CR2: A Compressed ReRAM-based DNN Accelerator by Combining Computation and Read operation

Shi-Hao Hong and Yeh-Ching Chung
The Chinese University of Hong Kong, Shenzhen

Background

- ReRAM-based PIM: A non-von neumann architecture performing Matrix-Vector Multiplication in memory cell arrays.

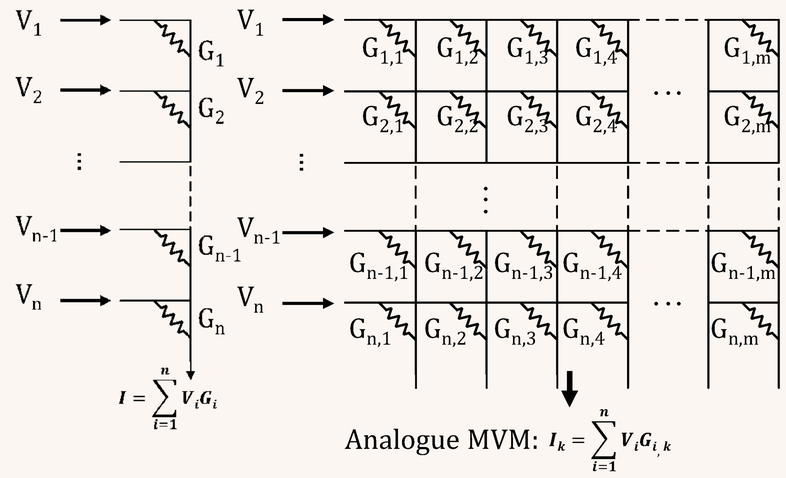

Figure 1: ReRAM-based MVM computation

- Operation Unit (OU): Due to hardware limits, it is necessary to split an MVM into multiple operation units (OUs).
- Sparsity pruning and repetition sharing are common methods to compress crossbar arrays in ReRAM-based PIM.

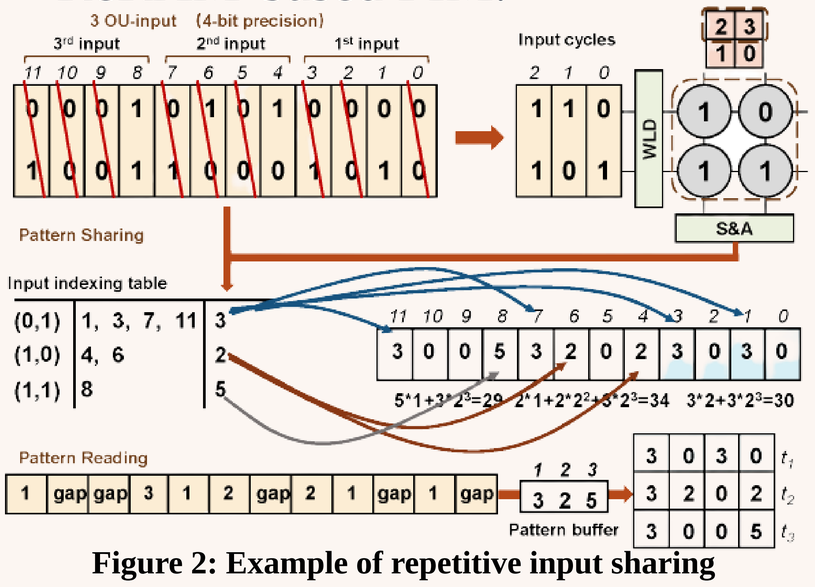

Figure 2: Example of repetitive input sharing

Motivation

Current studies have not explored the concentrated distribution of inputs. Preliminary experiments reveal that certain input patterns are frequently repeated while changing datasets. It allows us to replace repetitive computations with reading buffer.

Design of CR2

- Architecture: CR2 consists mainly of weight pattern compute (WPC), weight pattern reuse (WPR) and input pattern reuse (IPR).

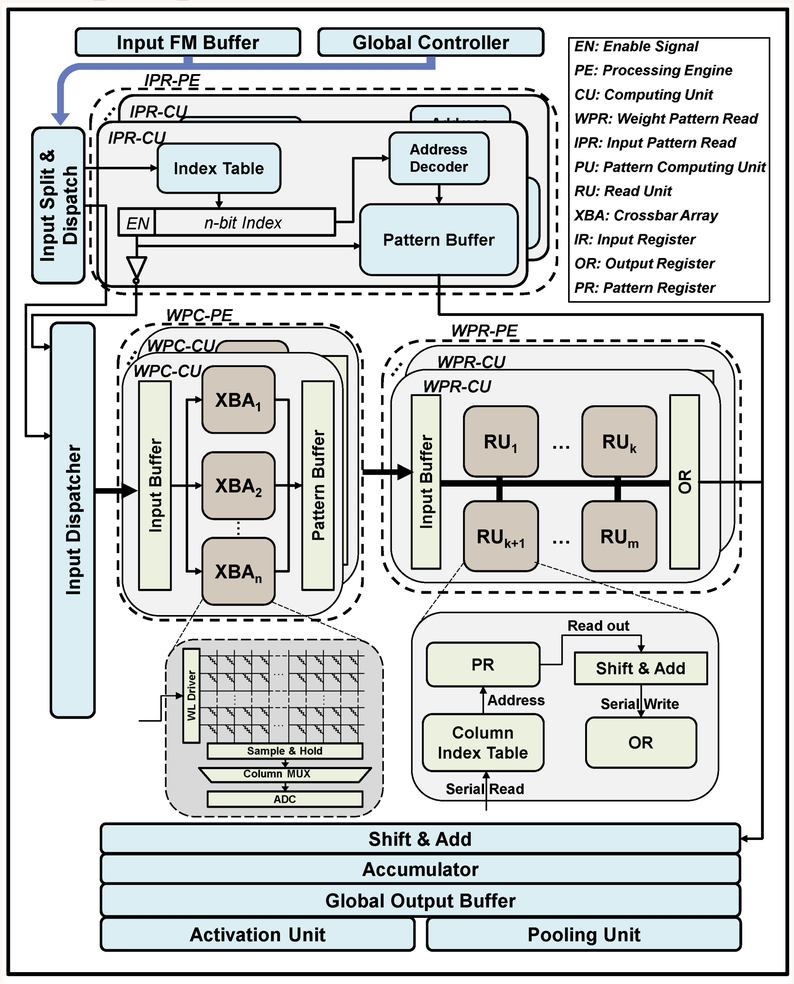

Figure 4: CR2 Architecture

Size of Learning data (SL), Size of Testing data (ST) are parameters in learning allocation strategy

we propose a buffer allocation strategy tailored to repeating input patterns at each layer

$$\max \sum_j P_j[Bj] \ s.t. B_j \geq 0, \sum_j B_j W_j \leq B$$

Recurrence searching: $Pmax(i,w) = \max(Pmax(i-1, w-kwi) + Pi[k]), \ k \in [0, \min(xi, \lfloor \frac{w}{w_i} \rfloor)]$

Evaluation

Benchmark: Imagenet task
Models: AlexNet, VGG16 ResNet50 and GoogLeNet, where weights are quantized to 8-bit.

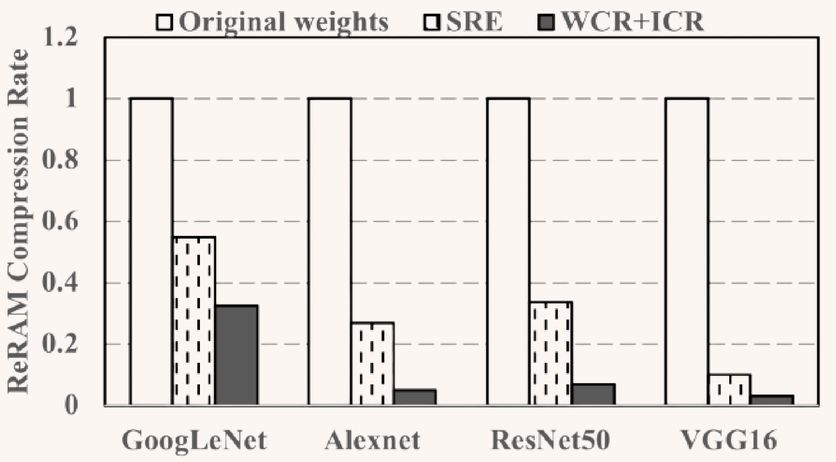

Figure 5: ReRAM compression rate of CR2 and SRE normalized to baseline

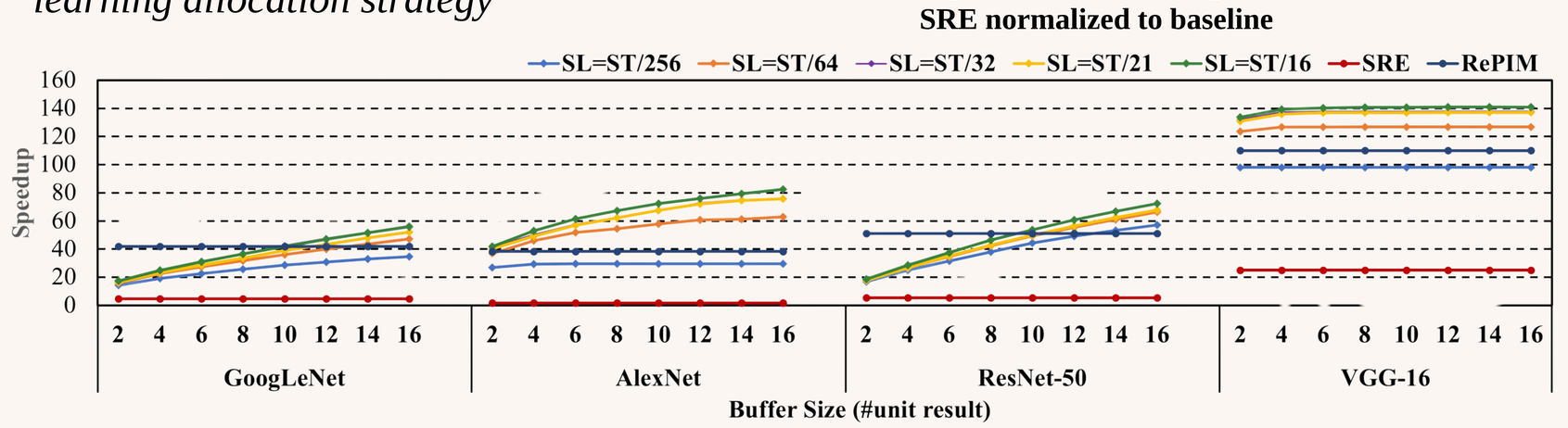

Figure 6: Performance speedup of different approaches

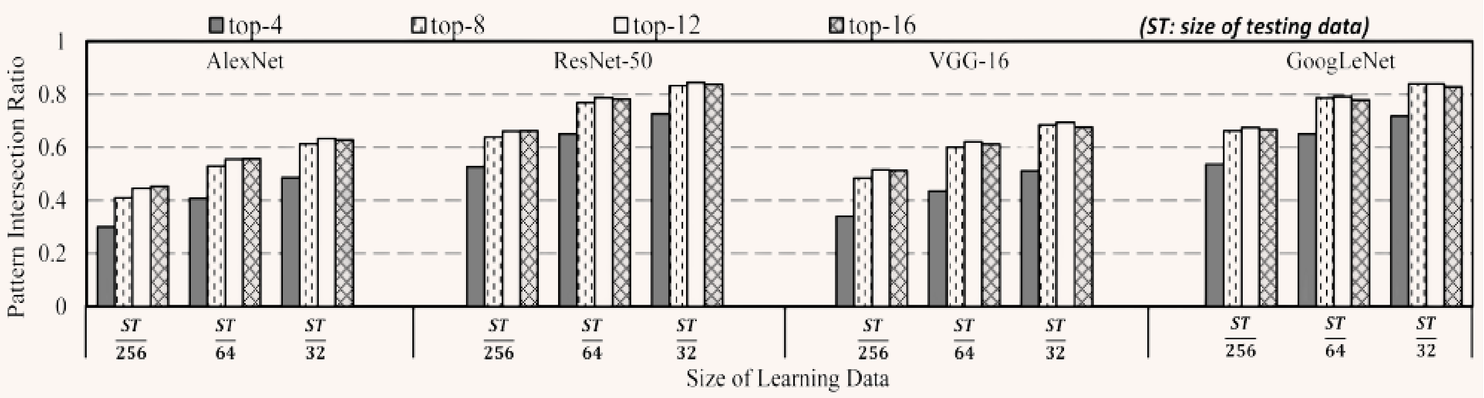

Figure 3: Intersection Ratio of Input Patterns with different top-k and learning data sizes.