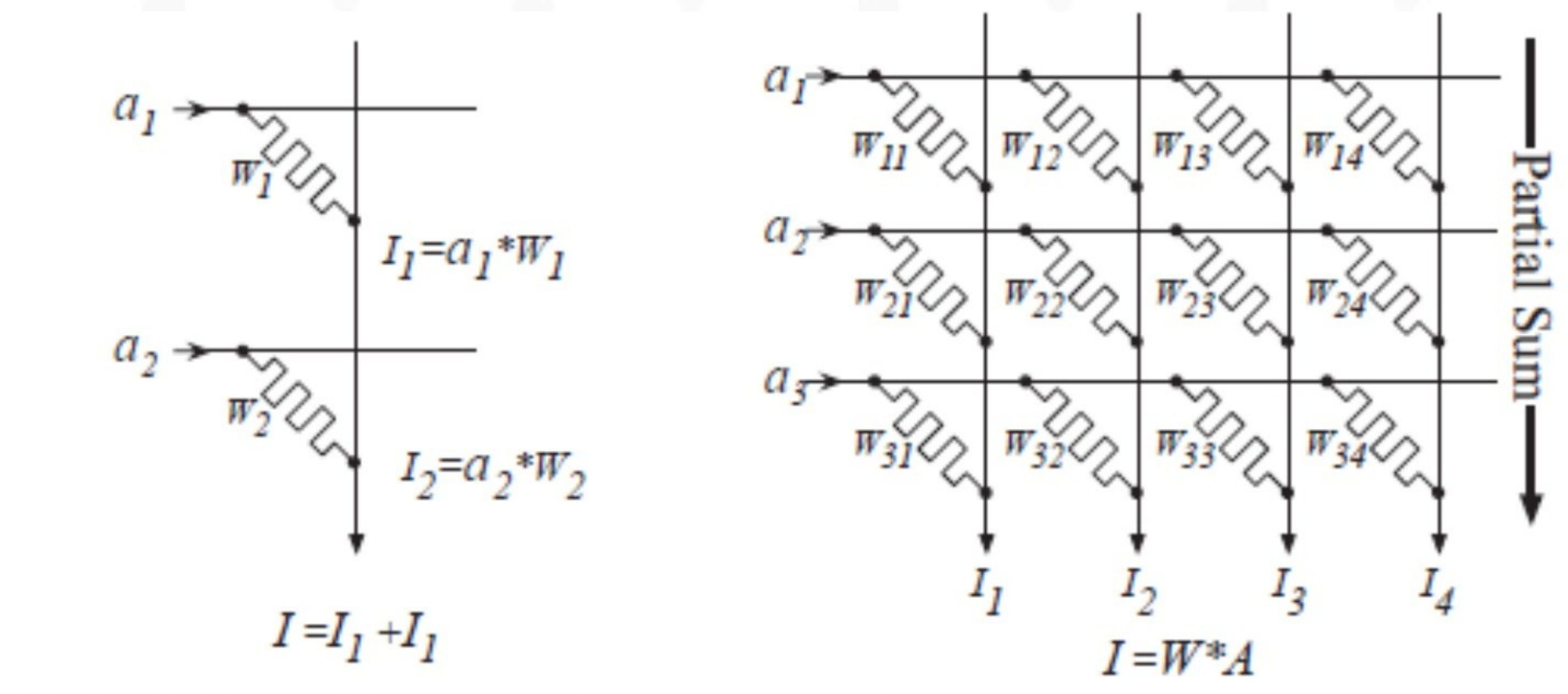
首届致远学术节
学生科研成果展示

ReCom: An Efficient Resistive Accelerator for Compressed Deep Neural Networks

Houxiang Ji[1], Lianghao Song[2], Li Jiang[1], Hai(Halen) Li[2], Yiran Chen[2]
1. Department of CSE, Shanghai Jiao Tong University
2. Department of ECE, Duke University

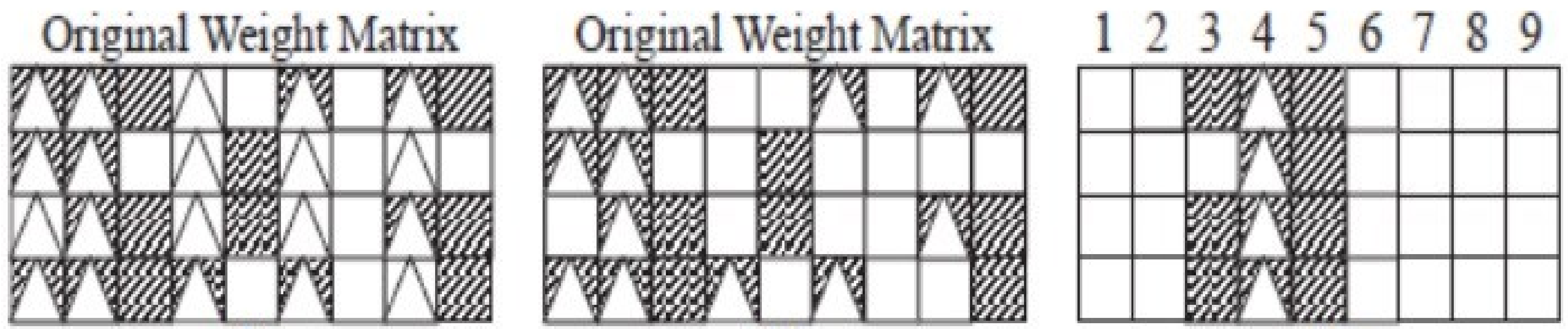
I. Introduction & Motivation

1.ReRAM-based Accelerator for Deep Learning Algorithms



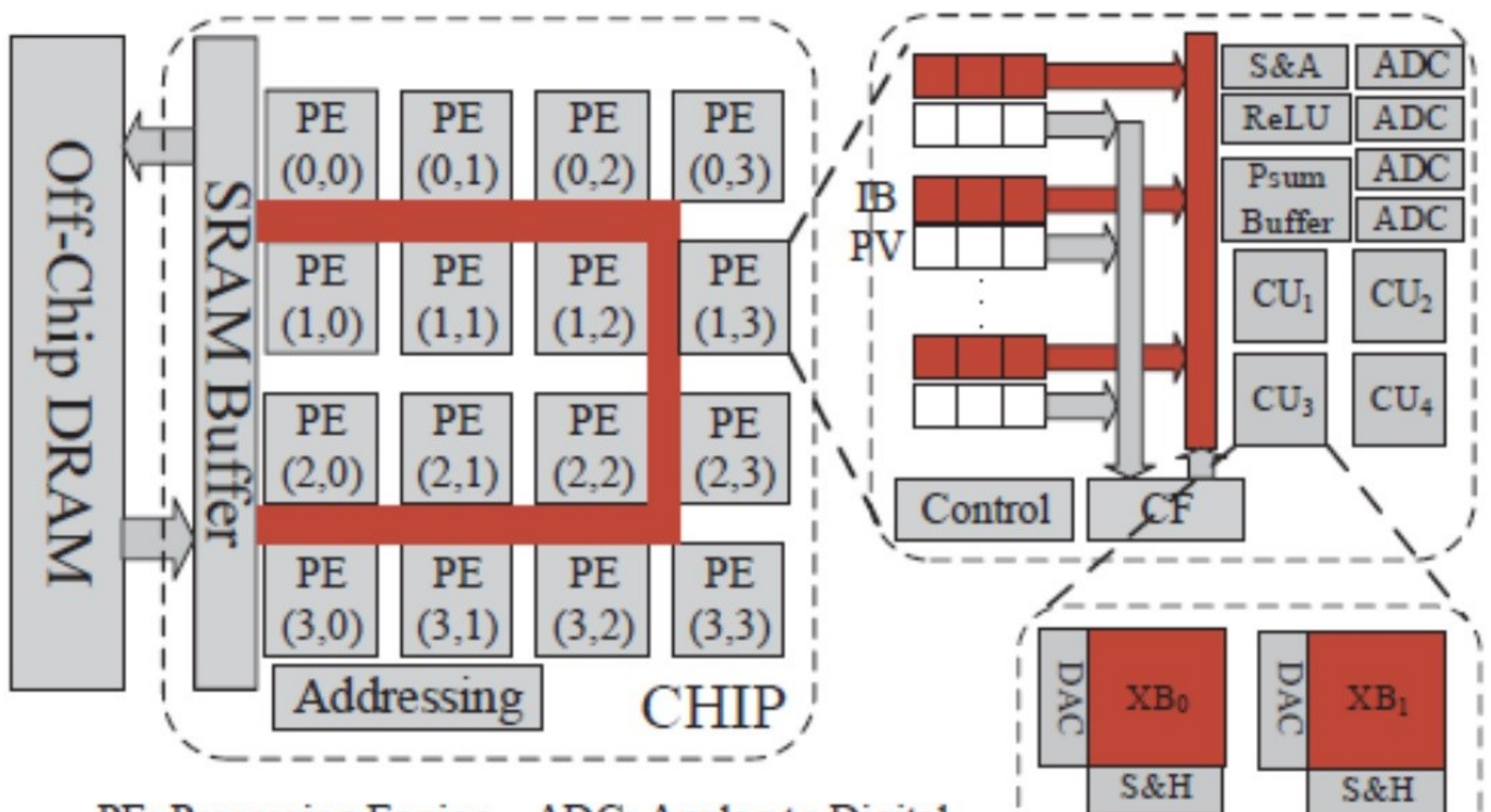
(a) Multiply and Accumulation
(b) Matrix-Vector Multiplication

2. Sparse Neural Network Acceleration



(b) (c) (d)

II. ReCom Architecture



PE: Processing Engine
S&A: Shift and Add
IB: Input Buffer
PV: Position Vector
S&H: Sample and Hold
ADC: Analog to Digital
DAC: Digital to Analog
CF: Column Fetching
CU: Computation Unit
XB: Memristor Crossbar

1. Structural Compression on Weight Matrix

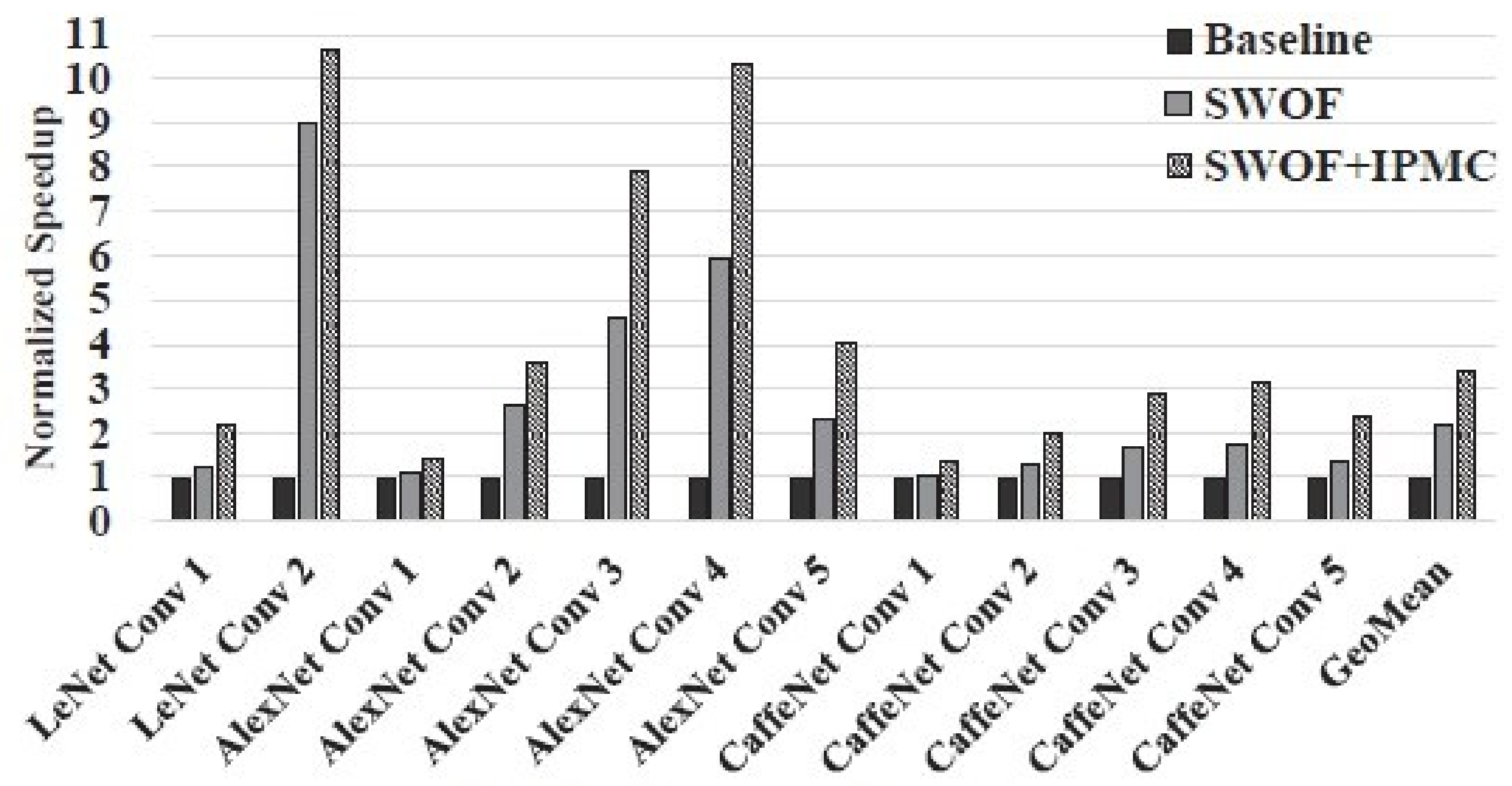
2. Structurally-compressed Weight Oriented Fetching (SWOF)

3. In-layer Pipeline for Memory and Computation (IPMC)

III. Evaluation

3.37x speedup (up to 10.66x)

2.41x energy saving (up to 9.43x)



IV. Conclusion

RECOM, the *first* accelerator to support the sparse DNN processing in ReRAM. Deep neural networks are compressed structurally by specific regularization on each layer with little or no loss in accuracy. Our experiments show notable improvement on processing speed and energy-saving.

个人信息：2014级 工科
邮　　箱：hxji1997@gmail.com

致远学院 上海交通大学致远荣誉计划
ZHIYUAN HONORS PROGRAM