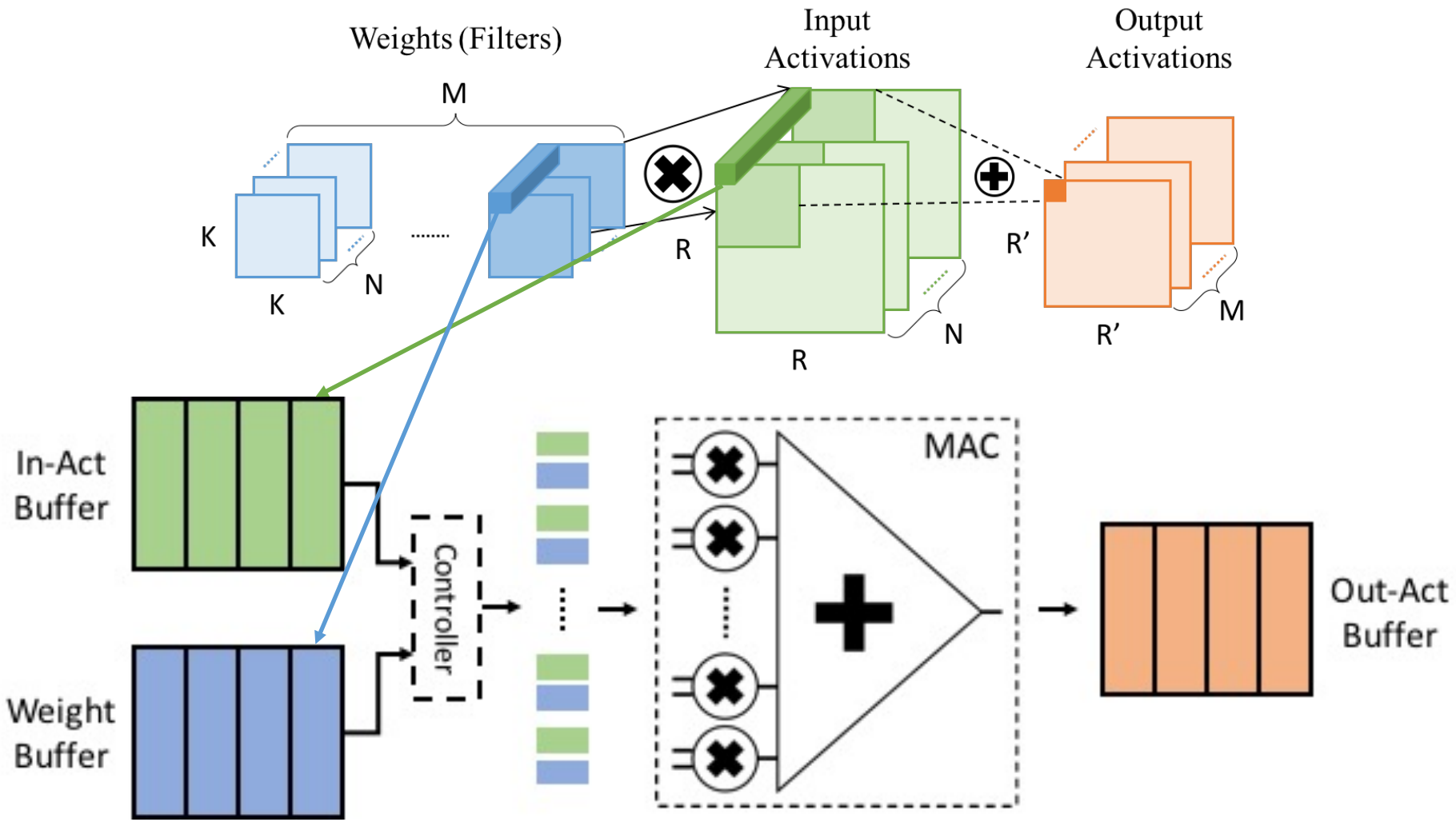
N. P. Jouppi et al., In-Datacenter Performance Analysis of a Tensor Processing Unit, ISCA 2017
T. Chen et al., DianNao: a small-footprint high-throughput accelerator for ubiquitous machine learning, ASPLOS 2014
C. Zhang et al., Optimizing fpga-based accelerator design for deep convolutional neural networks, FPGA 2015
S. Zhang et al., Cambricon-X: An accelerator for sparse neural networks, MICRO 2016

CNN Acceleration with SIMD



SIMD-like Architecture

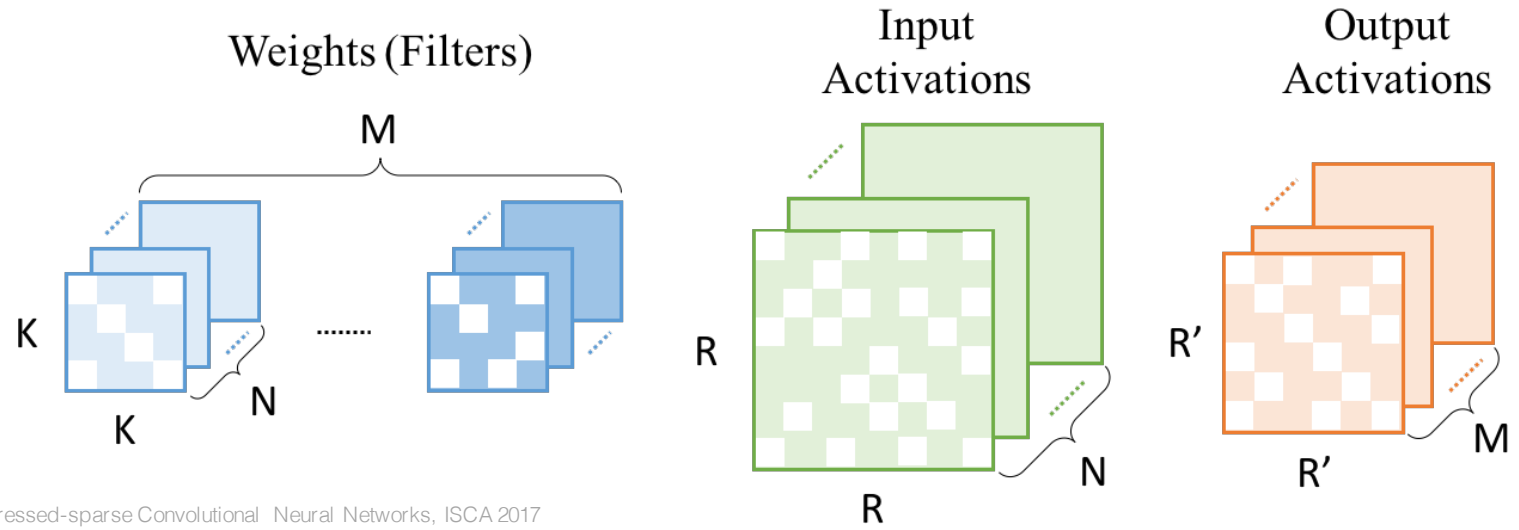
CNN Acceleration with SIMD



SIMD-like Architecture

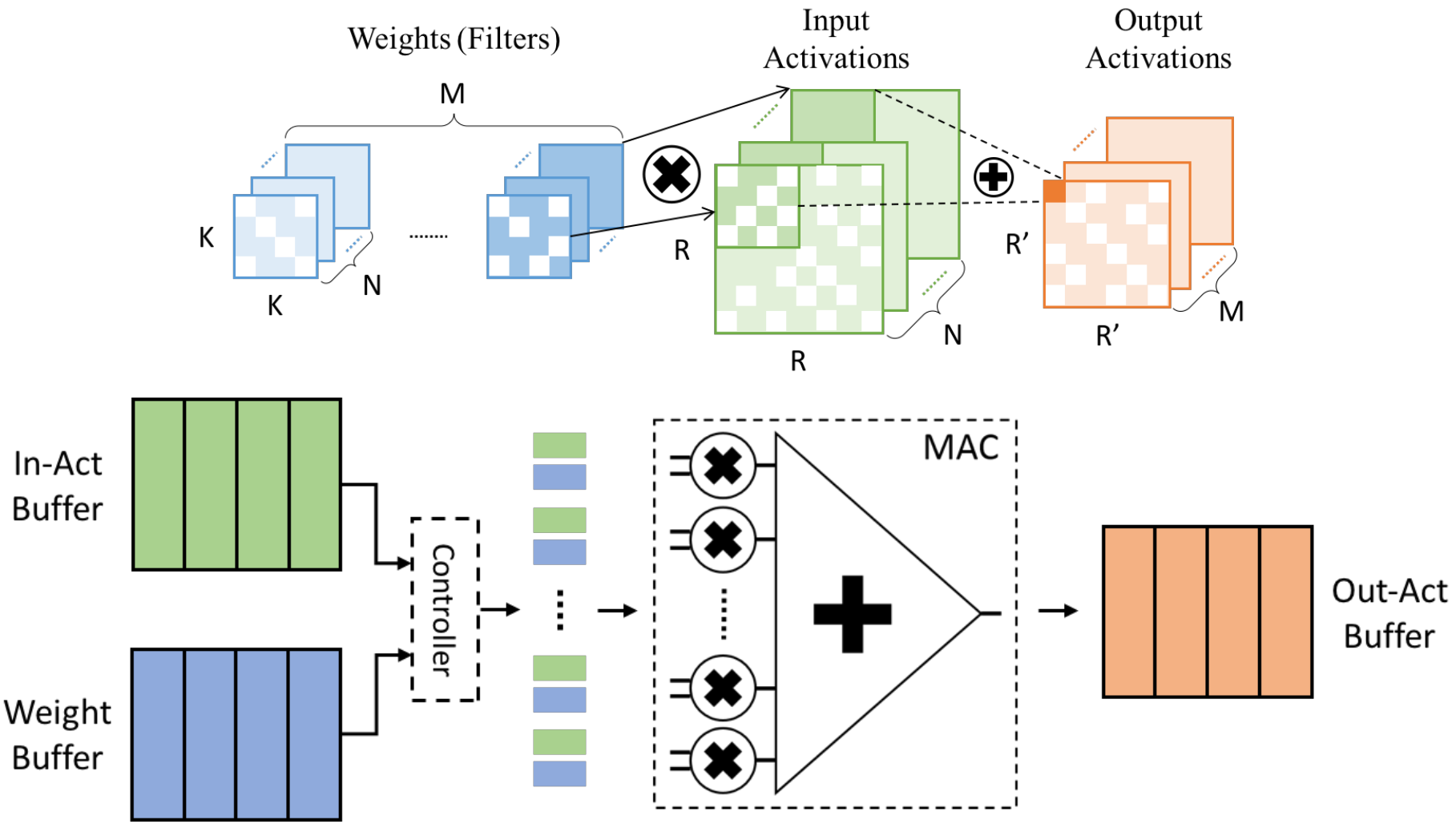
CNN is Sparse

- About **60%** of weights and activations are **ZEROs**
  - Zeros in activations are dynamically generated after **ReLU**
  - Zeros in weights are obtained with **Network Pruning**
- Sparsity is promising for **speedup** (Zero-skipping) and **energy reduction** (smaller memory footprint)



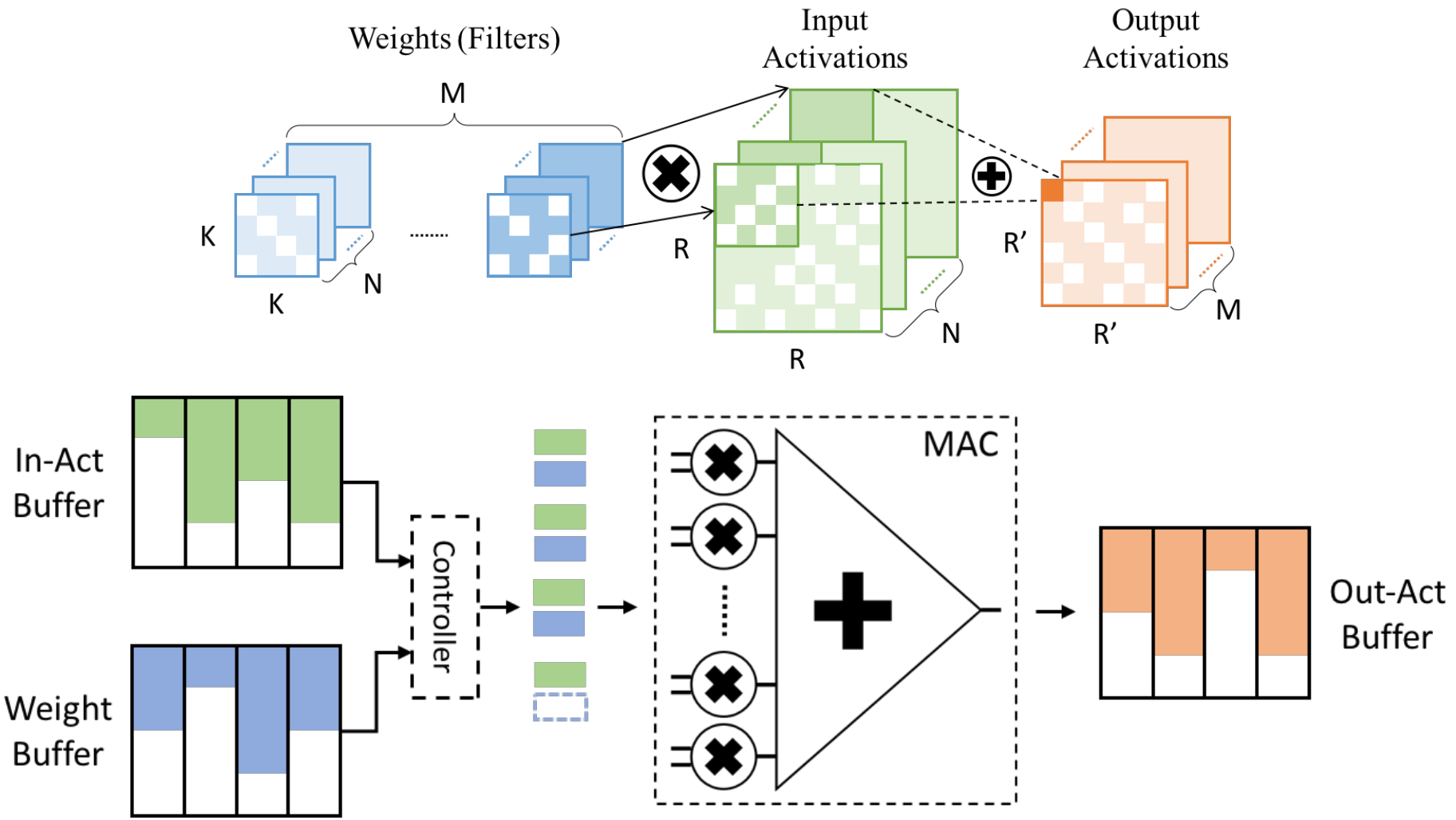
A. Parashar et al., SCNN: An Accelerator for Compressed-sparse Convolutional Neural Networks, ISCA 2017
S. Han et al., Learning both weights and connections for efficient neural network, NIPS 2015

Sparse CNN on SIMD?



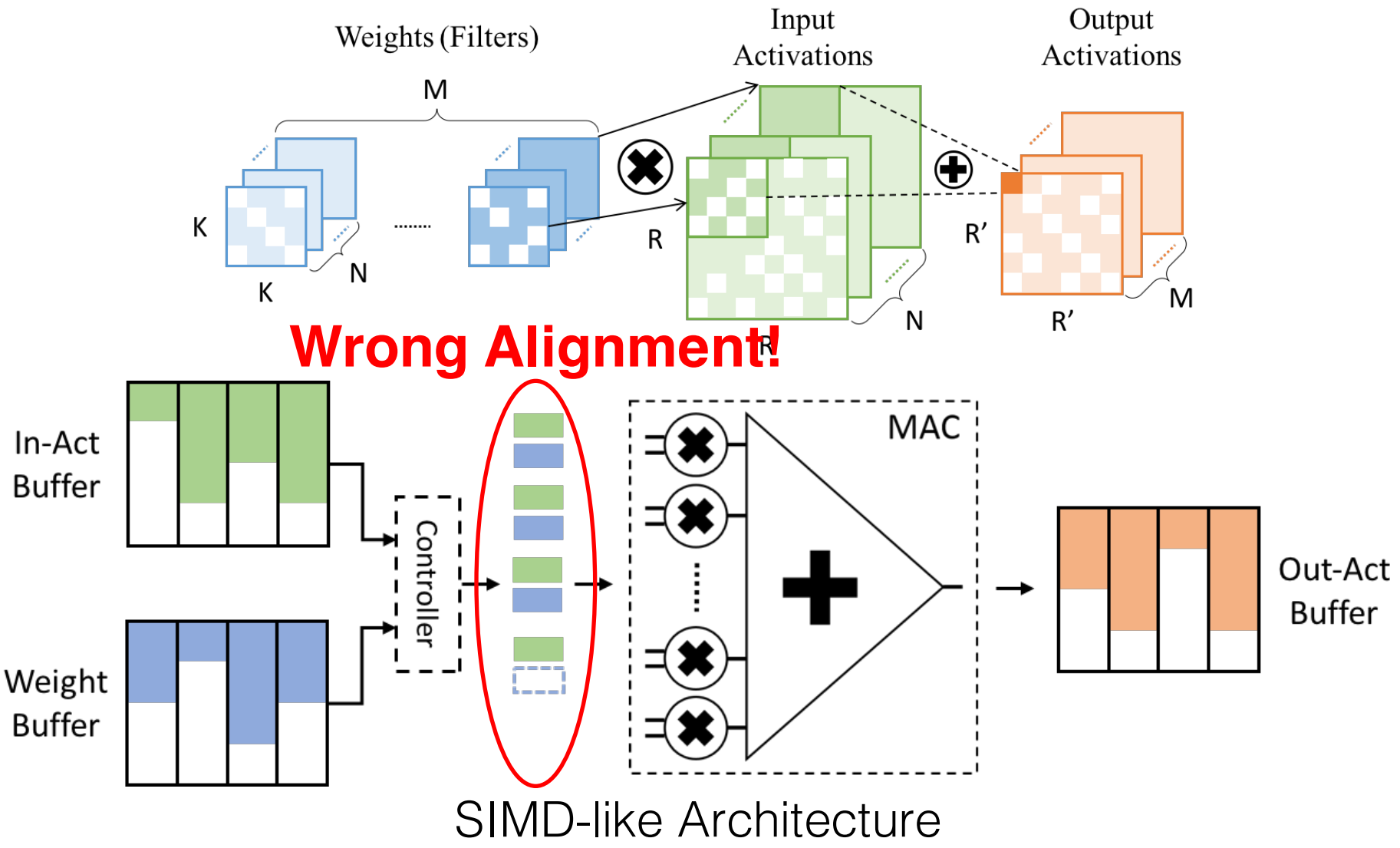
SIMD-like Architecture

Sparse CNN on SIMD?

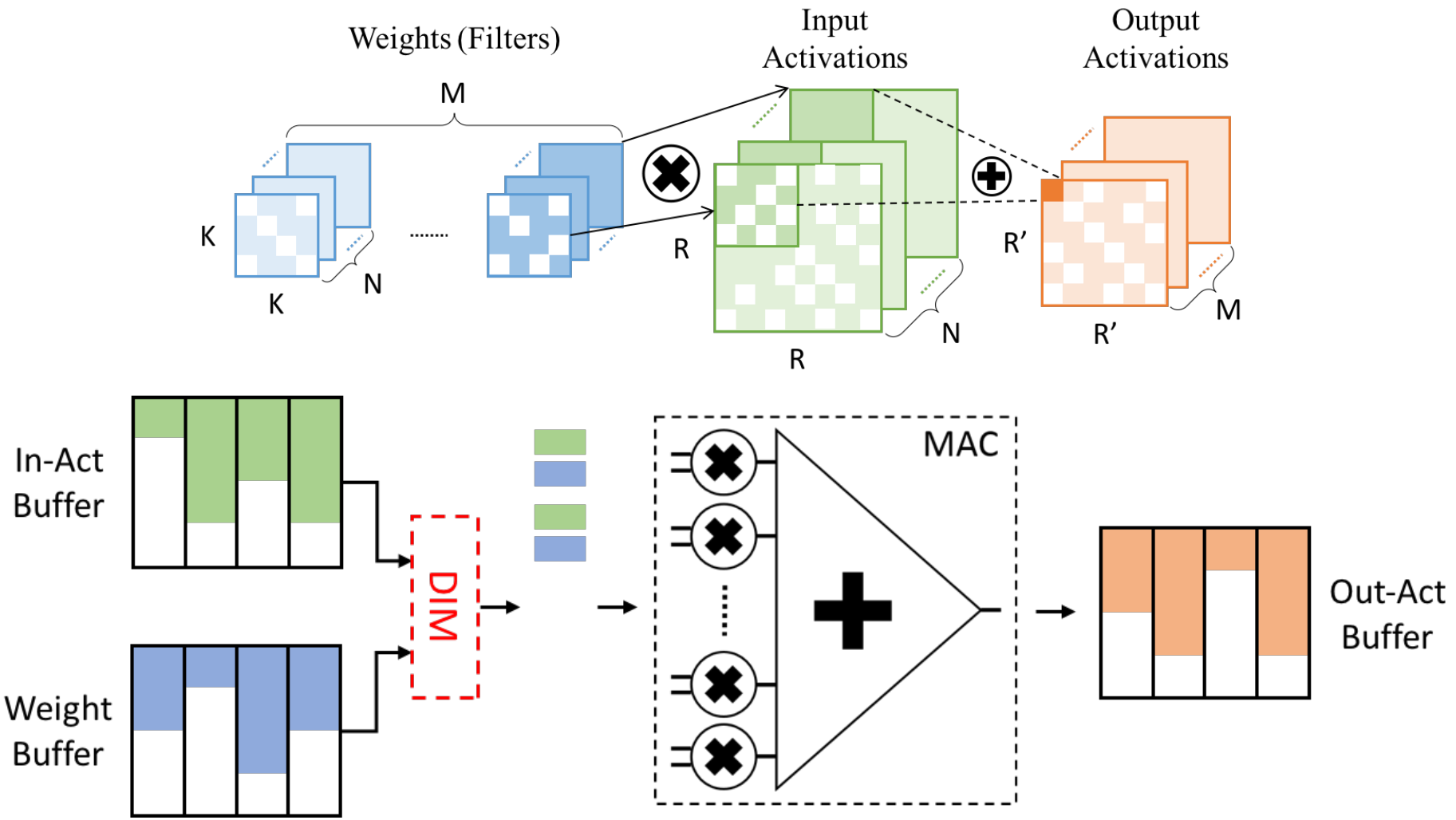


SIMD-like Architecture

Sparse CNN on SIMD?

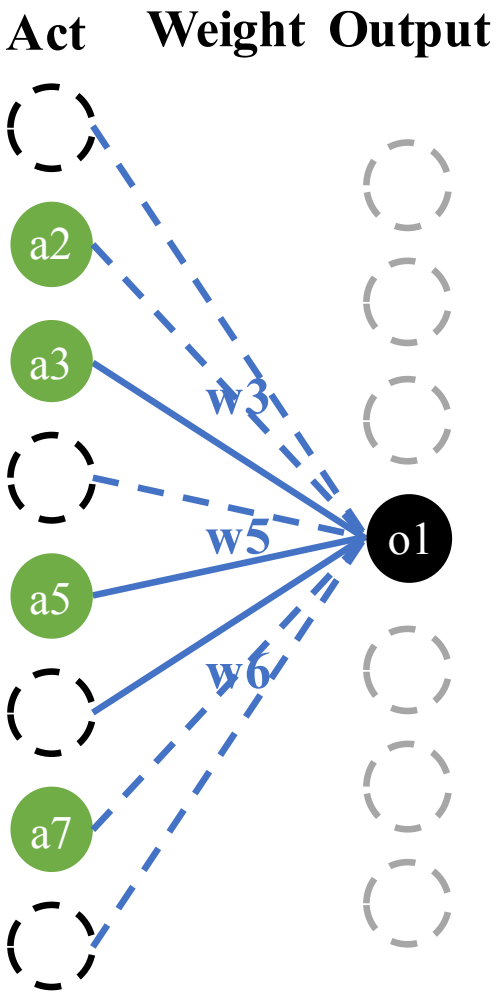


Sparse CNN on SIMD!

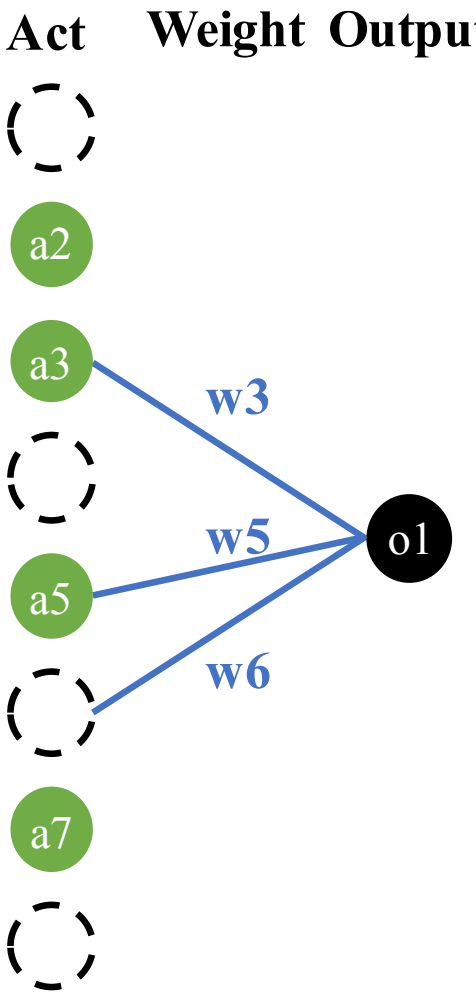


SIMD-like Architecture

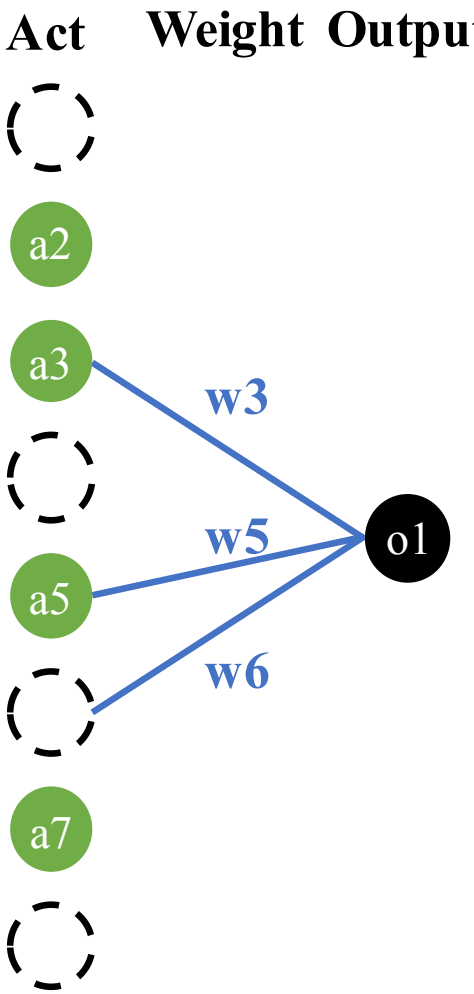
A Simple Sparse Layer



A Simple Sparse Layer



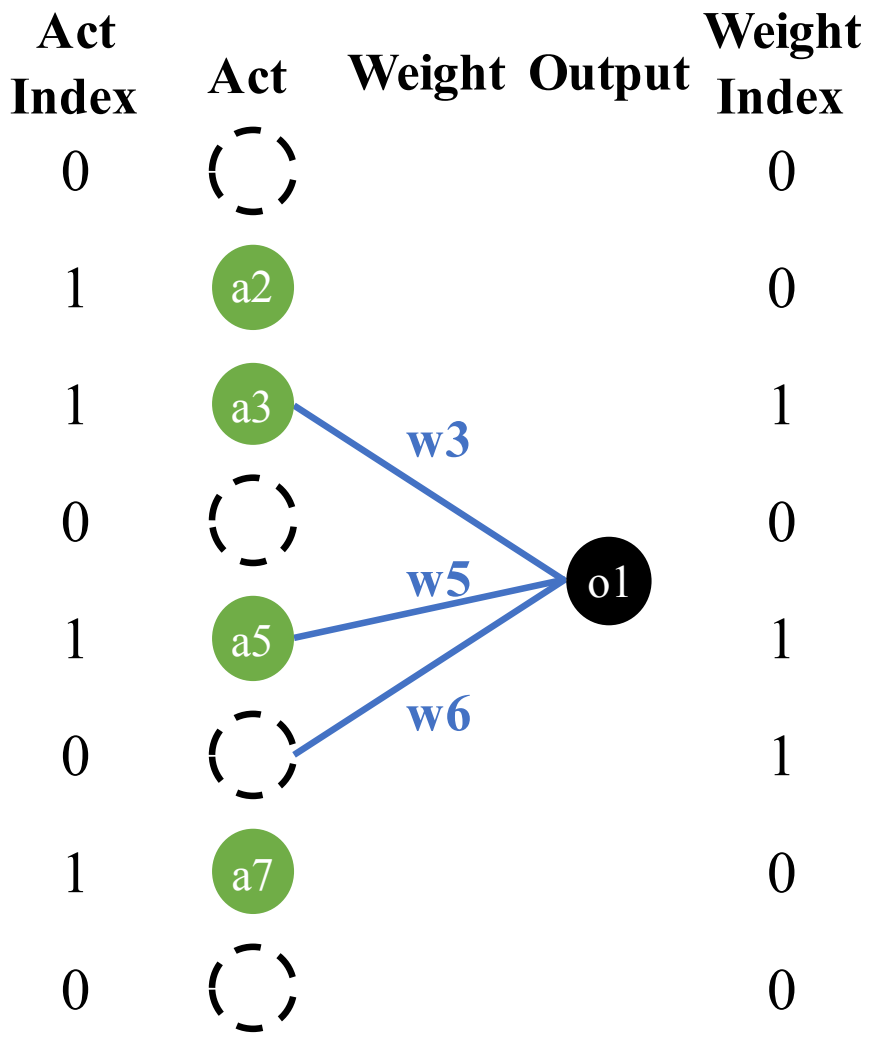
Compressed-Sparse Data: Only Keep Non-Zeros



Stored Activations

Stored Weights

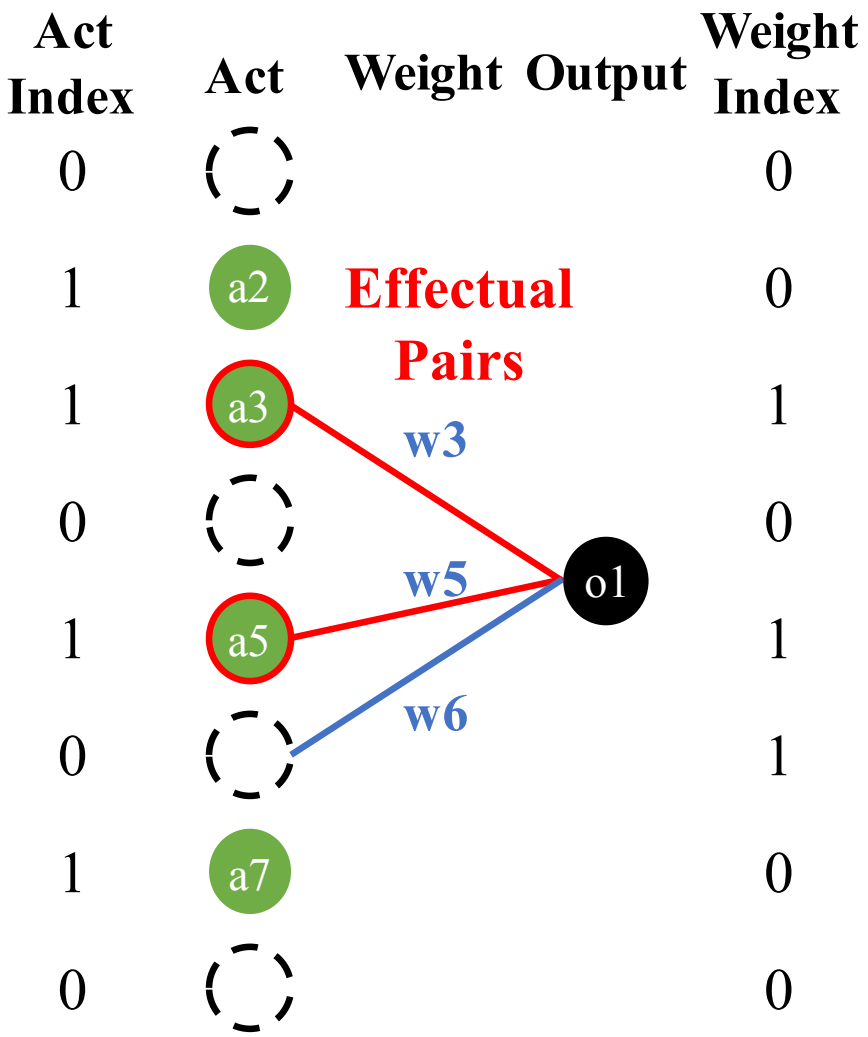
Plus Index: Bit-Vector Recording Zero/Non-Zero



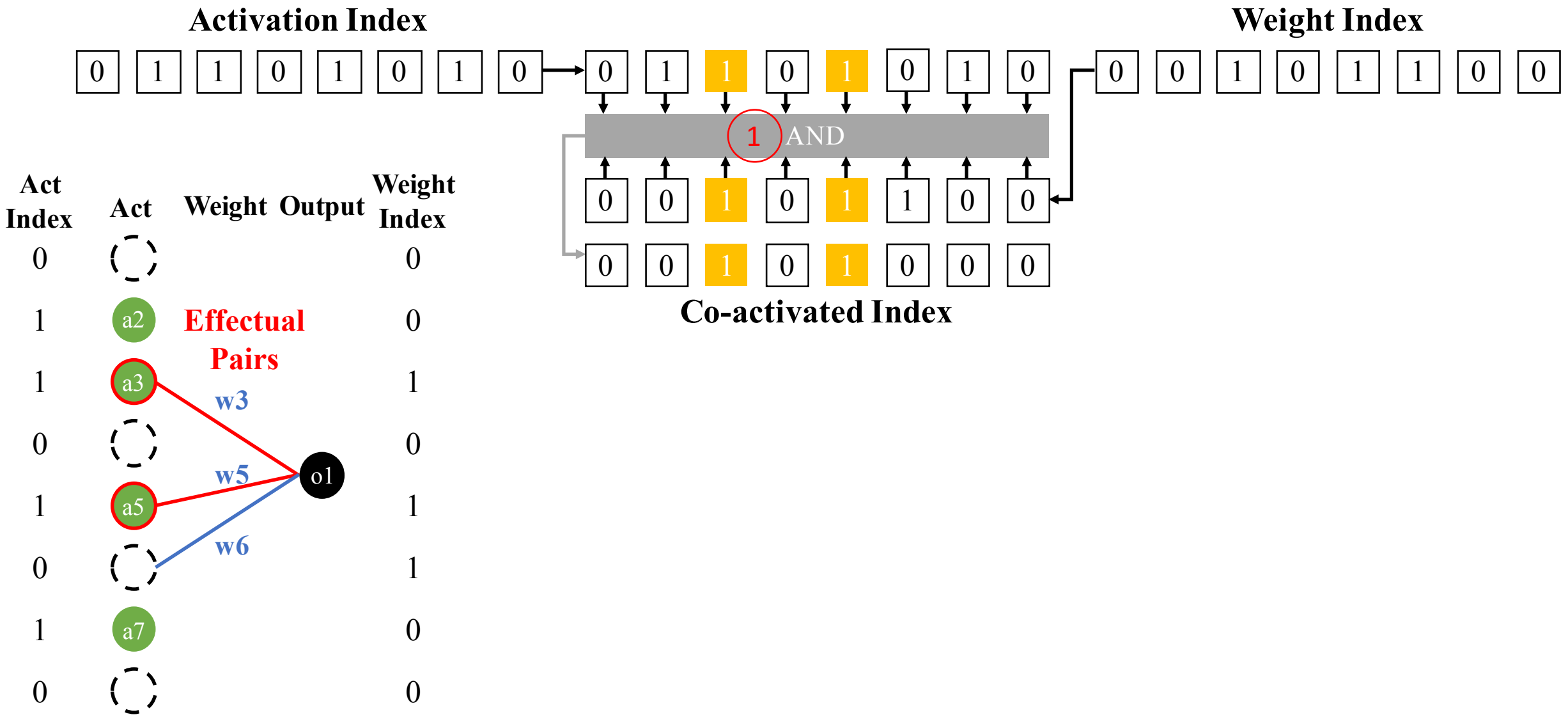
Stored Activations: a2 a3 a5 a7

Stored Weights: w3 w5 w6

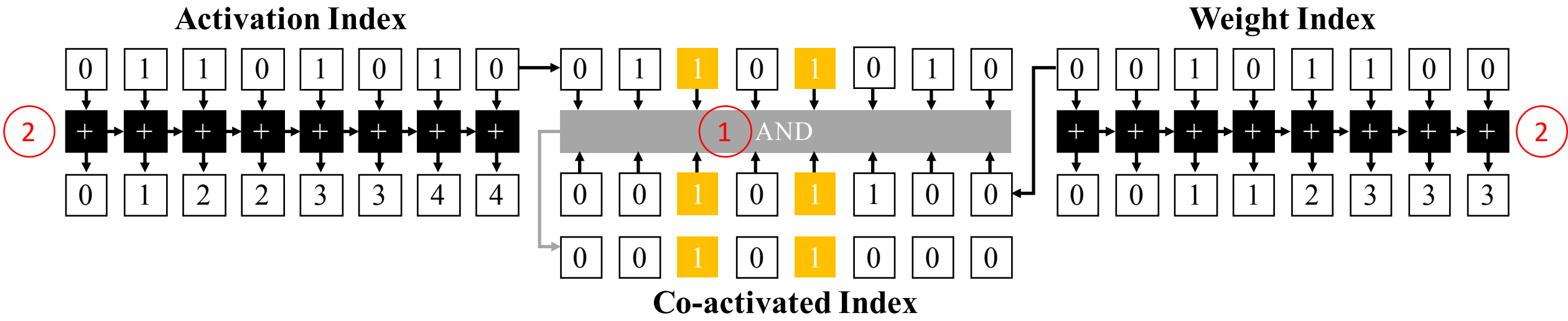
Target of DIM: Find Out Effectual Pairs



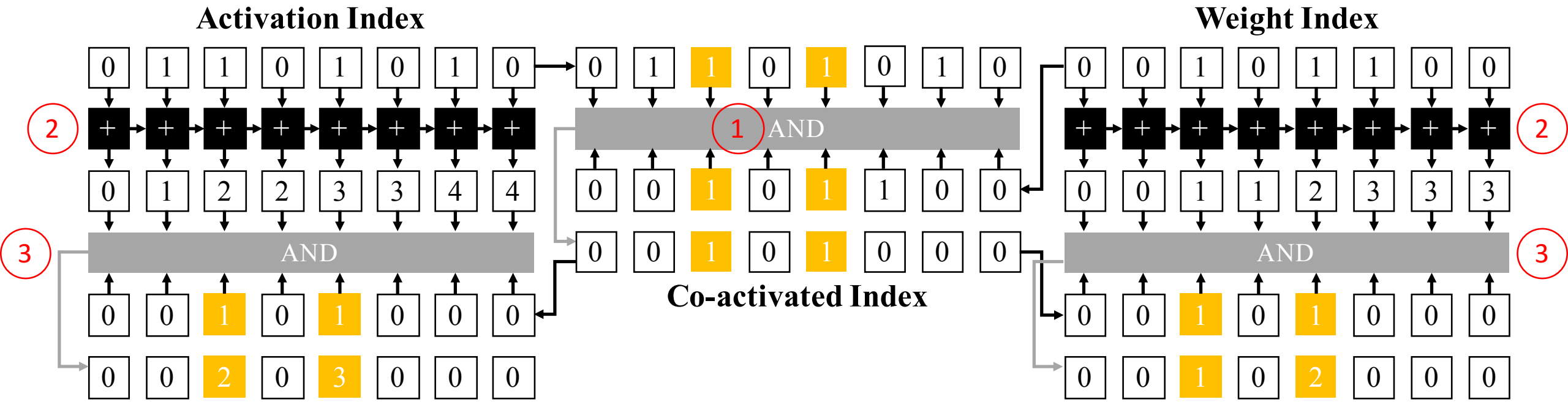
Dual Indexing Module: Step1



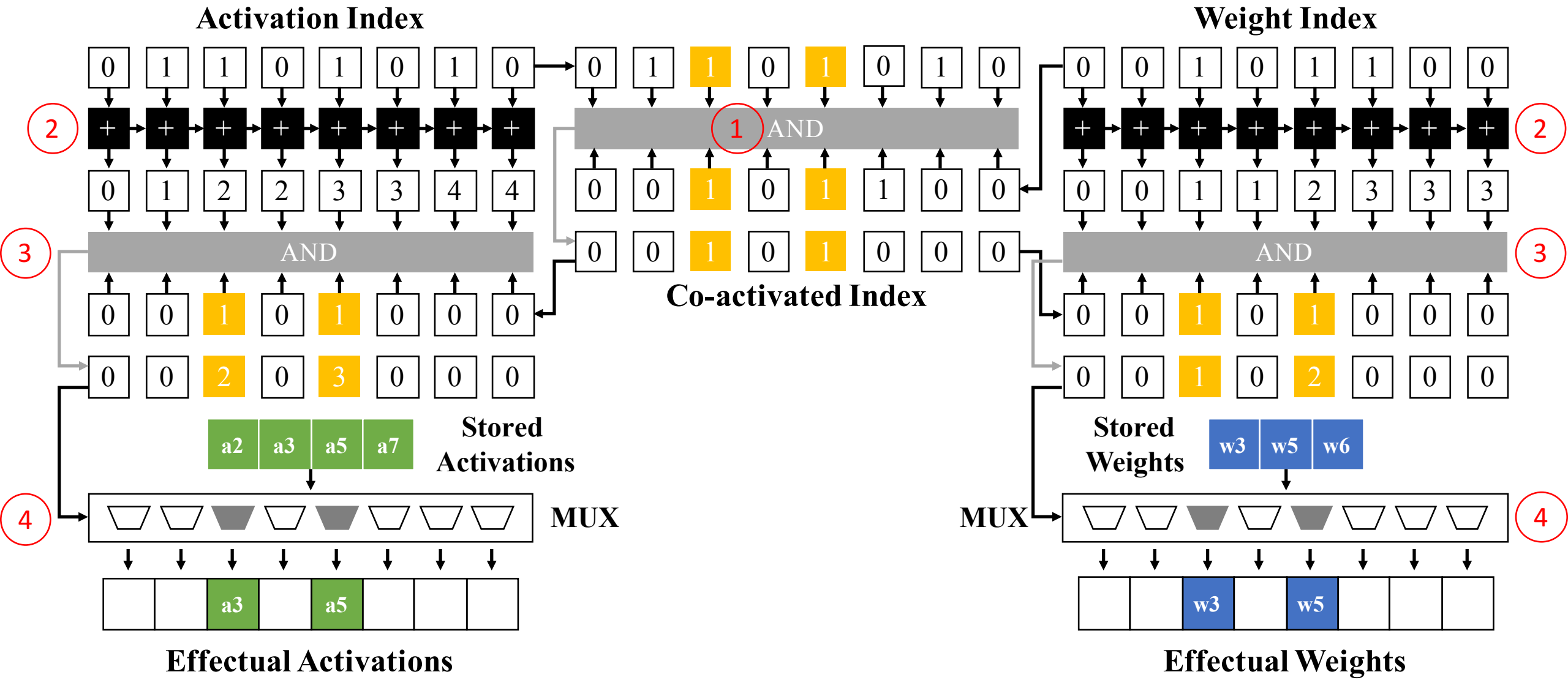
Dual Indexing Module: Step2



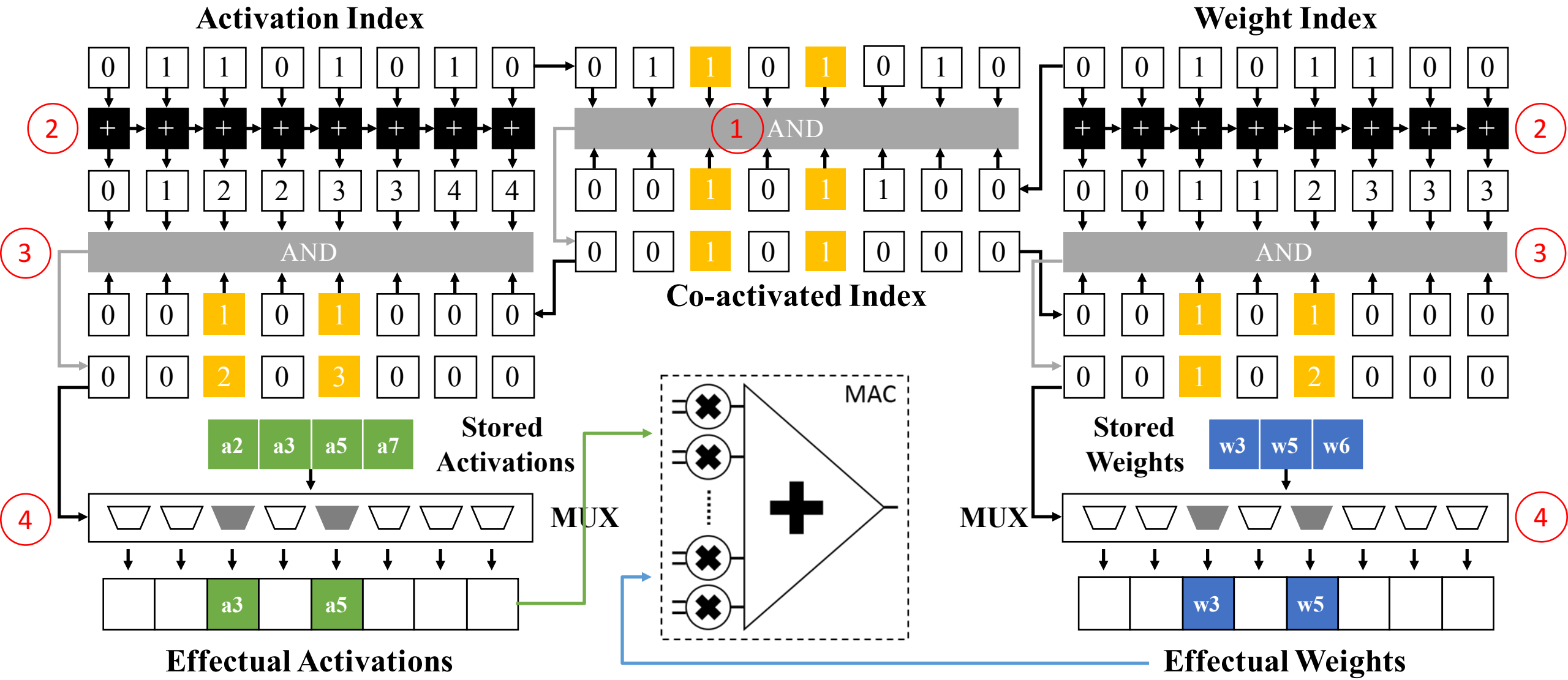
Dual Indexing Module: Step3



Dual Indexing Module: Step4

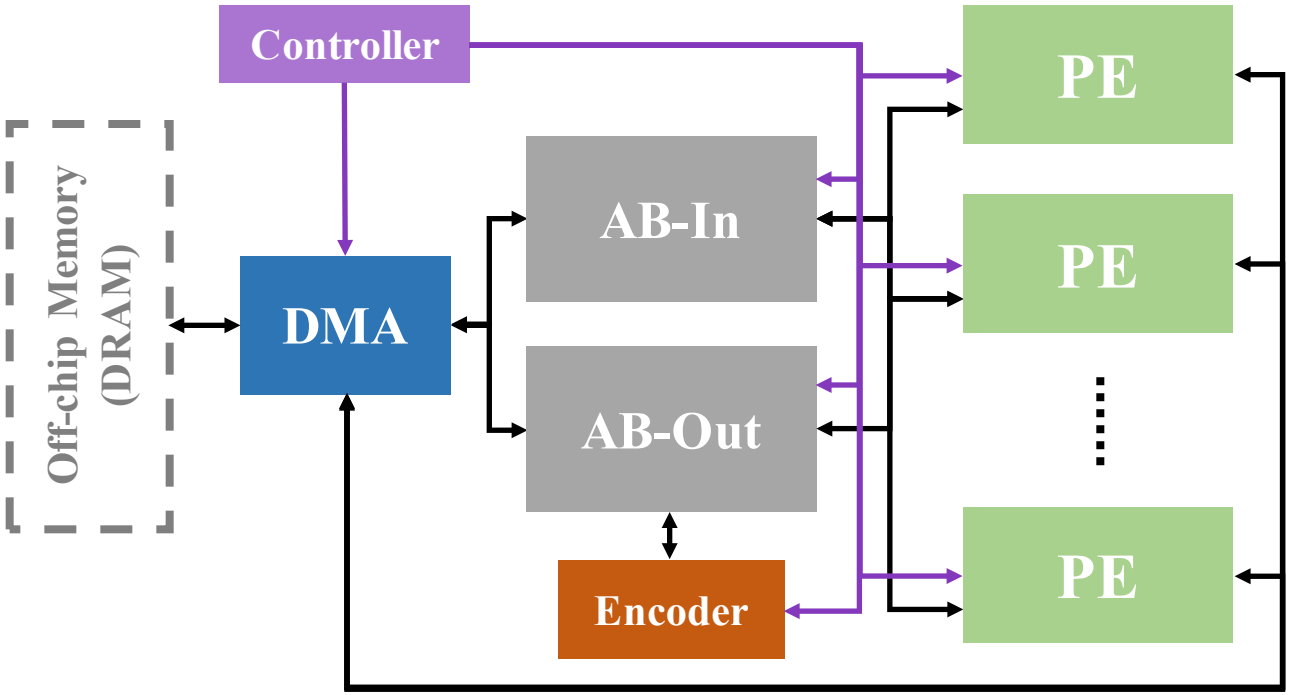


Alignment Issue Solved!



Accelerator Design

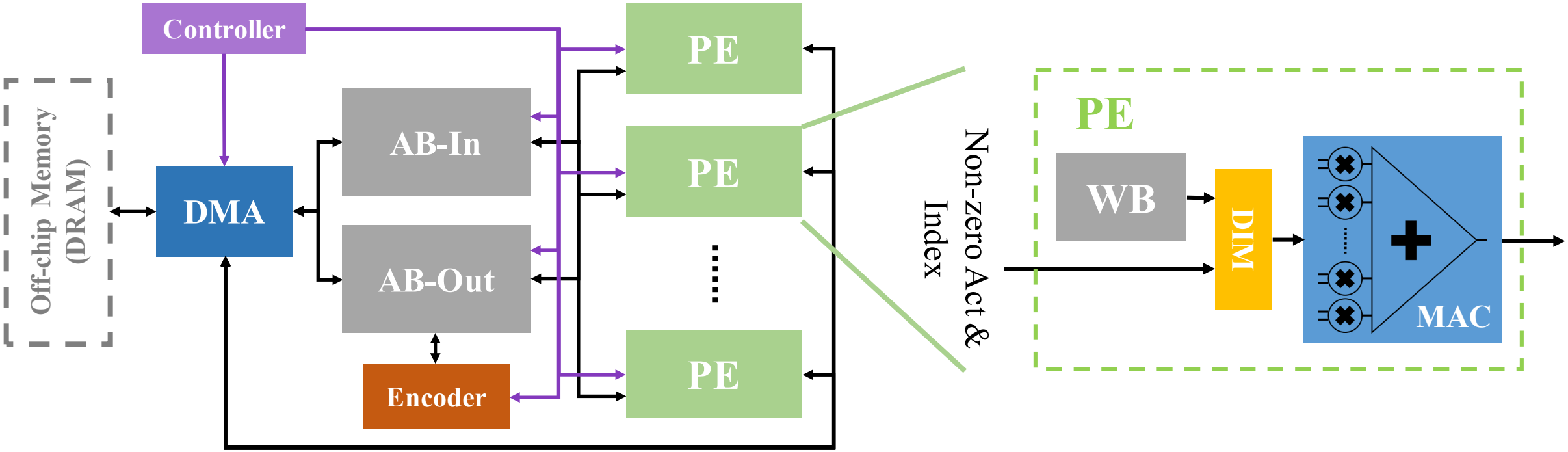
- Extended from Cambricon-X [MICRO 2016]



S. Zhang et al., Cambricon-X: An Accelerator for Sparse Neural Networks, MICRO 2016

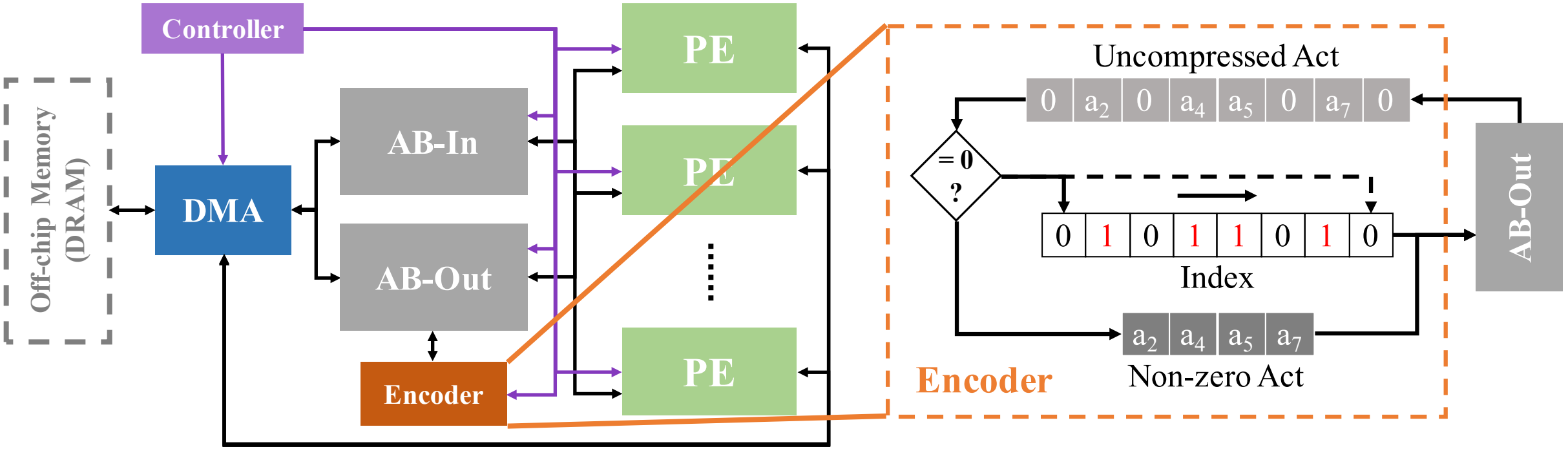
Accelerator Design

- Plug DIM into each PE



Accelerator Design

- Encode output activations on-the-fly



Evaluation Methodology

- Logic: Synthesis with TSMC 40nm
- SRAM and DRAM: CACTI
- Benchmark: Open Sparse-AlexNet + ImageNet Data
- Experiments: In-house performance simulator

M. Naveen et al., CACTI 6.0: A Tool to Model Large Caches, HP Laboratories, 2009
Sparse AlexNet, https://github.com/songhan/Deep-Compression-AlexNet
J. Deng et al., Imagenet: A large-scale hierarchical image database, CVPR, 2009

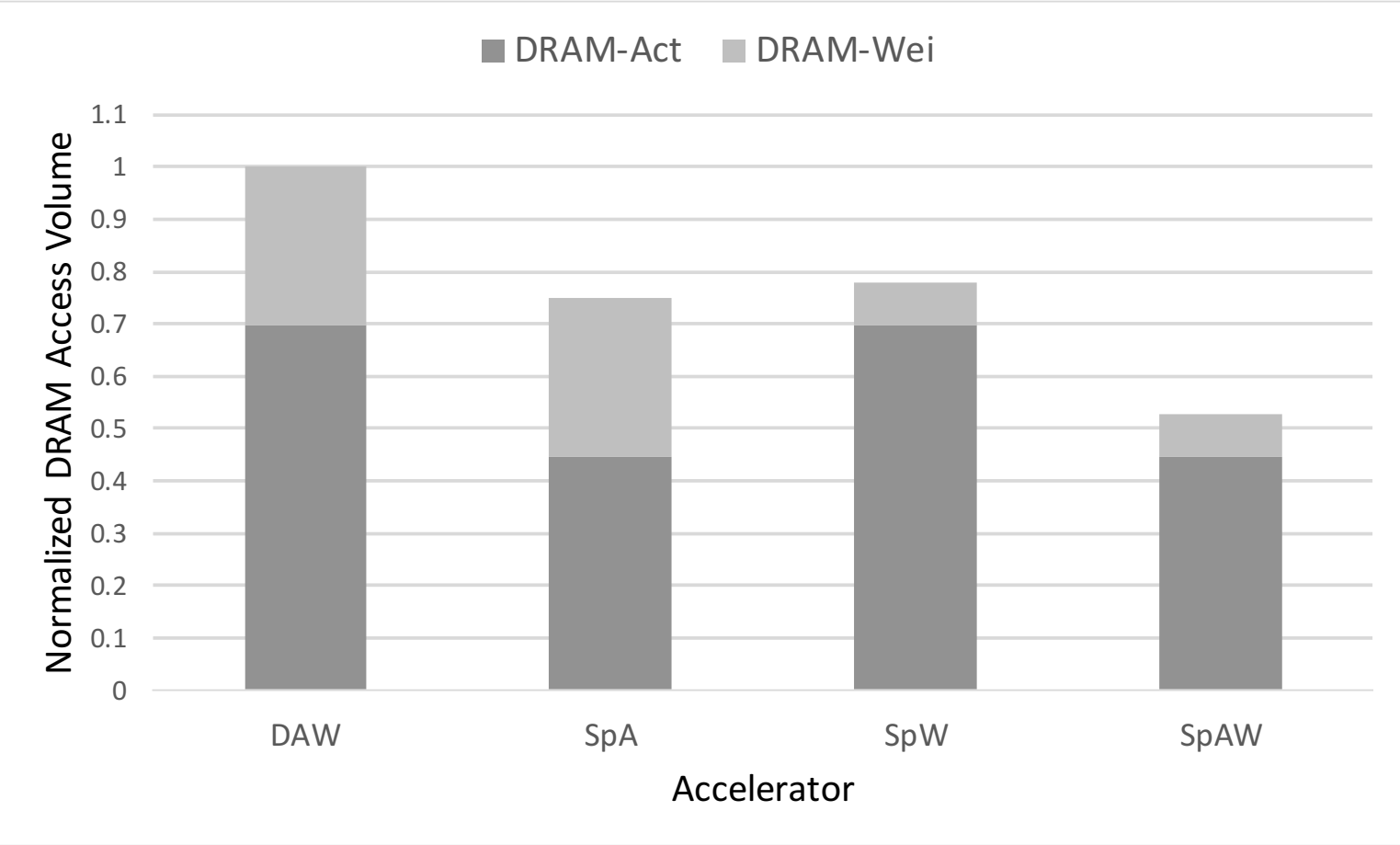
Accelerator Variants

- Overheads: 14.4% in Area and 19.5% in Power

Acc	Act	Weights	Index	Encoder	Area(mm^2)	Power(mW)
DAW	Dense	Dense	N/A	N/A	2.05	395
SpA	Sparse	Dense	IM	✔	2.15	428
SpW	Dense	Sparse	IM	N/A	2.23	441
SpAW	**Sparse**	**Sparse**	**DIM**	✔	**2.34**	**472**

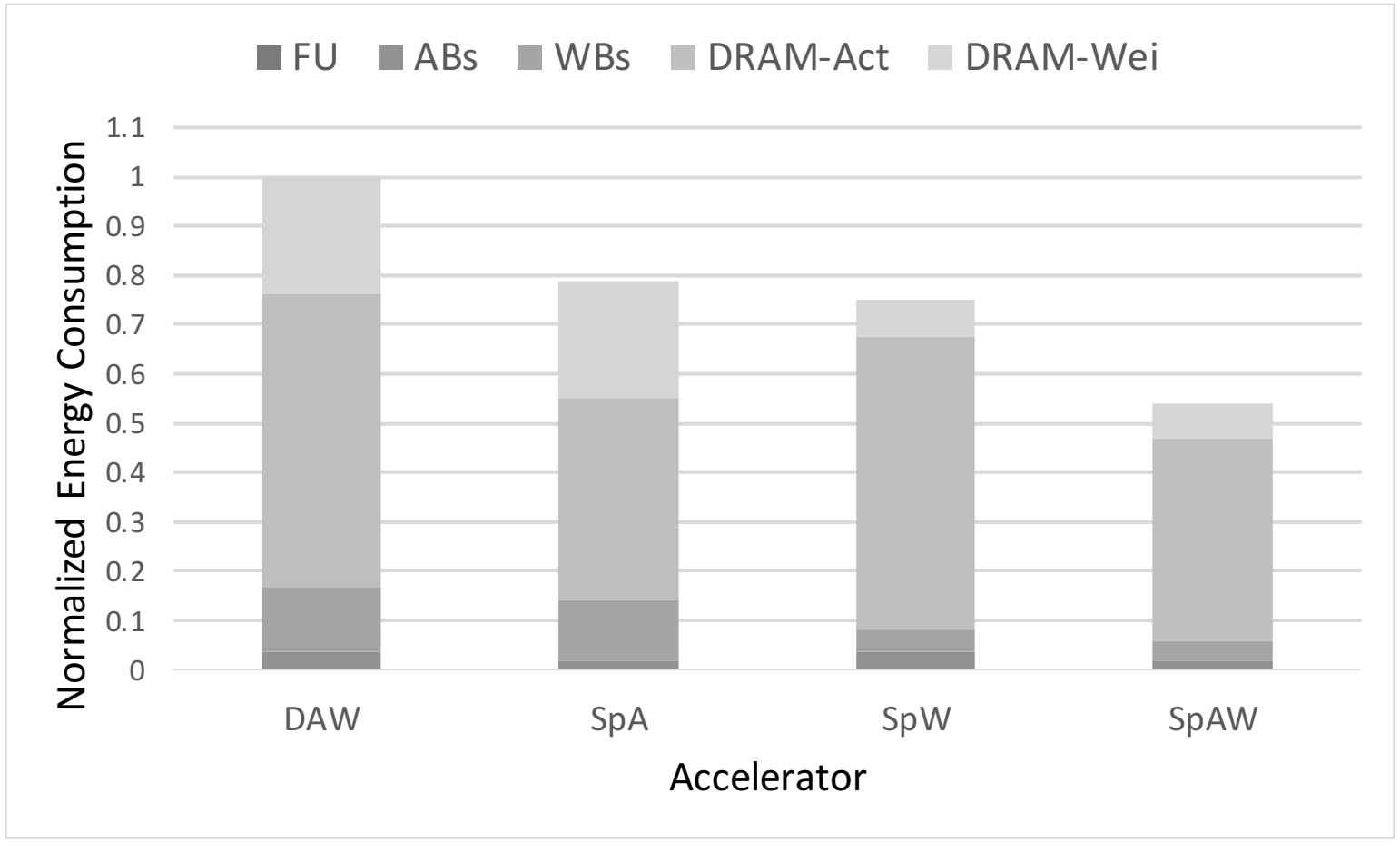
DRAM Access Volume

- 47.3% less in DRAM access volume compared to DAW



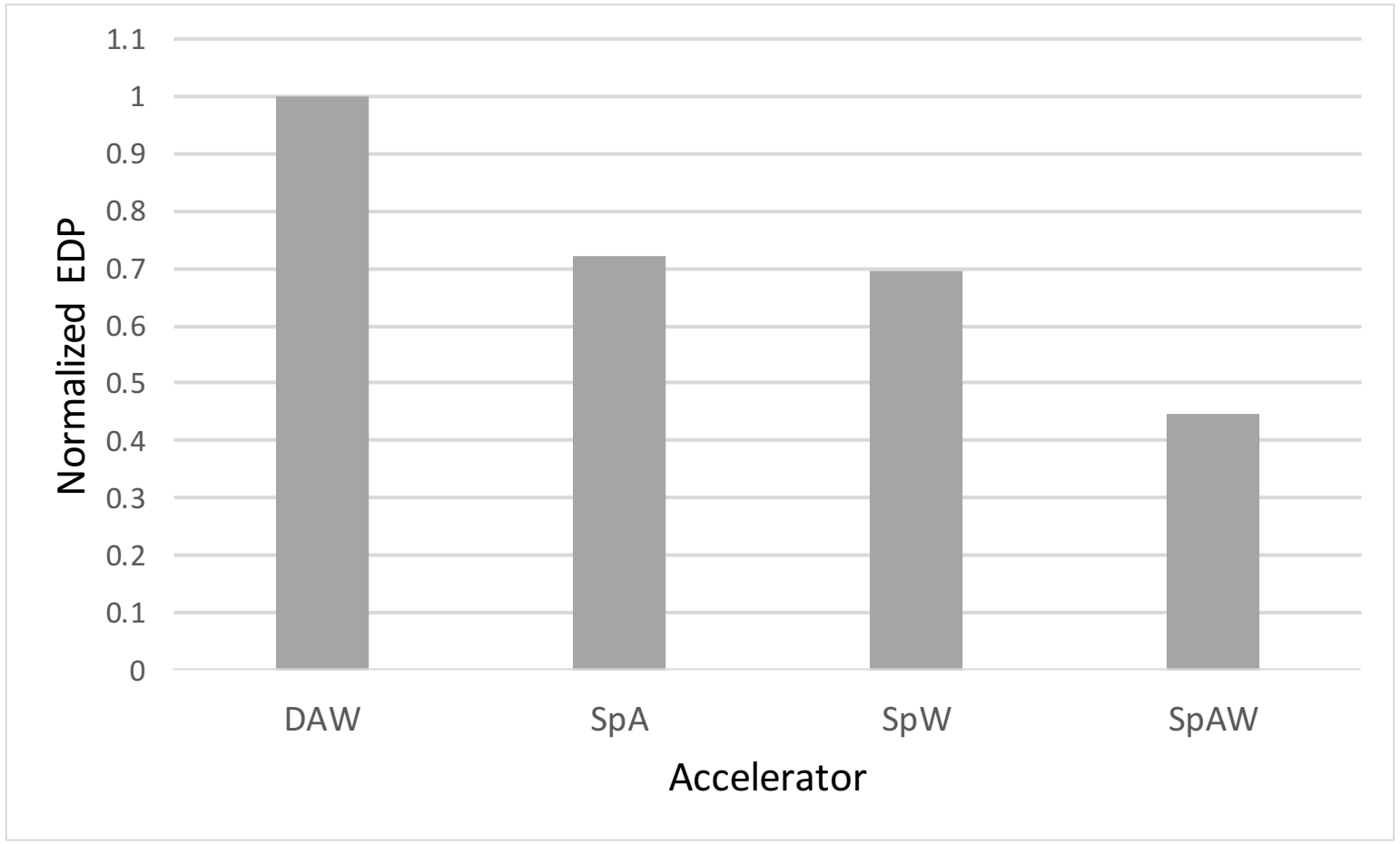
Energy Consumption

- 46% energy reduction compared to DAW



Energy-Delay-Product

- 55.4% EDP reduction compared to DAW



Summary

- SIMD-like accelerator has **alignment issue** while performing sparse CNN
- We propose a novel ***Dual Indexing Module*** (DIM) to handle the alignment issue efficiently
- By keeping data in a **compressed-sparse format**, a CNN accelerator with DIM can reduce DRAM access volume, energy consumption and EDP for 47.3%, 46% and 55.4%

Thank You!

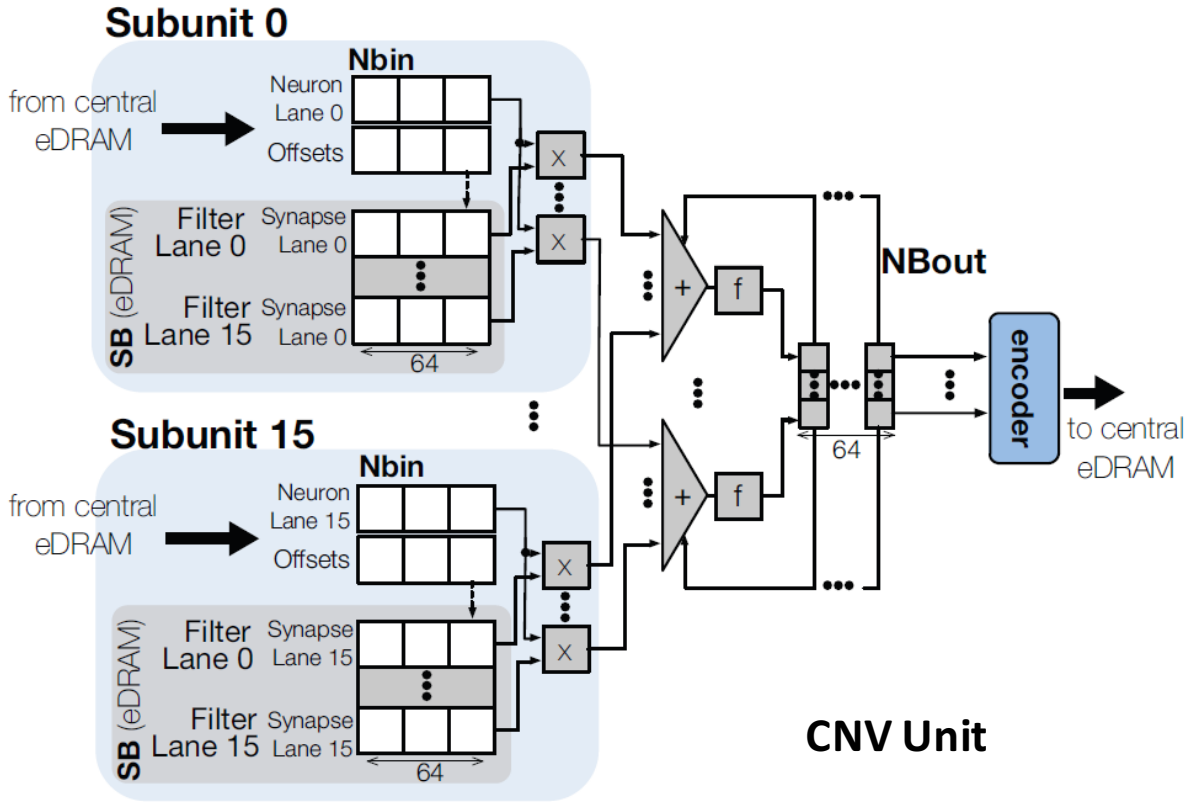
Additional Materials

Design Parameters of SpAW

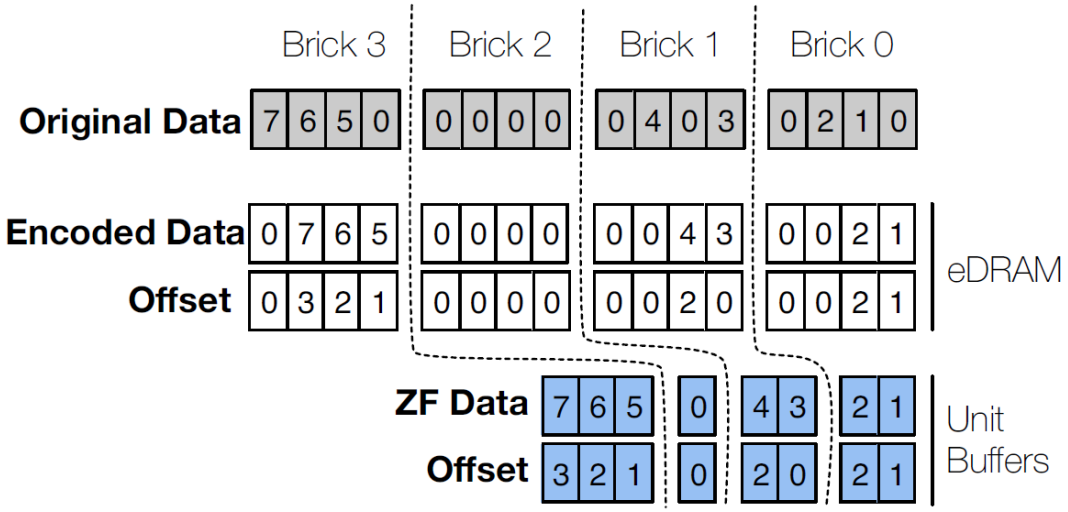
Accelerator Parameters	Value
Clock Rate	1 GHz
Number of PEs	16
WBs (Total)	32 KB
WBs-Idx (Total)	2 KB
AB-In/AB-Out (each)	8 KB
AB-In-Idx/AB-Out-Idx (each)	500 B
PE Parameters	**Value**
Multiplier Precision	16-bit
MAC Width	16 * 16-bit
WB	2KB
WB-Idx	128 B

Related Work - Cnvlutin

- Decouple neuron lanes to do zero-skipping in neurons



CNV Unit

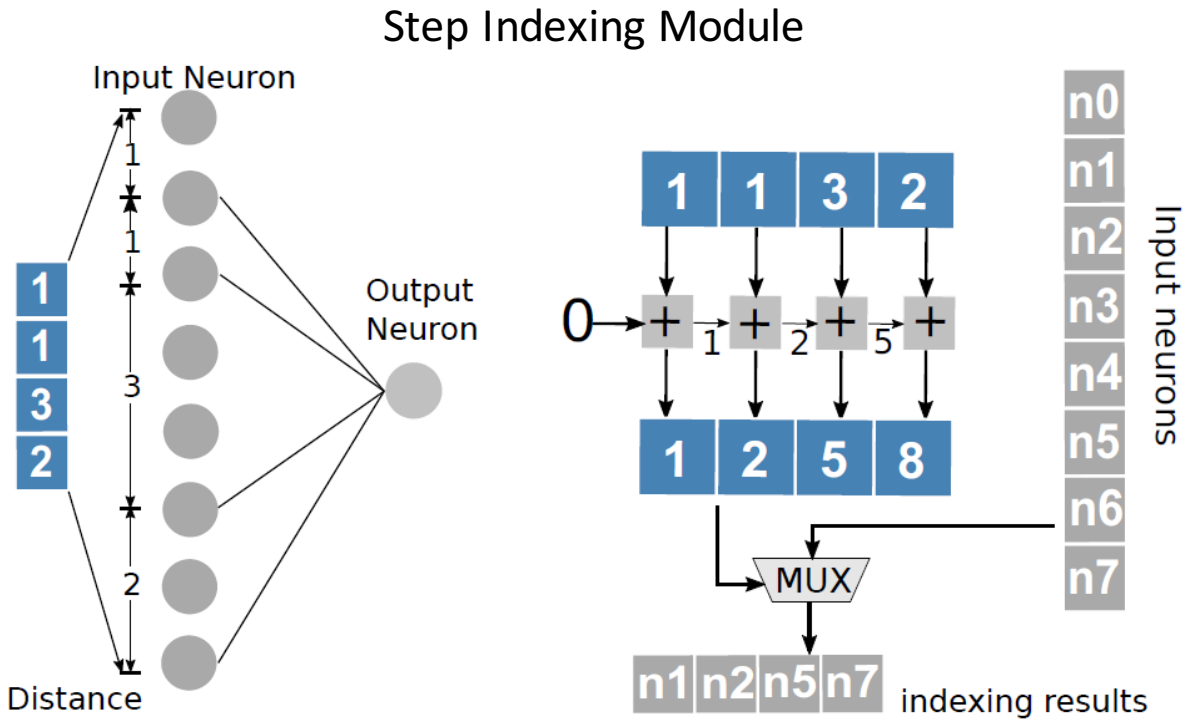


Zero-Free Neuron Array Format (ZFNAF)

J. Albericio et al., CNVLUTIN: Ineffectual-neuron-free DNN computing, ISCA 2016

Related Work - Cambricon-X

- Utilizing weight sparsity with step indexing (a compressed-sparse format)



S. Zhang et al., Cambricon-X: An Accelerator for Sparse Neural Networks, MICRO 2016