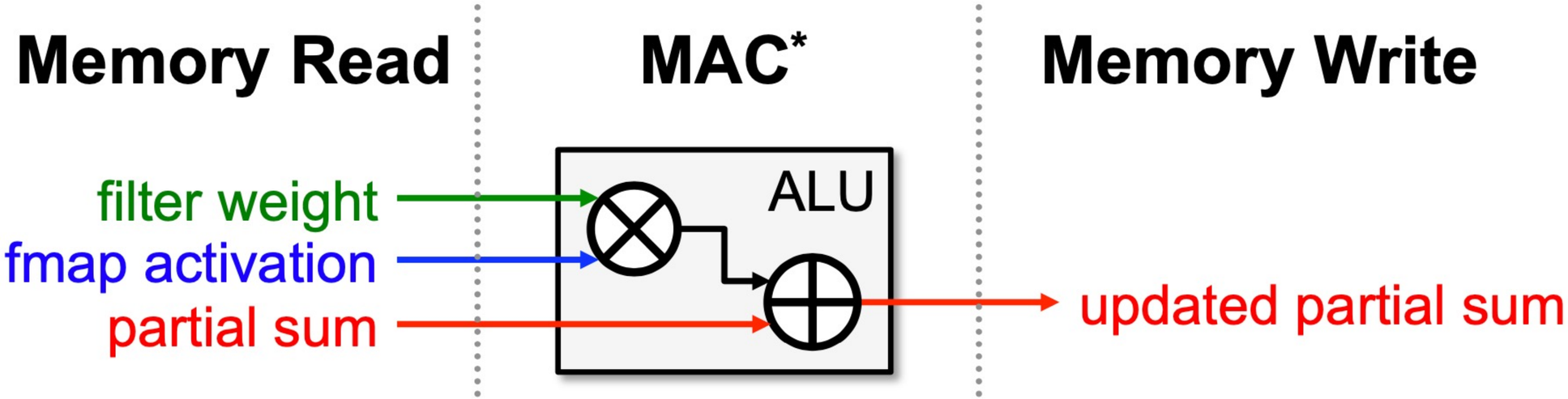
Electrical Engineering, National Chung Hsing University
Multimedia & Communication IC Design Lab

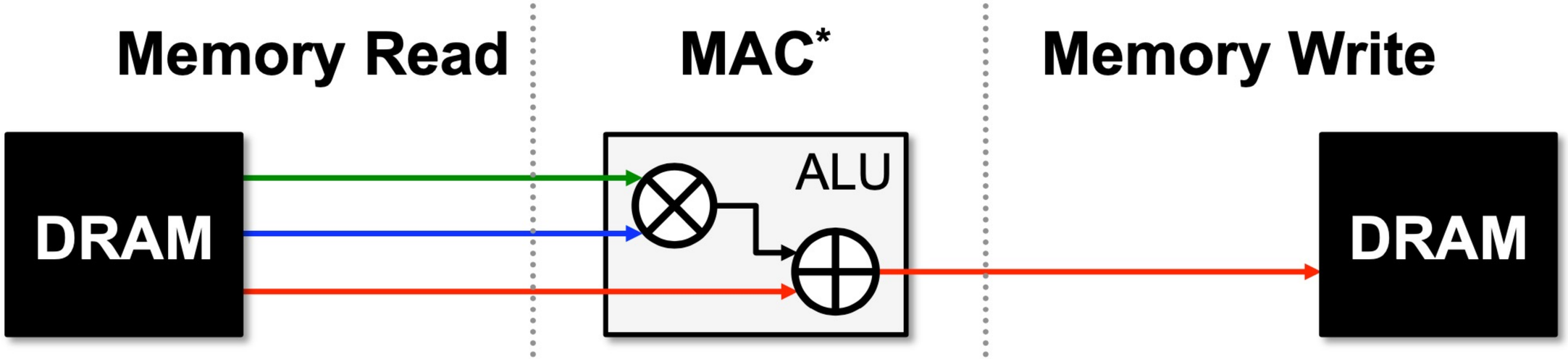
# Eyeriss

An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks

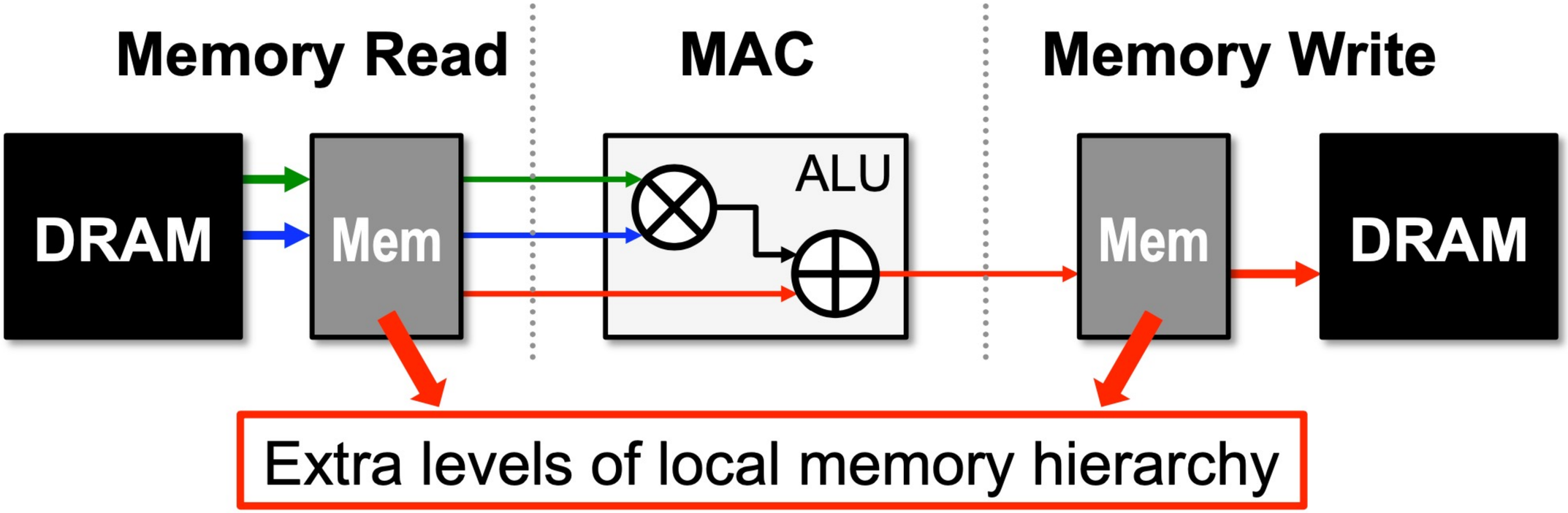
國立中興大學
National ChungHsing University

# Memory Access is the Bottleneck



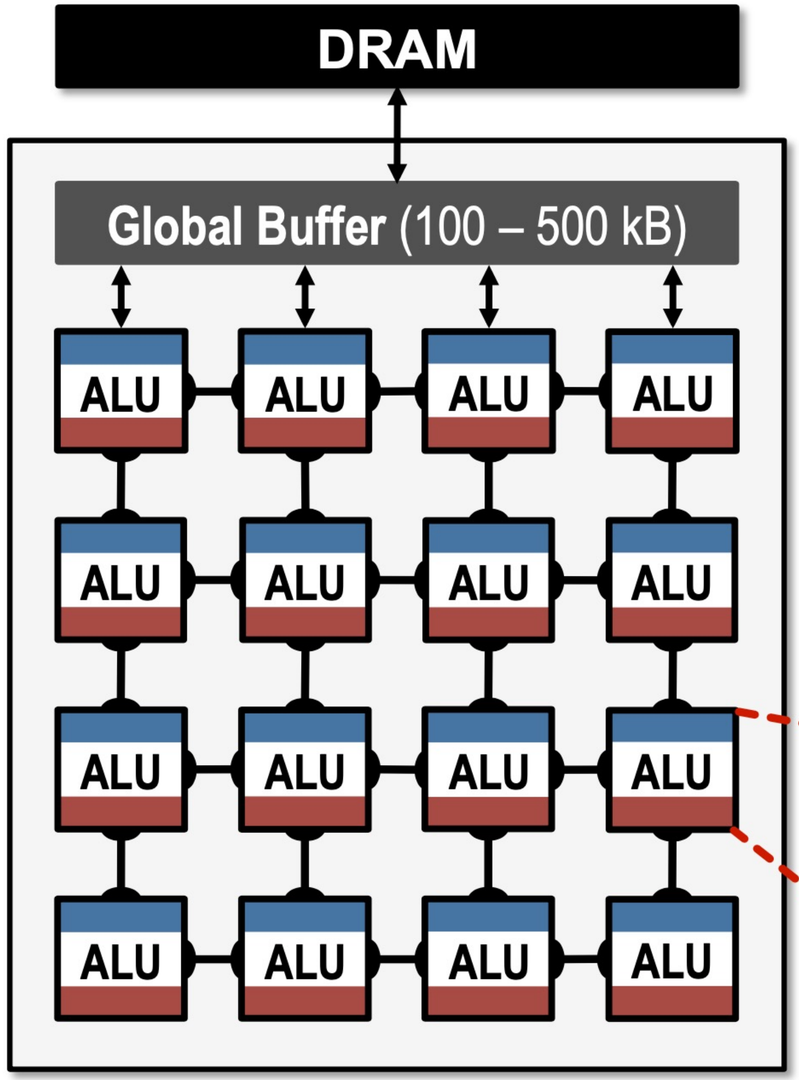



all memory R/W are **DRAM** accesses => **worst case**




- data reuse
- local accumulation

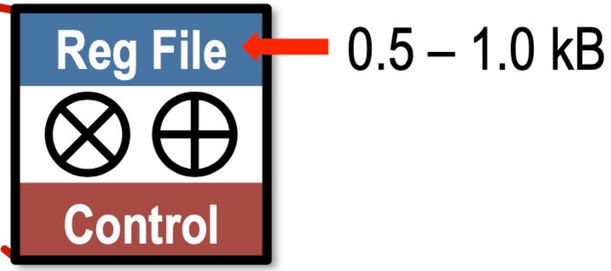
# Architecture

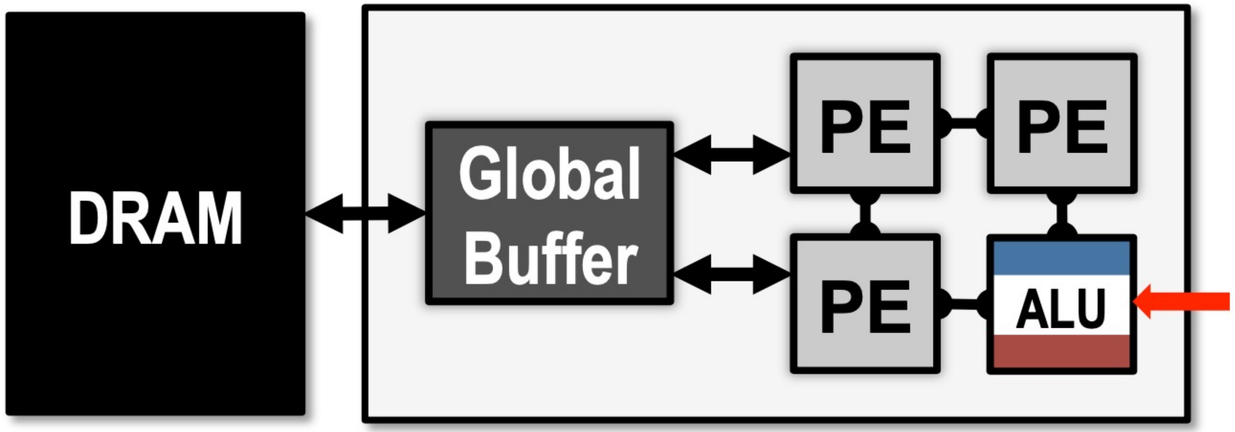


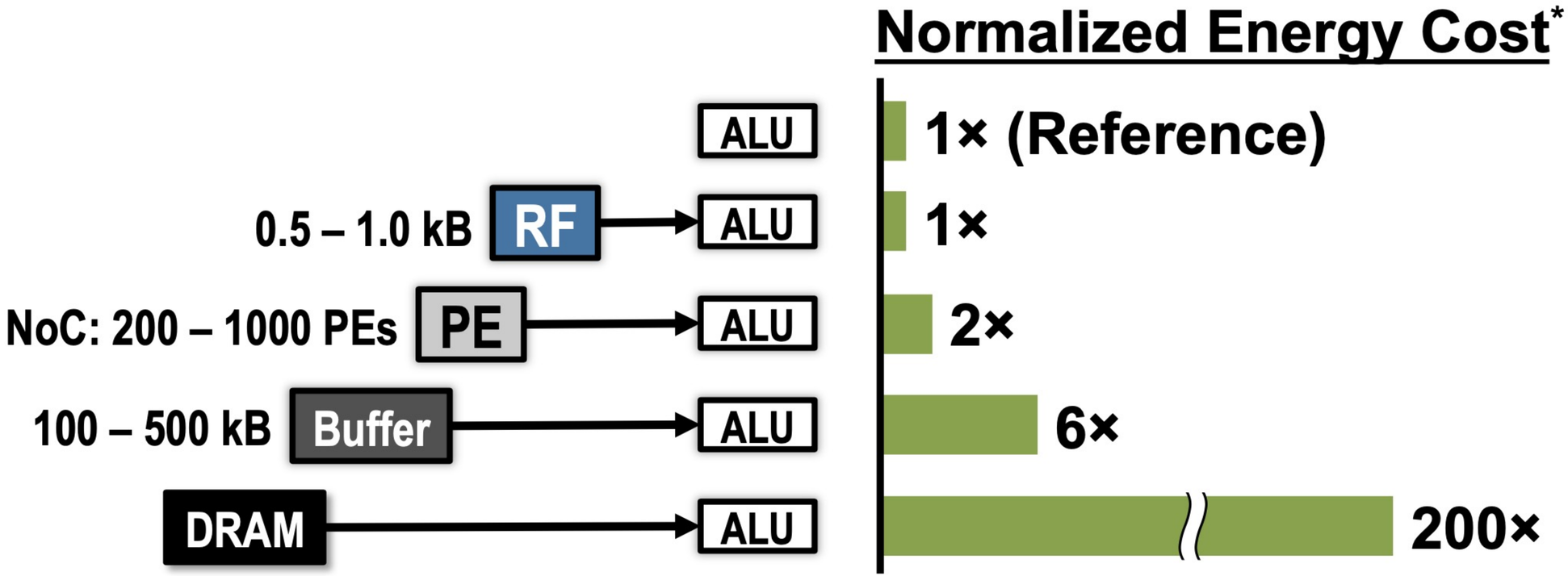

**Local Memory Hierarchy**

- Global Buffer
- Direct inter-PE network
- PE-local memory (RF)

**Processing Element (PE)**

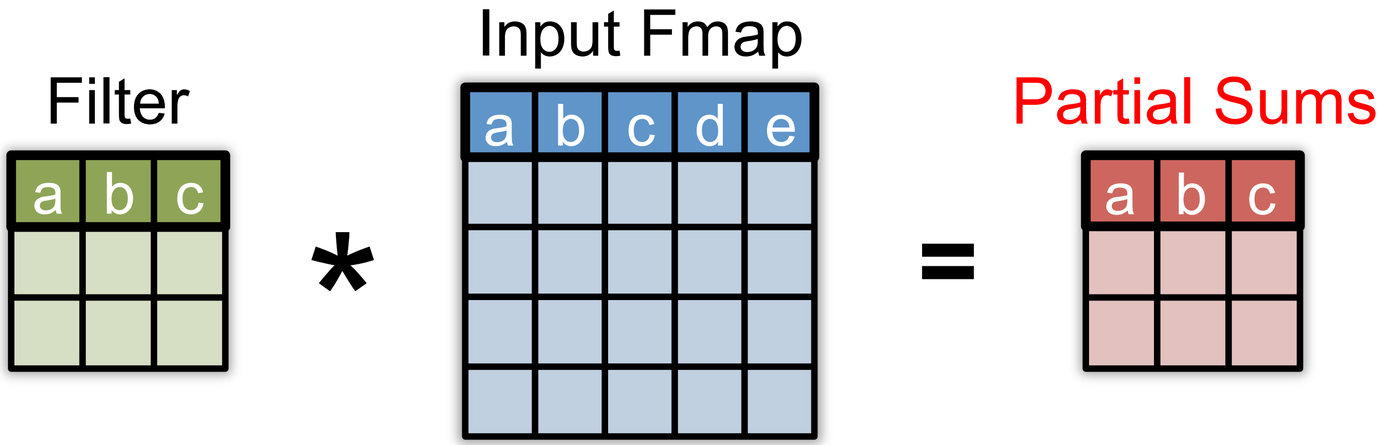


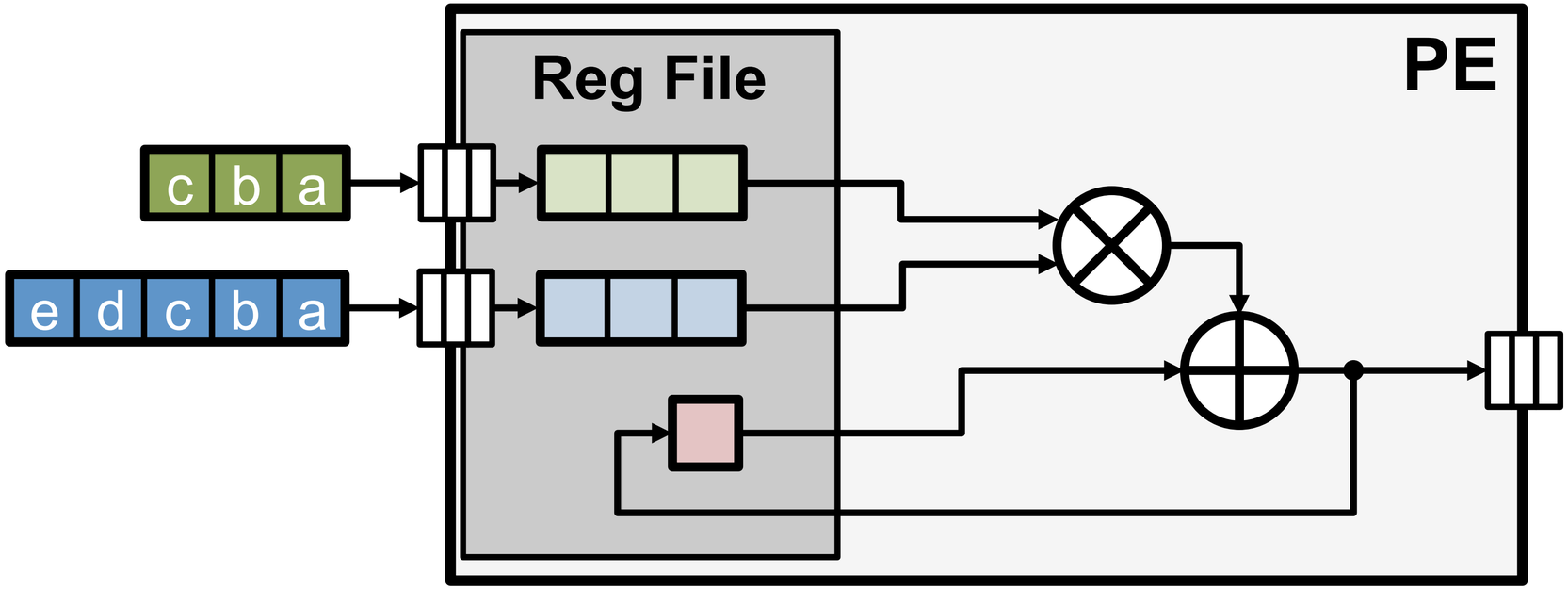


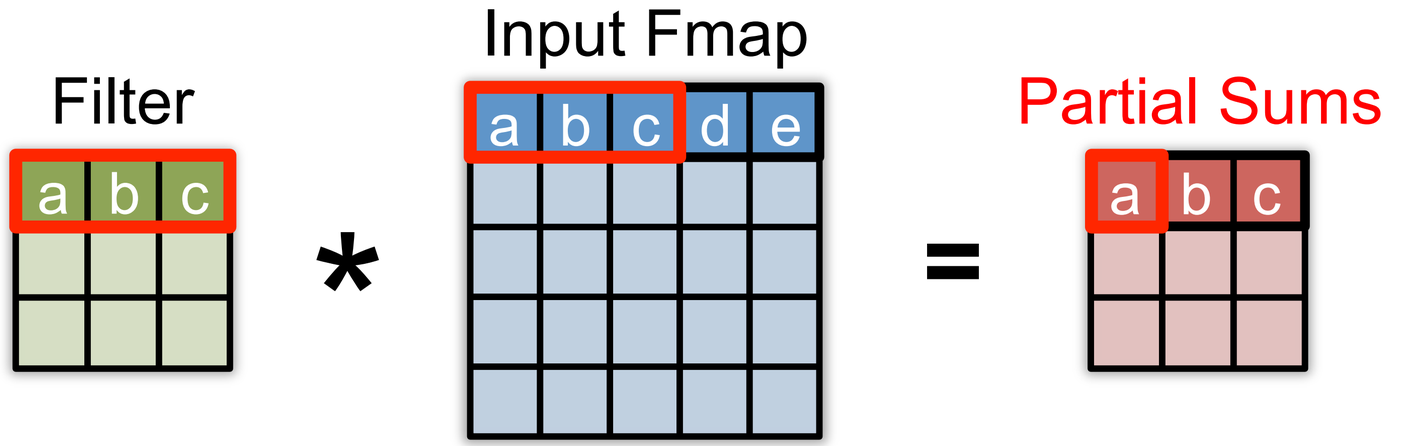



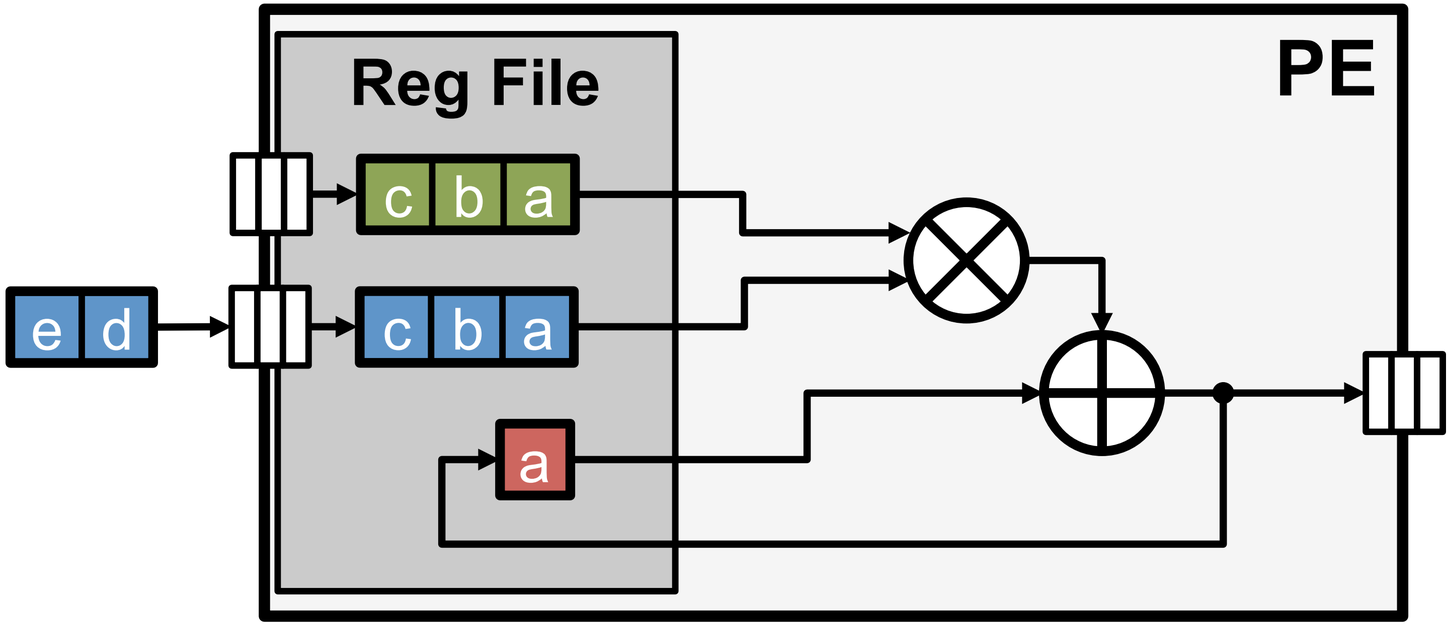
# Dataflow Taxonomy

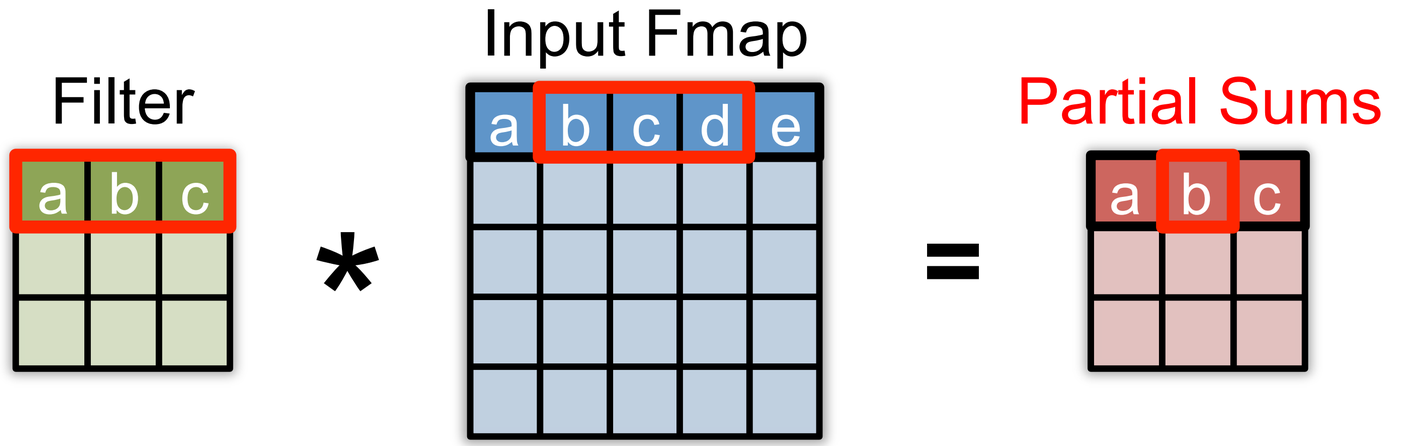
- Weight Stationary
- Output Stationary
- Row Stationary

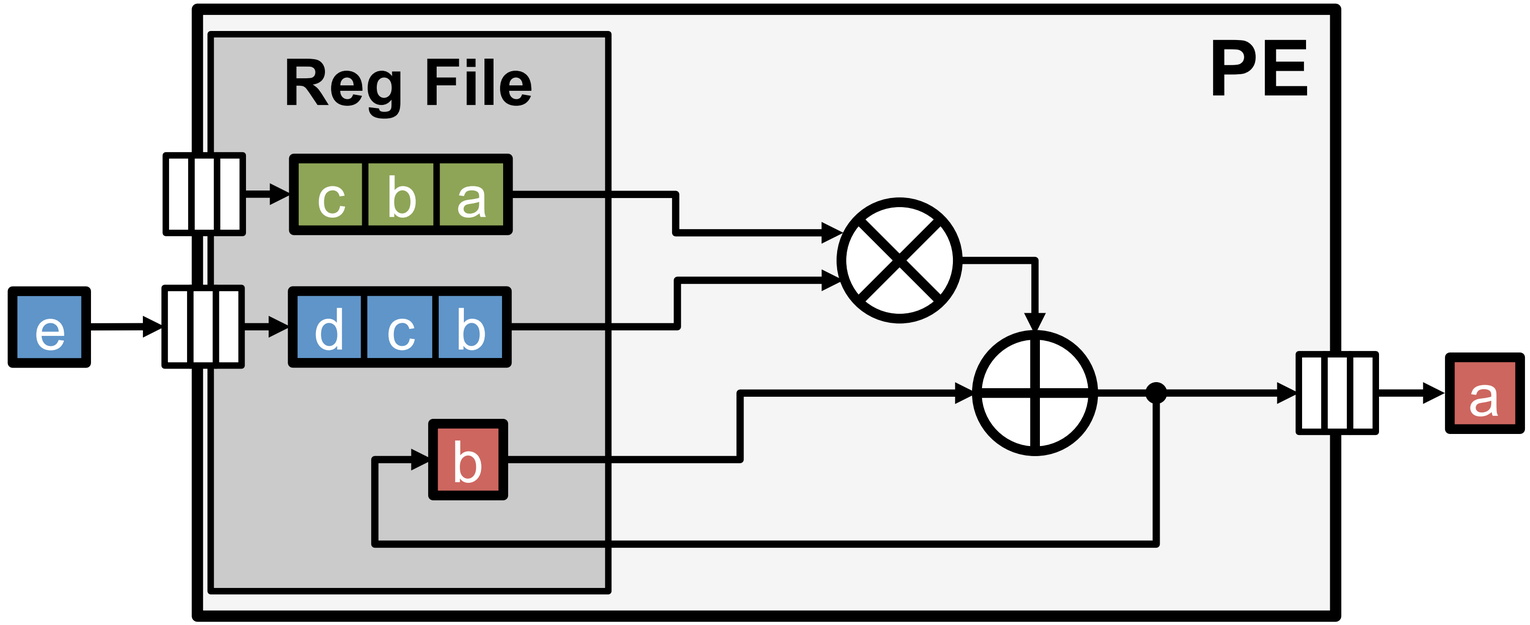


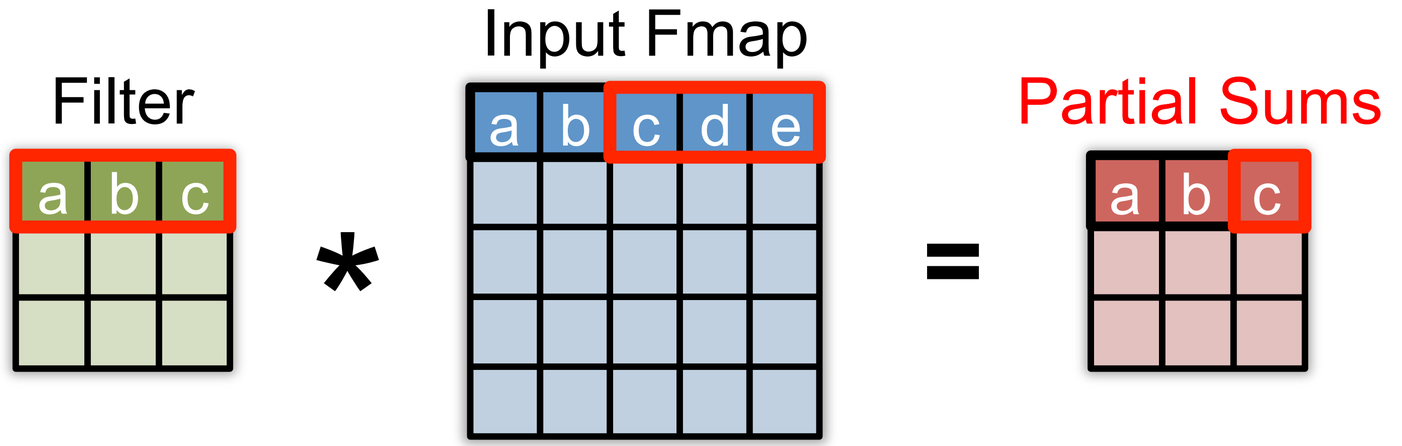



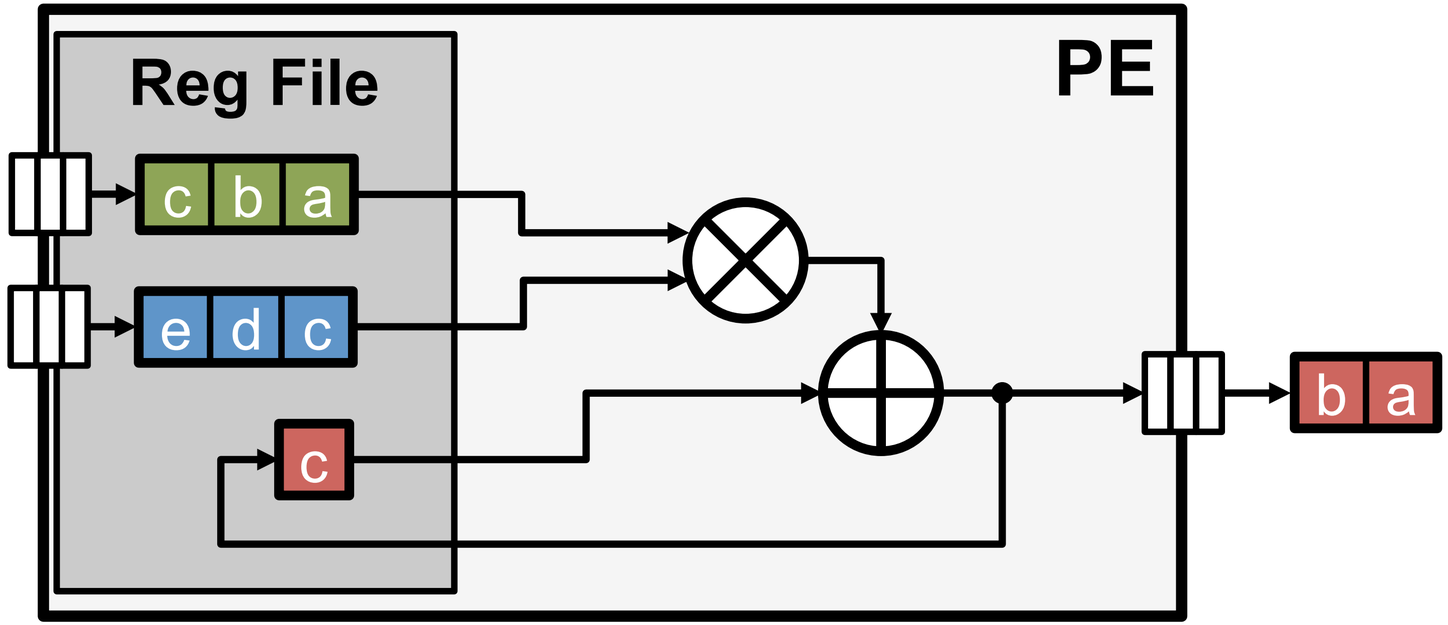


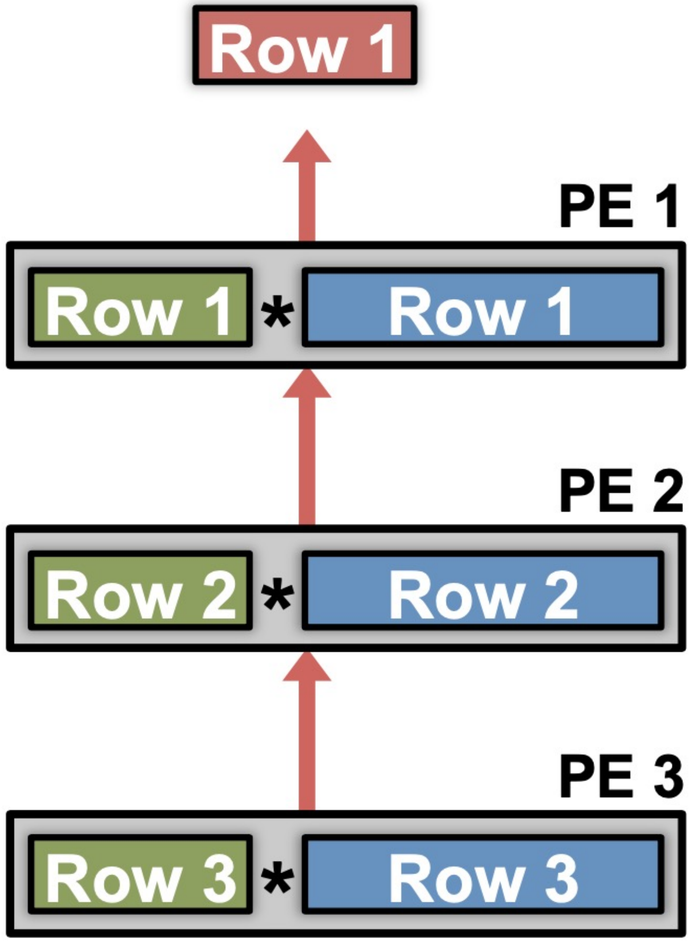



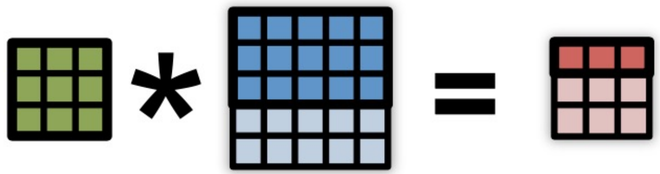


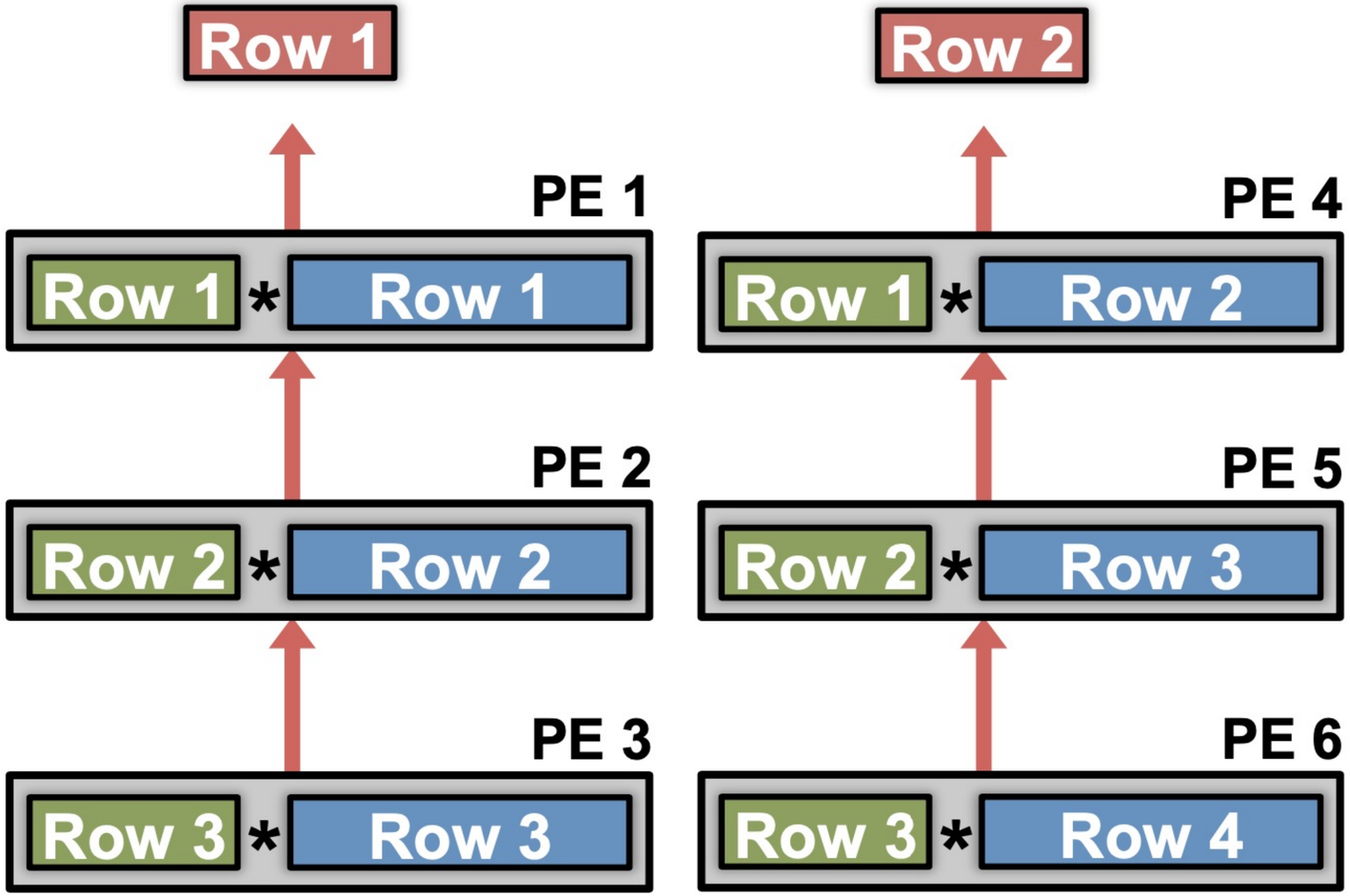



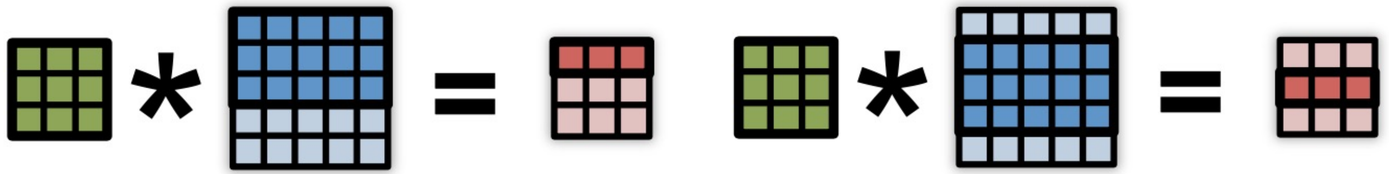


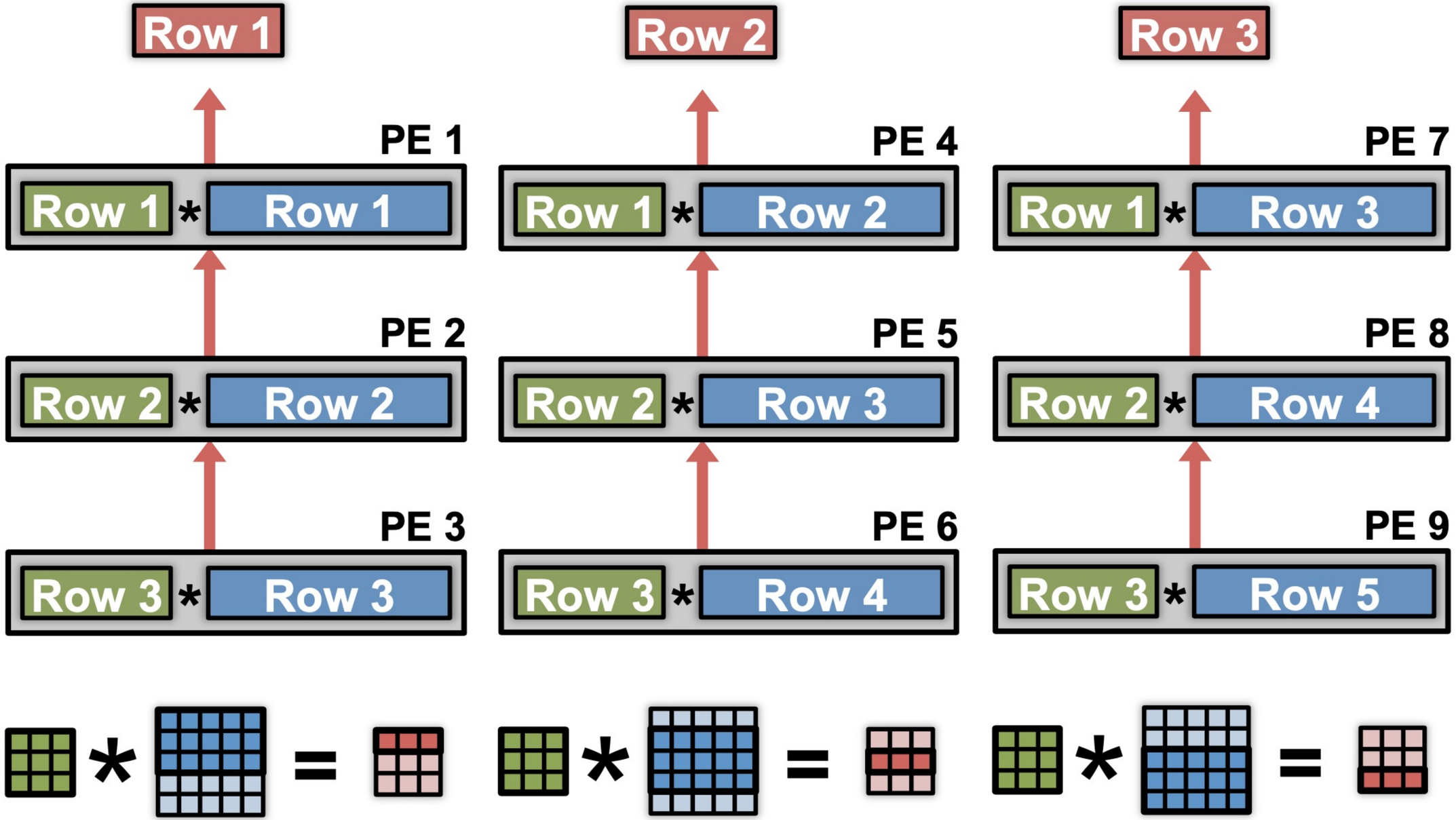



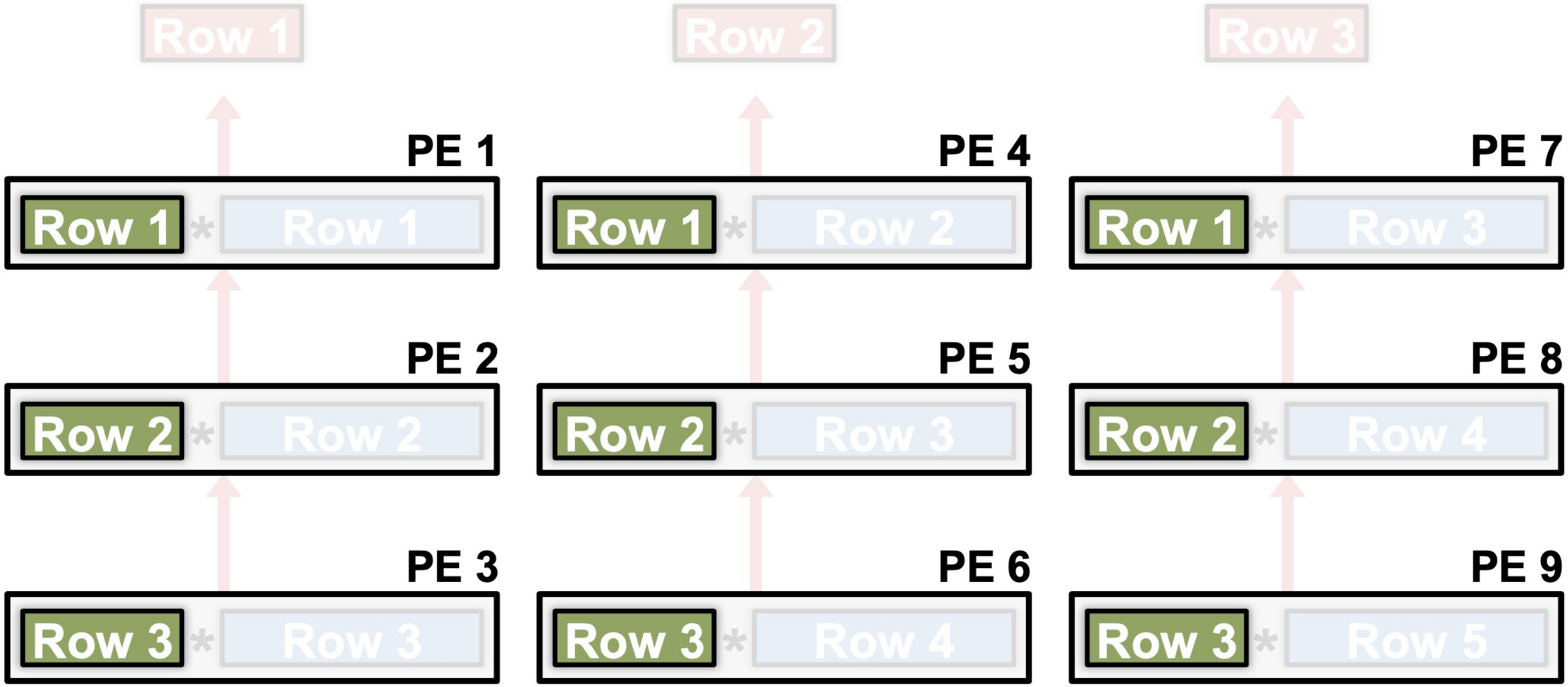

Row 1
PE 1
Row 1 * Row 1
PE 2
Row 2 * Row 2
PE 3
Row 3 * Row 3



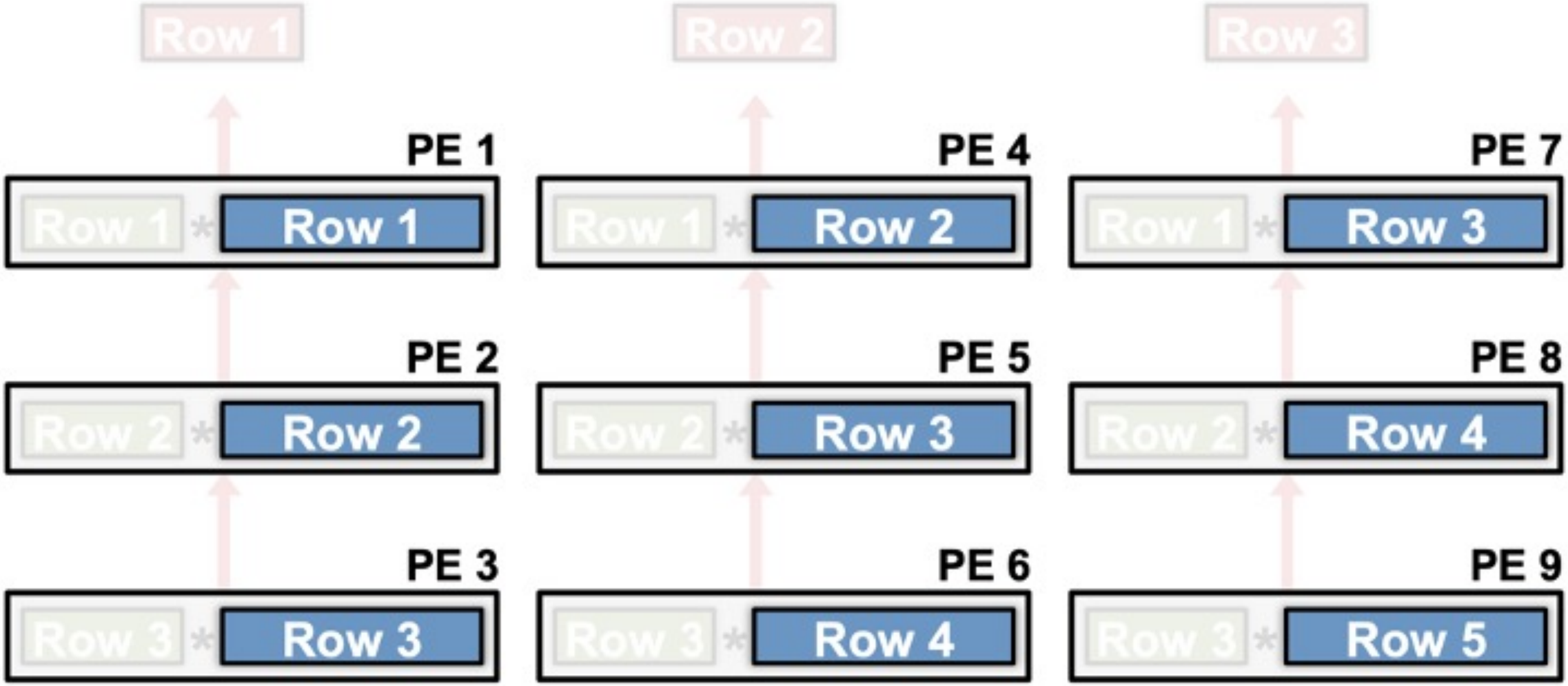

Row 1
Row 2
PE 1
Row 1 * Row 1
PE 4
Row 1 * Row 2
PE 2
Row 2 * Row 2
PE 5
Row 2 * Row 3
PE 3
Row 3 * Row 3
PE 6
Row 3 * Row 4





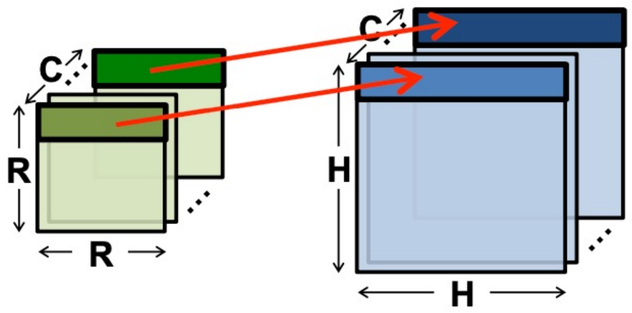


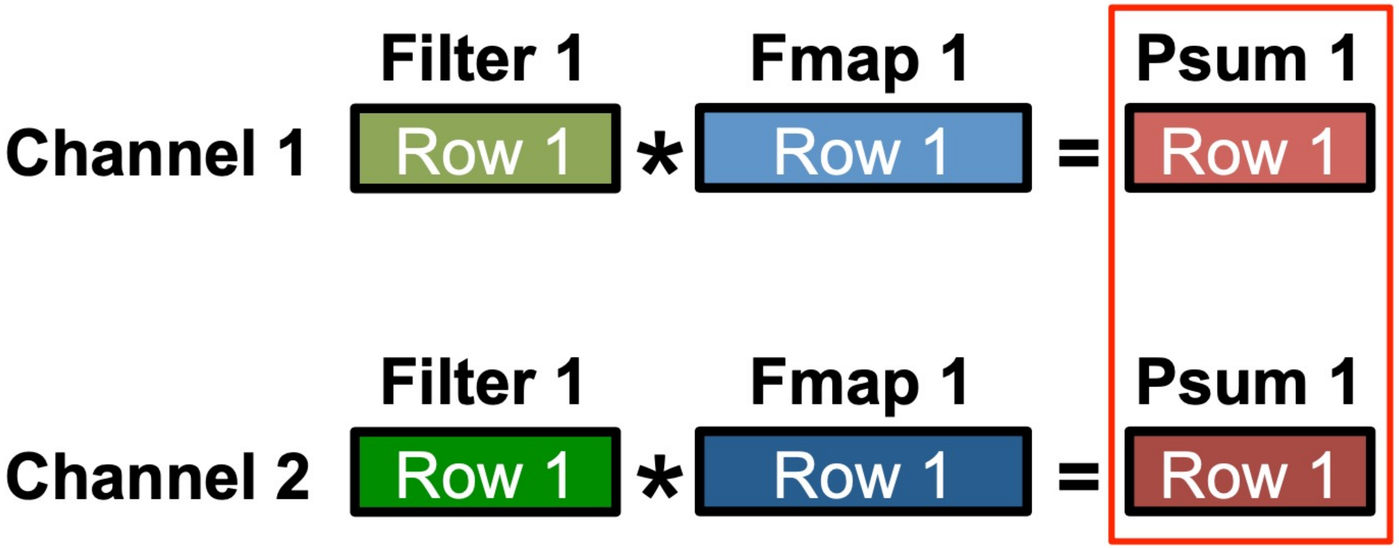

**Filter rows** are reused across PEs **horizontally**




**Fmap rows** are reused across PEs **diagonally**

# multiple channels



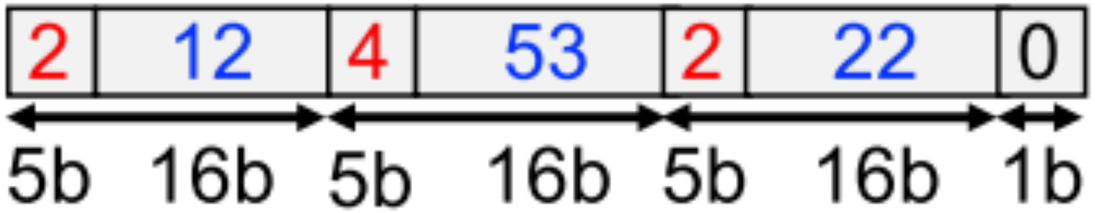

accumulate psums

# Run-Length Coding (RLC)

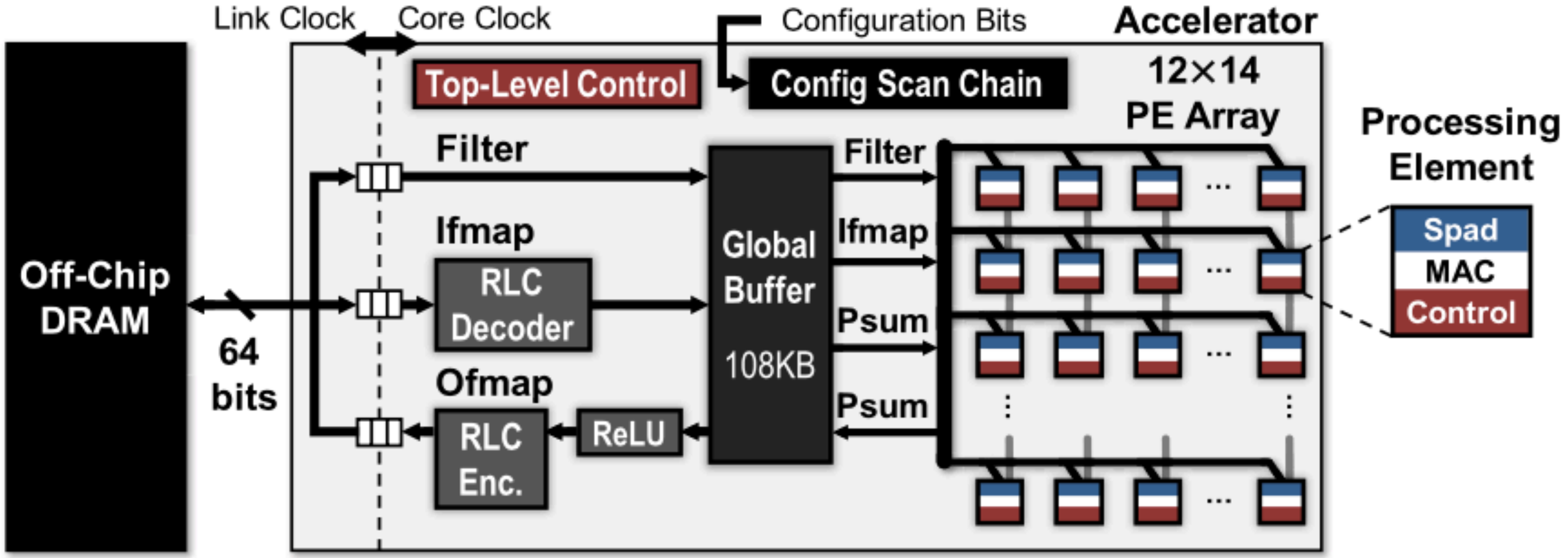
**Input:** 0, 0, 12, 0, 0, 0, 0, 53, 0, 0, 22, …

*Run Level Run Level Run Level Term*

**Output (64b):**



# Eyeriss system architecture



# Logical to Physical Mappings