Re2PIM: A Reconfigurable ReRAM-Based PIM Design for Variable-Sized Vector-Matrix Multiplication

Yilong Zhao (Speaker)

Zhezhi He, Naifeng Jing, Xiaoyao Liang, Li Jiang

Shanghai Jiao Tong University

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Outline

1 Background and Motivation

2 Architecture Design

3 Experiment Setup and Results

4 Conclusion

Outline

1 Background and Motivation

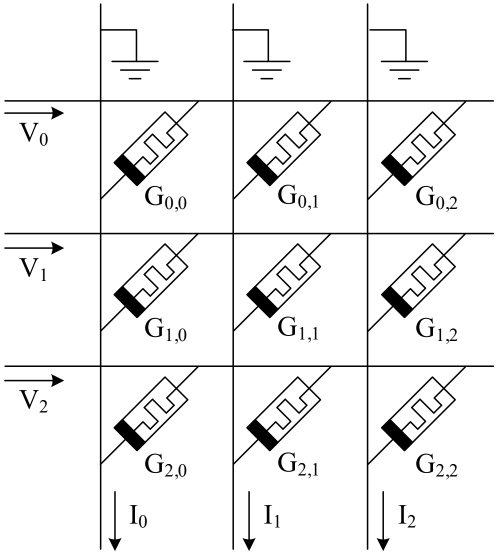
2 Architecture Design

3 Experiment Setup and Results

4 Conclusion

Background: PIM

- Process in Memory (PIM):
  - Computation happens in memory
  - Reduce data movement
- ReRAM-based Analog Vector-Matrix Multiplier
  - $O(1)$ time complexity
  - Suitable for DNN acceleration
- Advantages of ReRAM-based DNN accelerator:
  - High computational-density
  - Excellent energy efficiency
  - Superior parallelism
  - …

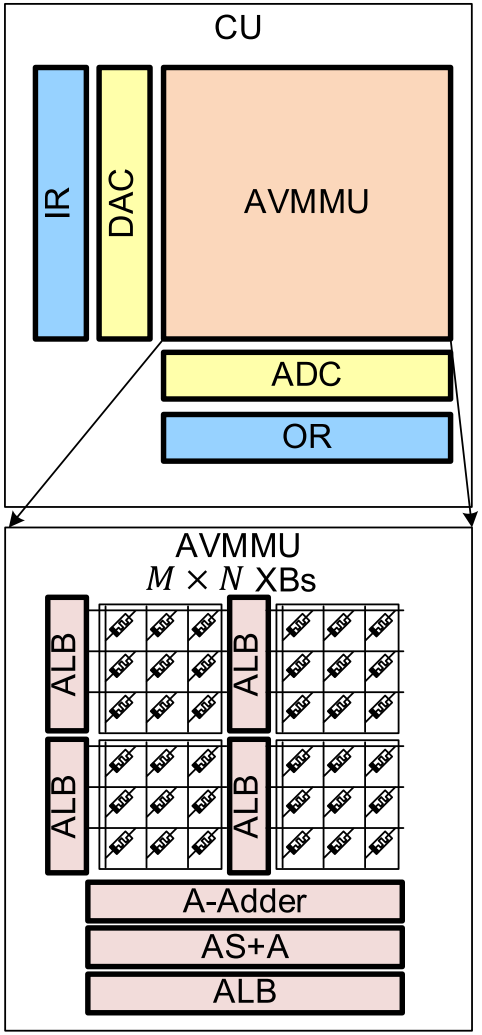


ReRAM-based Analog Matrix-vector multiplier

Vector: voltages on WLs
Matrix: conductance of ReRAM cells
Product: currents on BLs

Background: ReRAM-Based DNN Accelerator

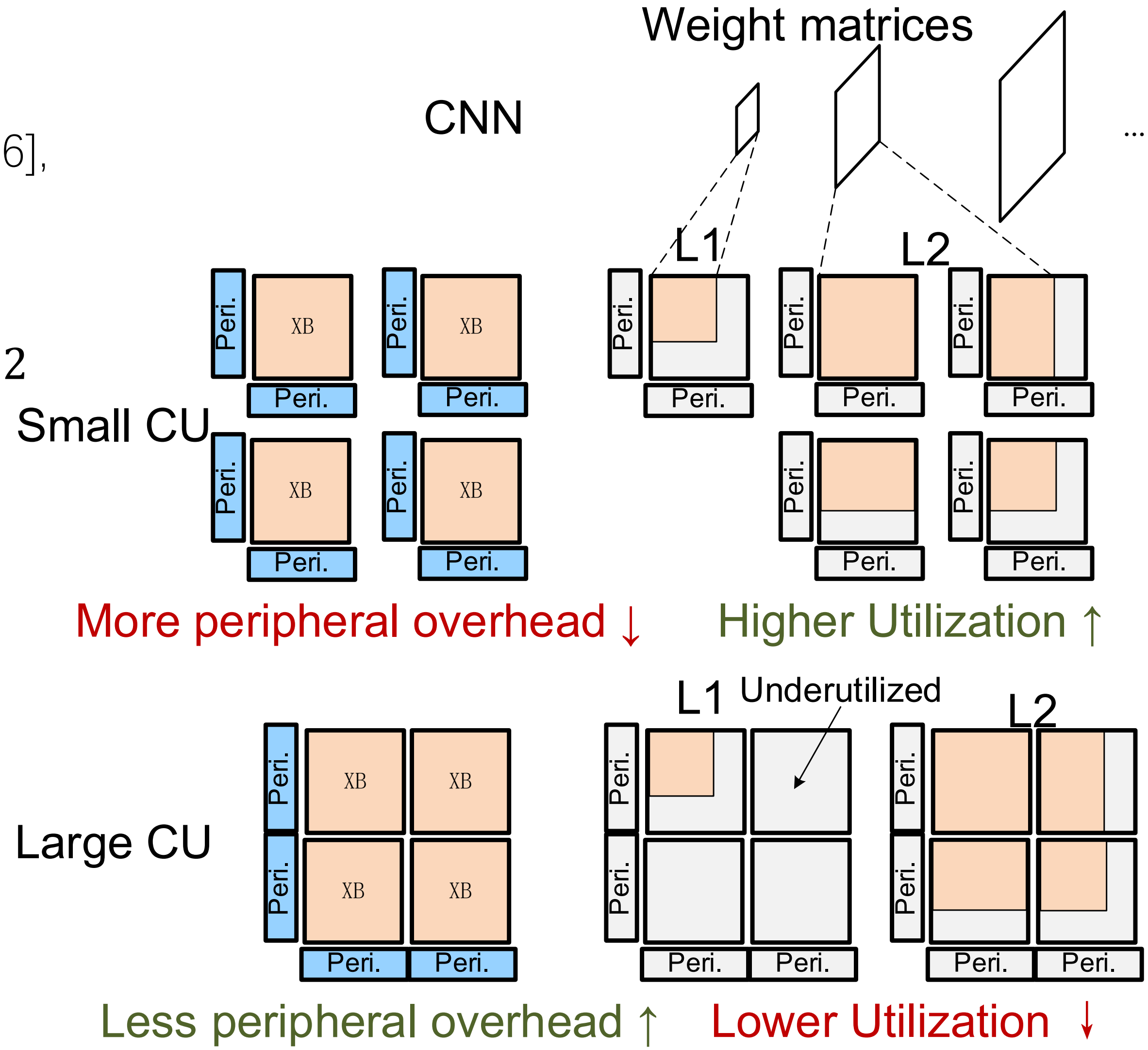
- Compute Unit (CU)
- Analog Vector-Matrix Multiplication Unit (AVMMU):
  - ReRAM Crossbar (XB)
  - Analog Local buffer (ALB)
  - Analog Adder (A-Adder)
  - Analog Shift & Add (AS+A)
- Peripheral Circuits:
  - Analog-to-Digital Converter (ADC)
  - Digital-to-Analog Converter (ADC)
  - Registers (IR, OR)
  - …



Motivation

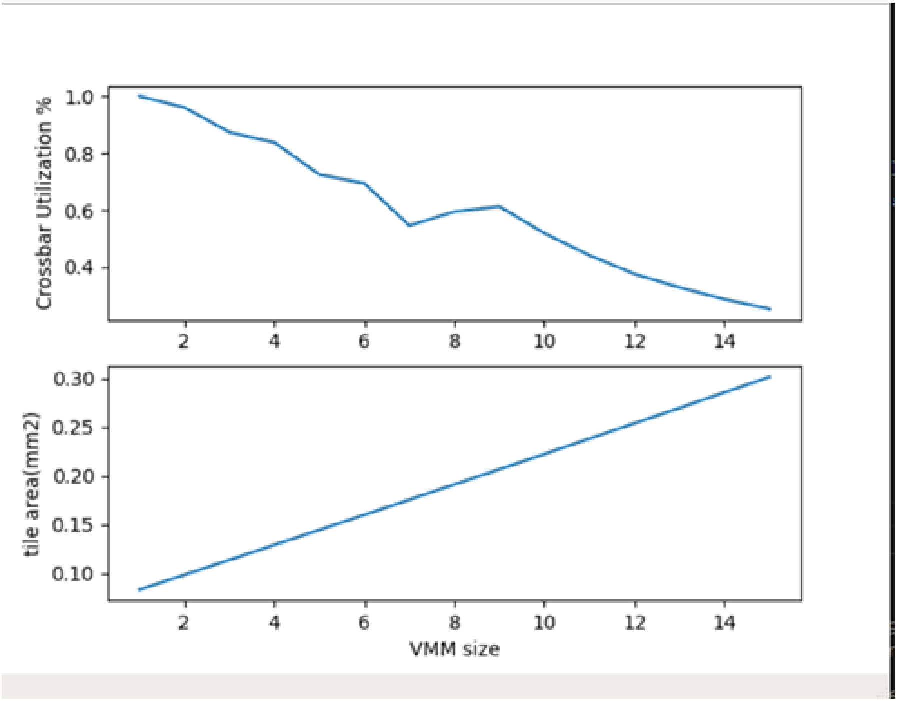
- Large CU v.s. Small CU
  - Small CU: PRIME[ISCA'16], ISAAC[ISCA'16]; one XB
  - Large CU: TIMELY[ISCA'20]; 16 × 12 XBs

Dilemma of trading off between energy efficiency and throughput !



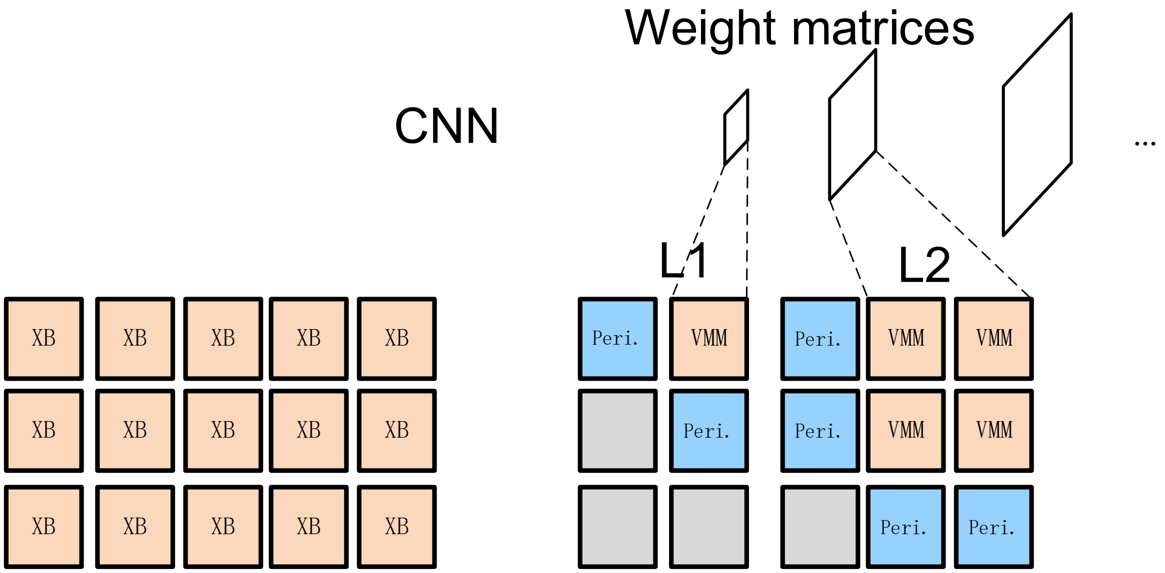
Energy efficiency-Throughput Trade-off

- Best CU size for ResNet-50 in fixed-size CU design:
  - The best VMM size is 9x16, while the utilization is less than 30%!



Main Idea

- Re2PIM (Proposed):
  - XB can be reconfigured into peripheral circuits or VMM
  - We can reconfigure variable-sized CUs which adapt to different weight matrices



Variable-sized CU

Adapt to different matrices

Outline

1 Background and Motivation

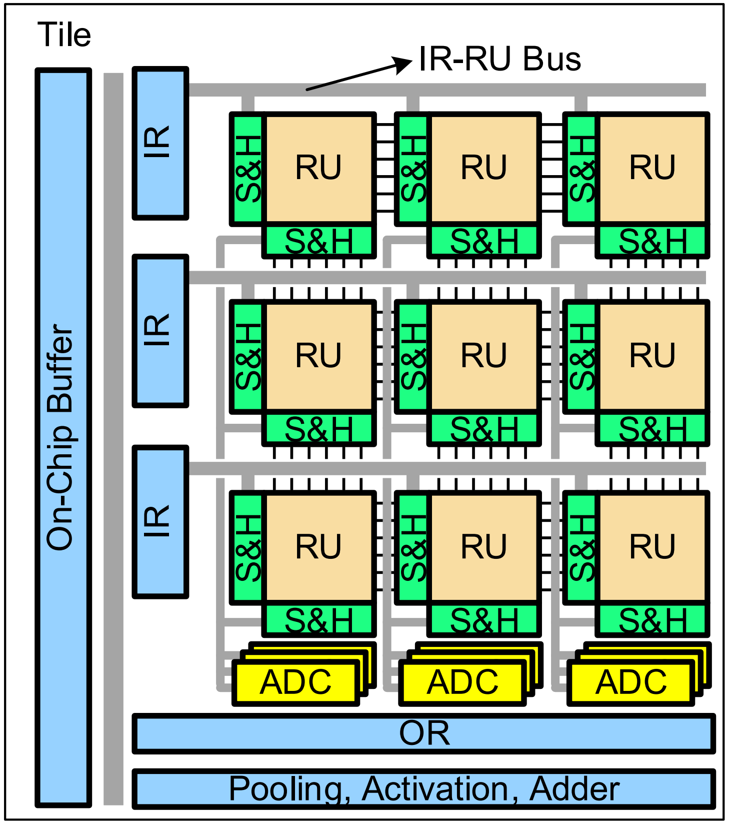
2 **Architecture Design**

3 Experiment Setup and Results

4 Conclusion

Re2PIM Architecture Overview

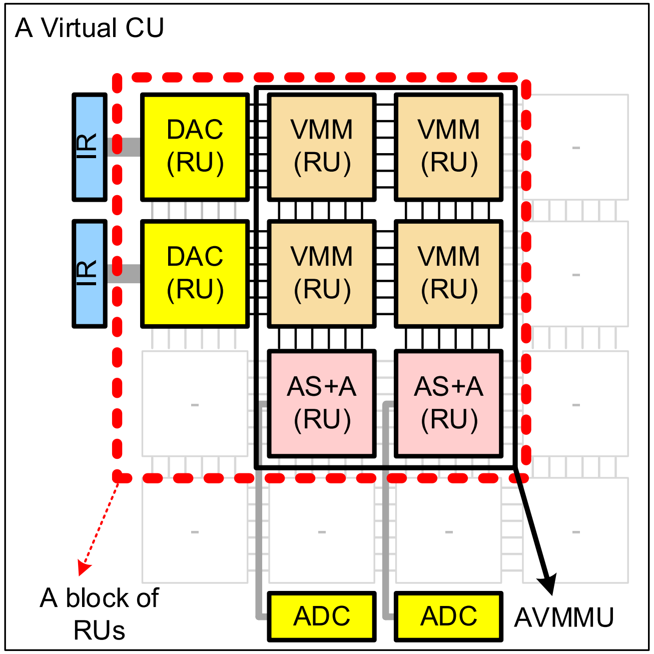
- Reconfigurable Unit (RU) , can be reconfigured into:
  - DAC
  - VMM
  - AS+A
- By reconfiguring RUs into different function, the RUs can be grouped into variable-sized CUs



(a)

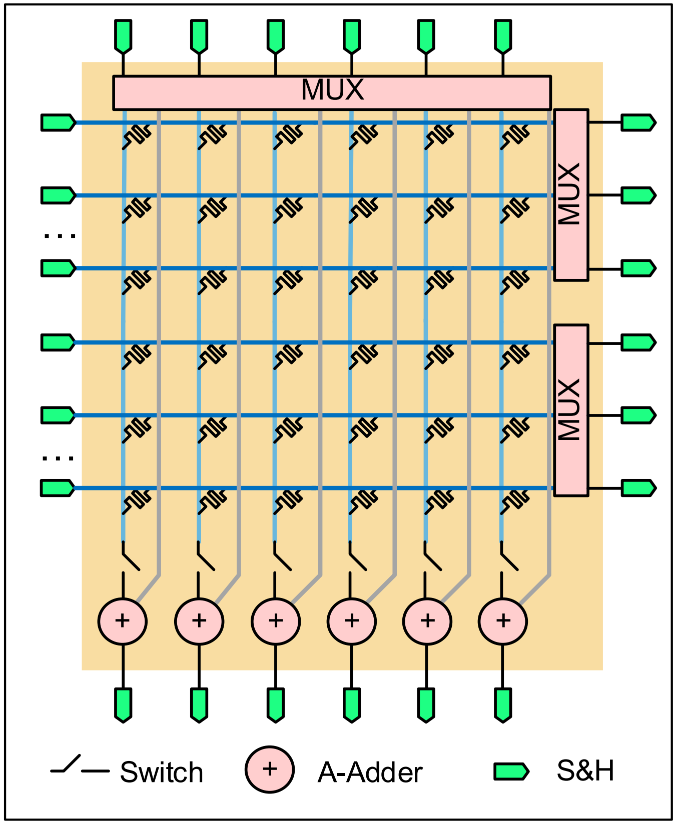
Reconfiguration

- Reconfigure RUs in to variable-sized CU:
  - RUs on the first column → DAC
  - RUs on the last row → AS+A
  - Other RUs → VMM



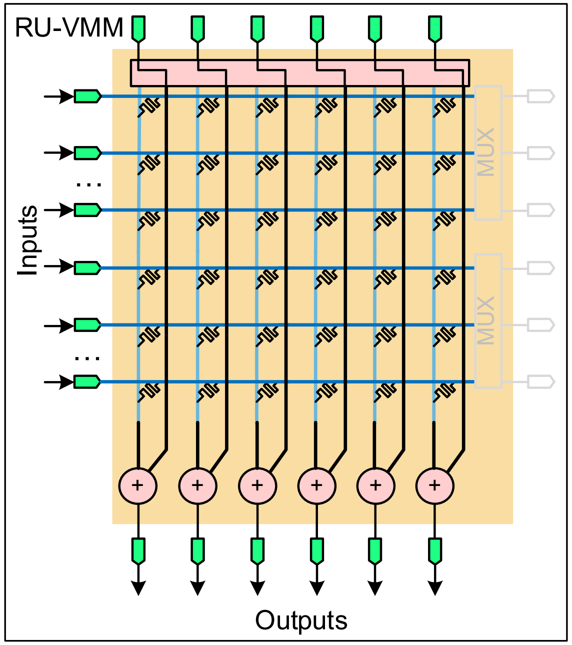
RU circuit

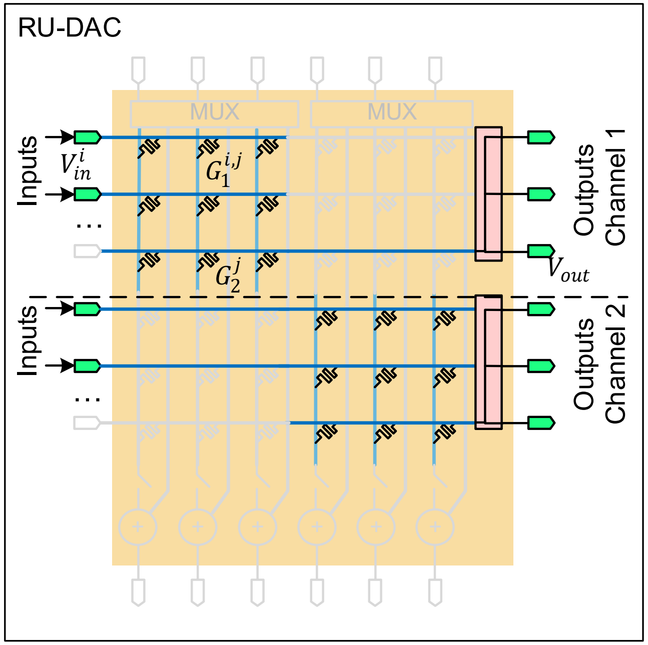
- ReRAM Crossbar
- A-Adder (mirror current source [ISCA'19])
- MUX & switches

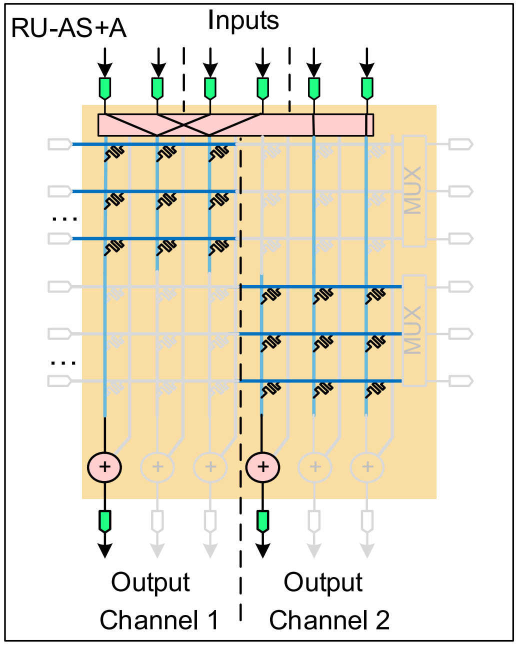


RU Circuit’s Reconfigurations

- VMM: Left → Bottom
- DAC: Left → Right
- AS+A: Top → Bottom

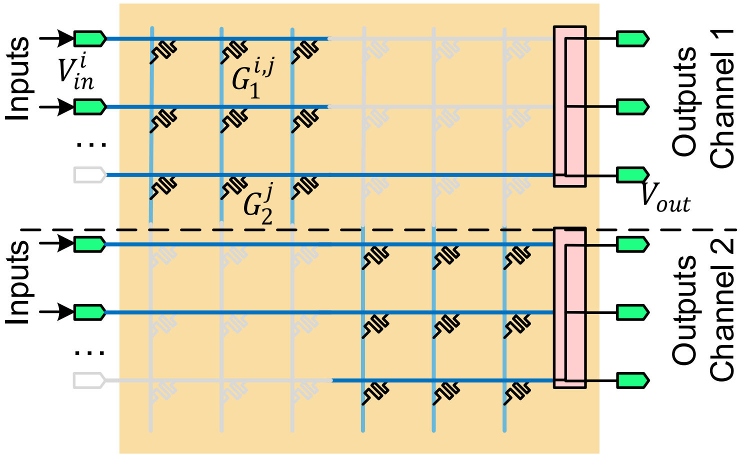






Reconfigure RU into DAC/AS+A

- How to decide the Value of ReRAM cells?
- The output voltage of DAC/AS+A are linear combination of input voltages: $V_{out} = \sum_{i=0}^{K-1} k_i \cdot V_{in}^{i}$
- Input-Output relation of the circuit: $V_{out} = \sum_i \frac{\sum_j G_1^{i,j} G_2^{j}}{\sum_{i,j} G_1^{i,j} G_2^{j}} \cdot V_{in}^{i}$
- Define a IP problem with constrains:
    - $0 \le G_1^{i,j}, G_2^{j} \le l$
    - Coefficients of V_{in}^{i} equals to k_i
    - A constrain we add for simplification: $\forall j, \sum_i G_1^{i,j} + G_2^{j}$ are equal
- Solve the IP problem with solvers



Outline

1 Background and Motivation

2 Architecture Design

3 **Experiment Setup and Results**

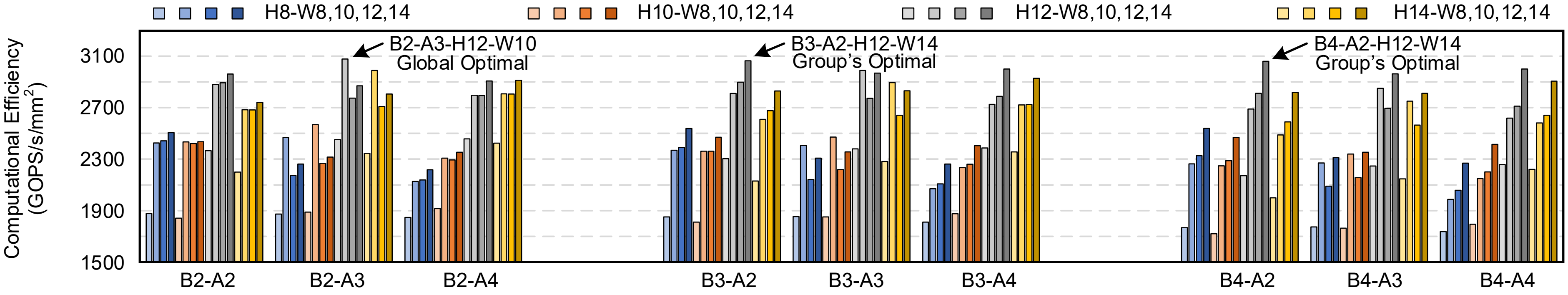
4 Conclusion

Experiment Setup

- Baselines:
  - Large CU: PRIME [ISCA’16], ISAAC [ISCA’16]
  - Small CU: TIMELY [ISCA’20]
- Networks:
  - CNNs: AlexNet, VGG, MSRA, ResNet, MobileNet
  - NeuralTalk
  - Bert
- AVMMU circuit: PySpice
- IP Solver: ScipOpt

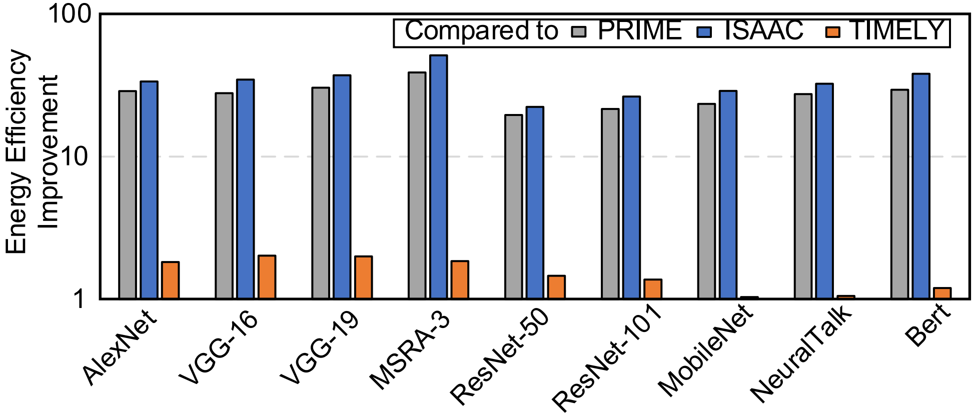
Design Space Exploration

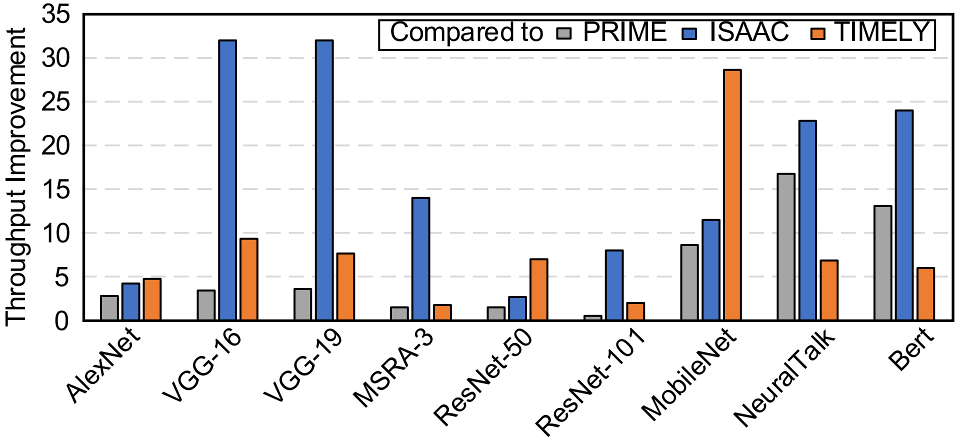
- Parameter:
  - H, W: RU number on a row/column
  - A: ADC number in a Tile
  - B: IR-RU bus width
- Criteria: Average Computational Efficiency over the benchmarks
- B2-A3-H12-W10 reaches the best



Throughput & Energy Efficiency Improvement

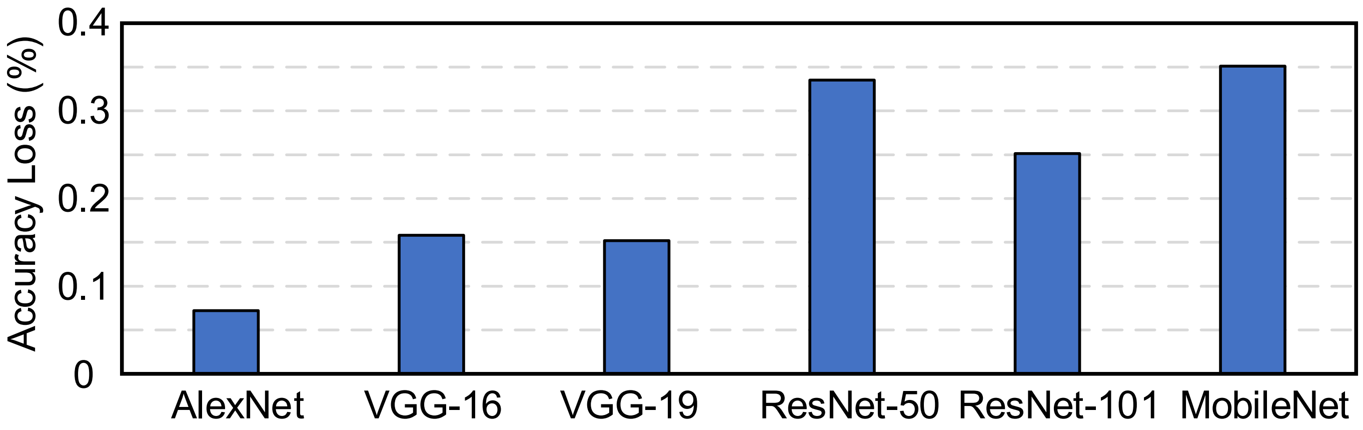
- 27×/34×/1.5× in energy efficiency
- 5.7×/17×/8.2× in throughput
- over PRIME/ISAAC/TIMELY





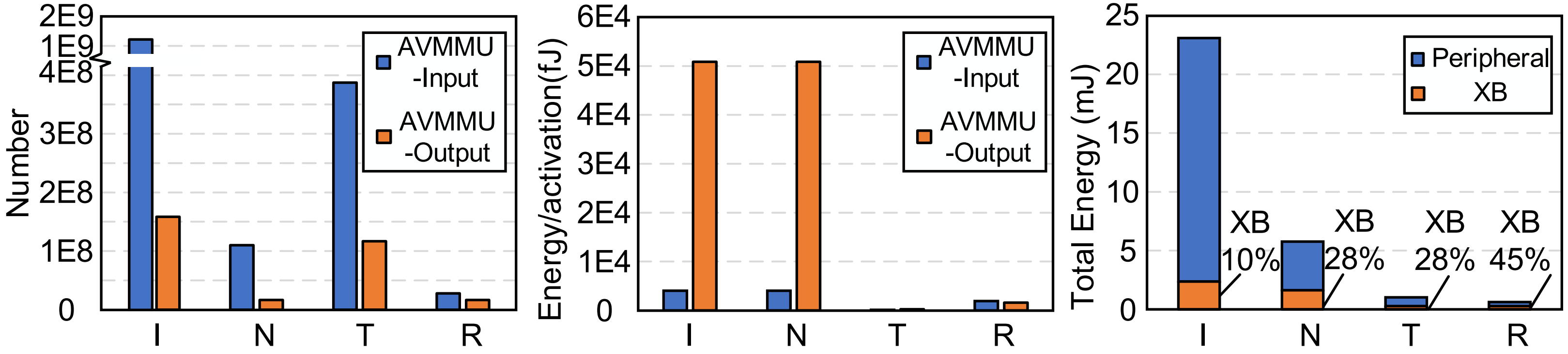
Accuracy Loss

- Simulate the RU circuit with PySpice
  - ReRAM variation: $\boldsymbol{\theta} = \mathbf{0.025}$ [ICCAD'19], $\boldsymbol{b} = \mathbf{0.0015}$ [DAC'19]
- Accuracy loss < 0.5% over all the CNN benchmarks



Analysis

- CU size's impact on energy efficiency
  - ISAAC: More AVMMU-Input/Output → Low energy efficiency
  - TIMELY & Re2PIM: Fewer AVMMU-Input/Output → High energy efficiency

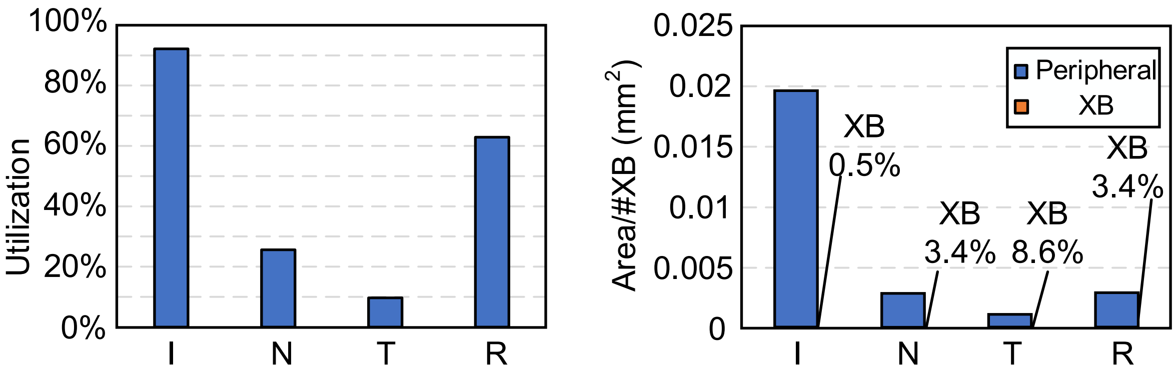


I:ISAAC   T: TIMELY   R: Re2PIM

N: A naïve design, directly enlarge ISAAC's CU size

Analysis (2)

- CU size's impact on throughput
  - ISAAC: Small CU size → Good utilization
  - TIMELY: Large CU size → Low utilization
  - Re2PIM: Variable-sized CU → Good utilization, best throughput



I:ISAAC   T: TIMELY   R: Re2PIM
N: A naïve design, directly enlarge ISAAC's CU size

Outline

1 Background and Motivation

2 Architecture Design

3 Experiment Setup and Results

4 Conclusion

Conclusion

- A reconfigurable ReRAM-based accelerator named Re2PIM:
  - Mainly composed of arrays of RUs
  - RU can be reconfigured into VMM, AS+A or DAC
  - RUs can be grouped into variable-sized CU
- Achieve high energy efficiency without damaging throughput

Q&A

Thank You!

Re2PIM: A Reconfigurable ReRAM-Based PIM Design for Variable-Sized Vector-Matrix Multiplication

Yilong Zhao (Speaker), Zhezhi He, Naifeng Jing, Xiaoyao Liang, Li Jiang

Shanghai Jiao Tong University