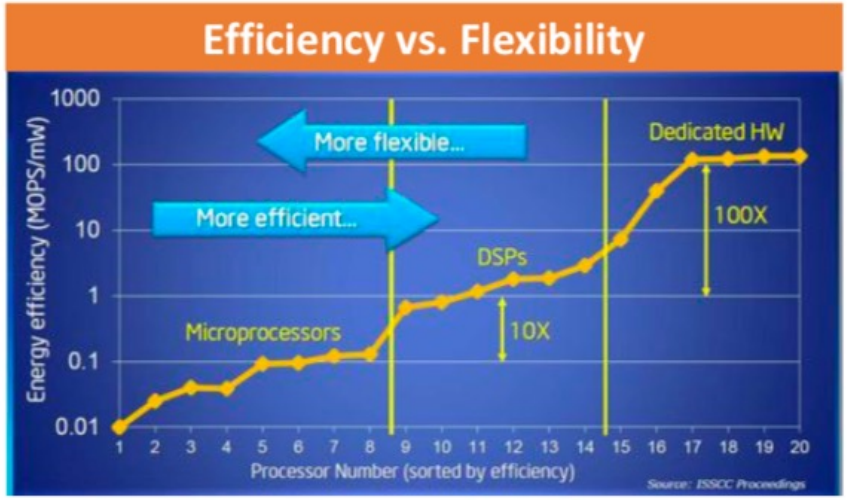
# ReRAM-Based Processing-In-Memory(PIM) Accelerators for AI Applications

Siling Yang

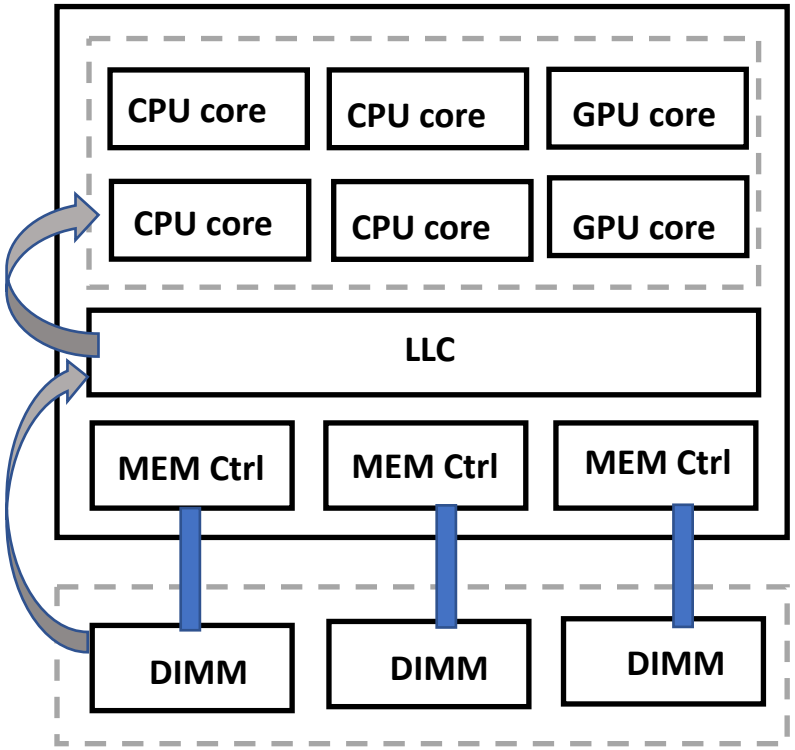
2021.04.17

# Existing hardware accelerators for deep learning

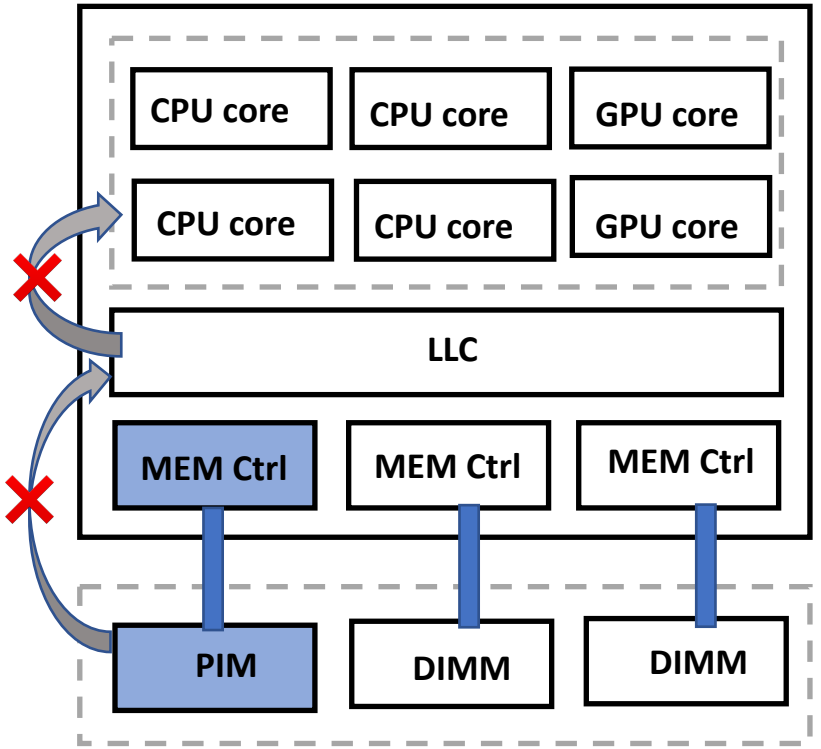
- **GPUs**
  - Fast, but high power consumption (~200W)
  - Training DNNs in back-end GPU clusters
- **FPGAs**
  - Massively parallel + low-power (~25W) + reconfigurable
  - Suitable for latency-sensitive real-time inference job
- **ASICs**
  - Fast + energy efficient
  - Long development cycle
- **Novel architectures and emerging devices**



# Why PIM?




| Operation | Energy(pJ) |
|---|---|
| 16b Add | 0.05 |
| 32b Add | 0.1 |
| 16b FP Add | 0.4 |
| 32b FP Add | 0.9 |
| 32b Mult | 3.1 |
| 16b FP Mult | 1.1 |
| 32b FP Mult | 3.7 |
| 32b SRAM Read(8KB) | 5 |
| 32b DRAM Read | 640 |




**Von Neumann architecture**

**Memory Wall**

**Processing in Memory**

# Why PIM benefits NN?

$$X_i \cdot W = X_o$$

**Input Feature Map:**

fetched from off-chip;
exhausted used through
exploring reuse pattern

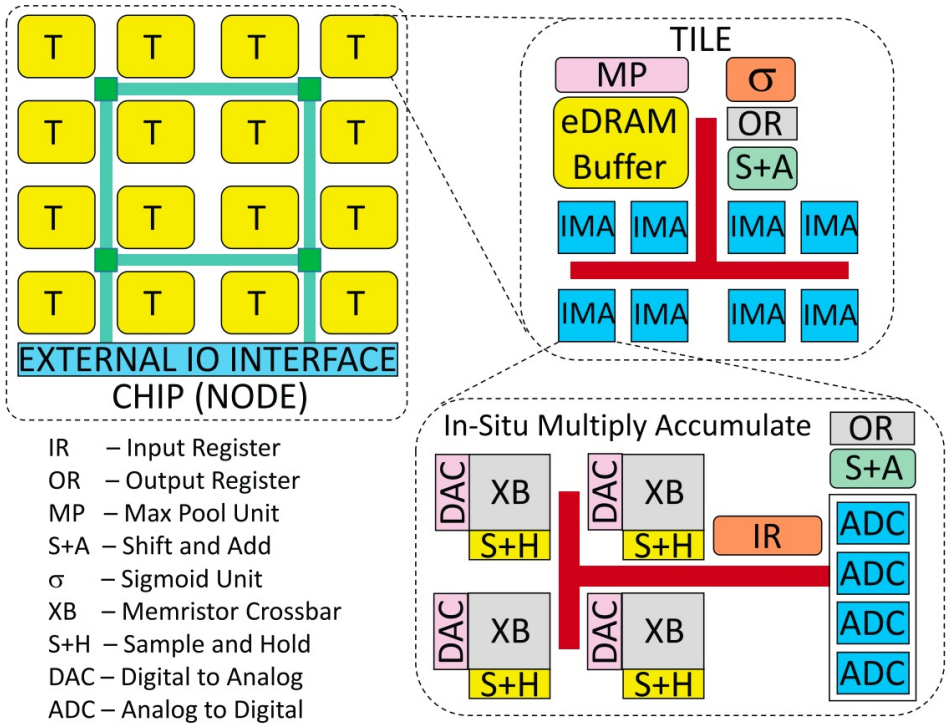
**Weight matrix:**

store in memory;
in-situ computed

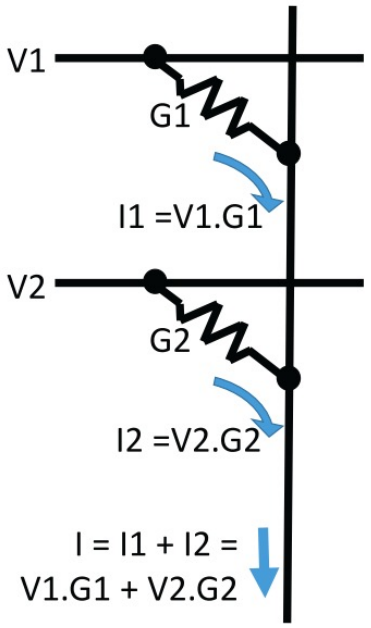
**Output Feature Map:**

Stored in on-chip buffer;
served as input of next layer

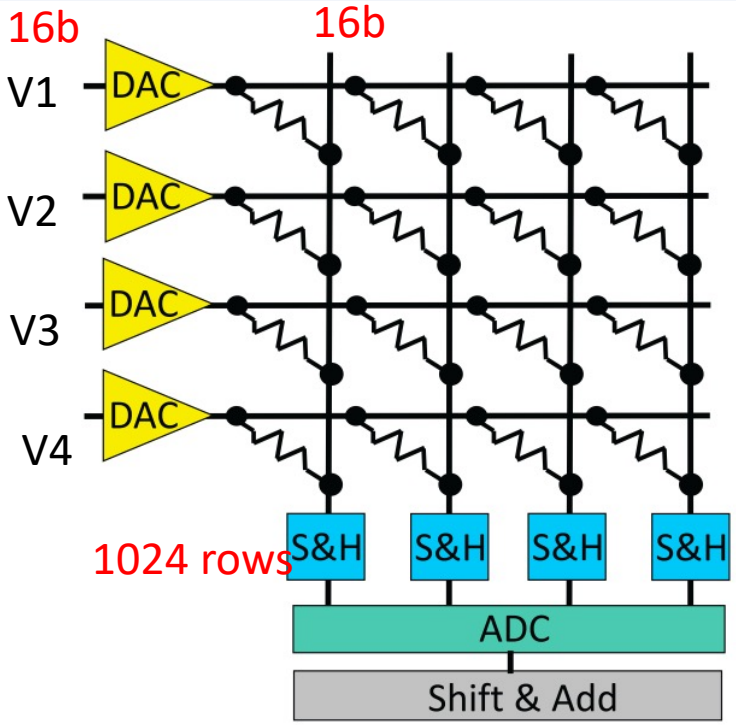
# Analog Dot Product in crossbars




ISAAC architecture hierarchy. [ISCA16]



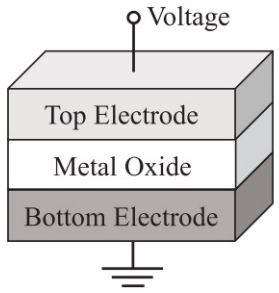

(a) Multiply-Accumulate operation

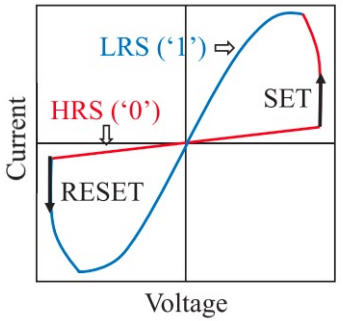


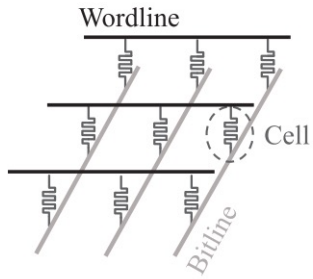

(b) Vector-Matrix Multiplier

Word line

Bit line



# Mapping Filter Weights of DNNs in crossbar-based Accelerators

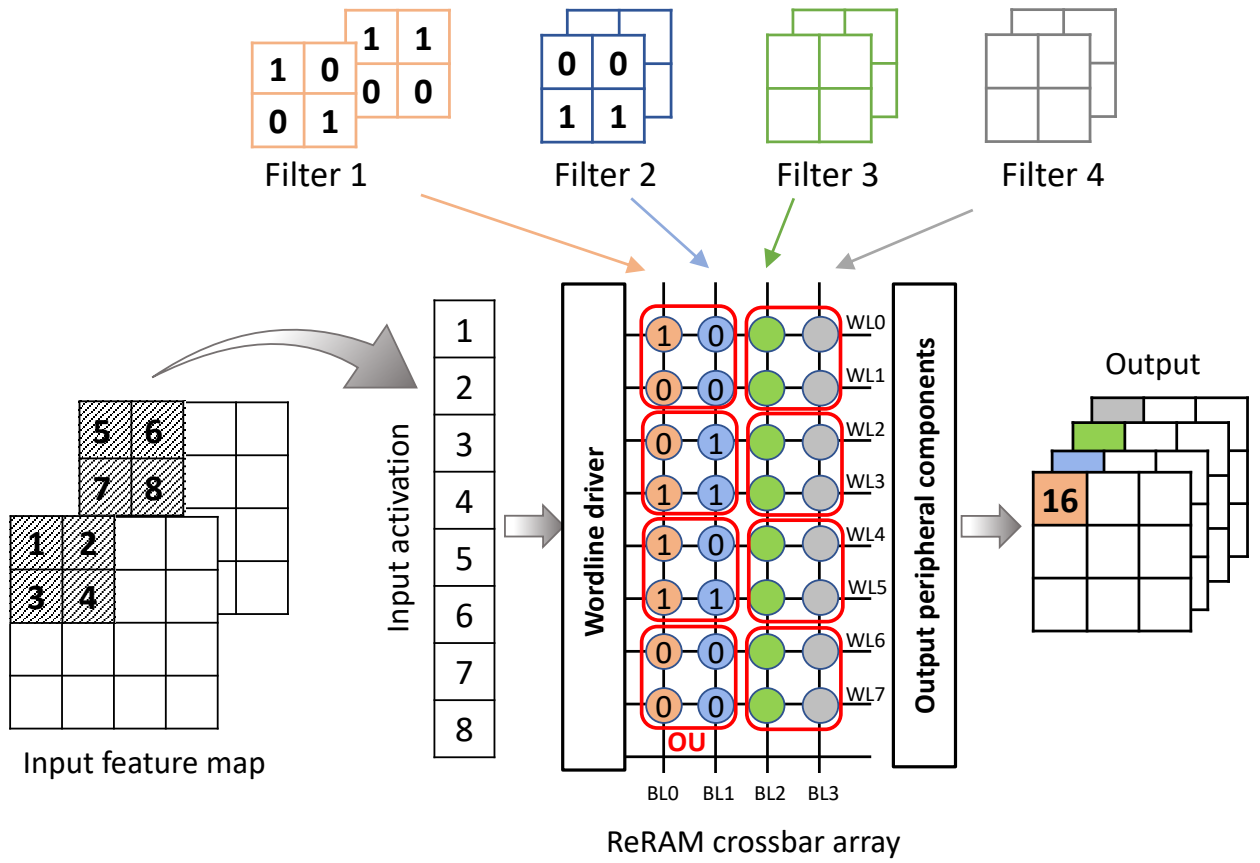


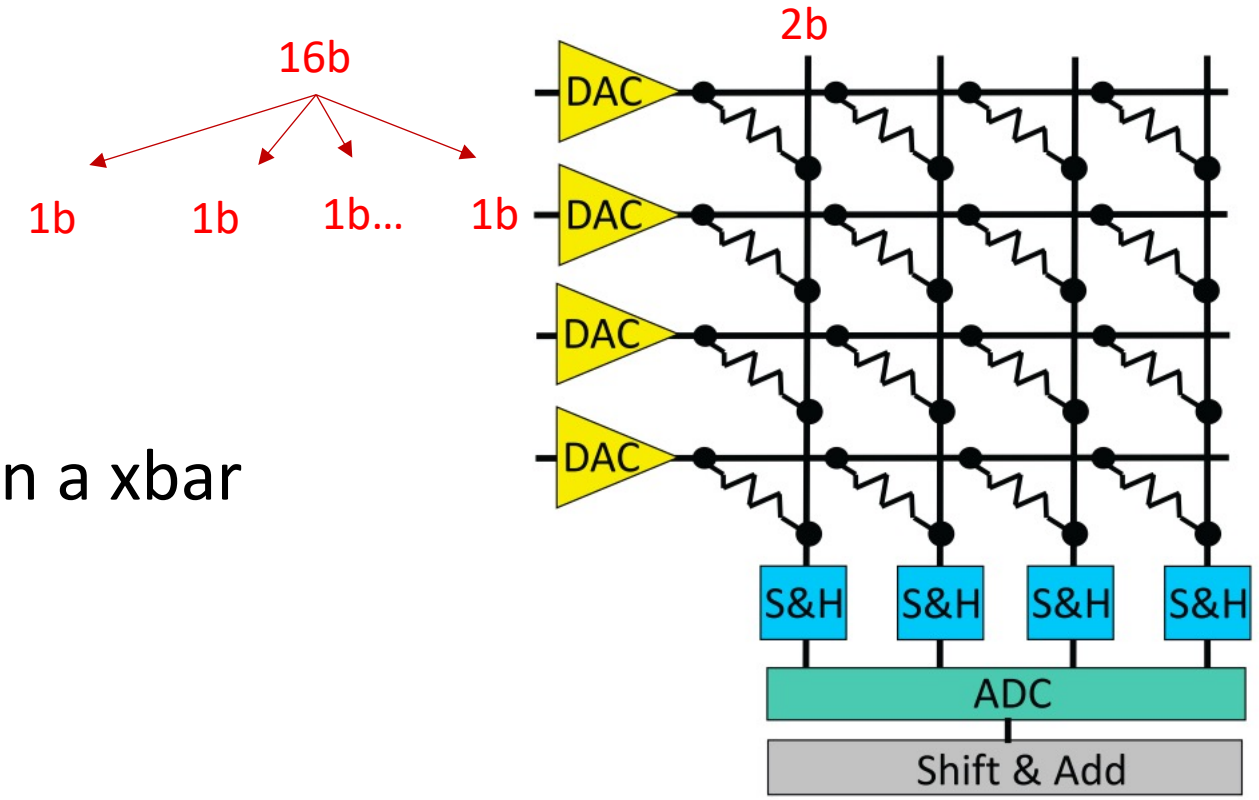

Illustration of mapping filter weights to a crossbar array used in the architecture of ReRAM-based accelerators. BL: bitline; WL: wordline; OU: operation unit.

# Design in ISAAC: simplifying the ADC

**[Never drop bits]**

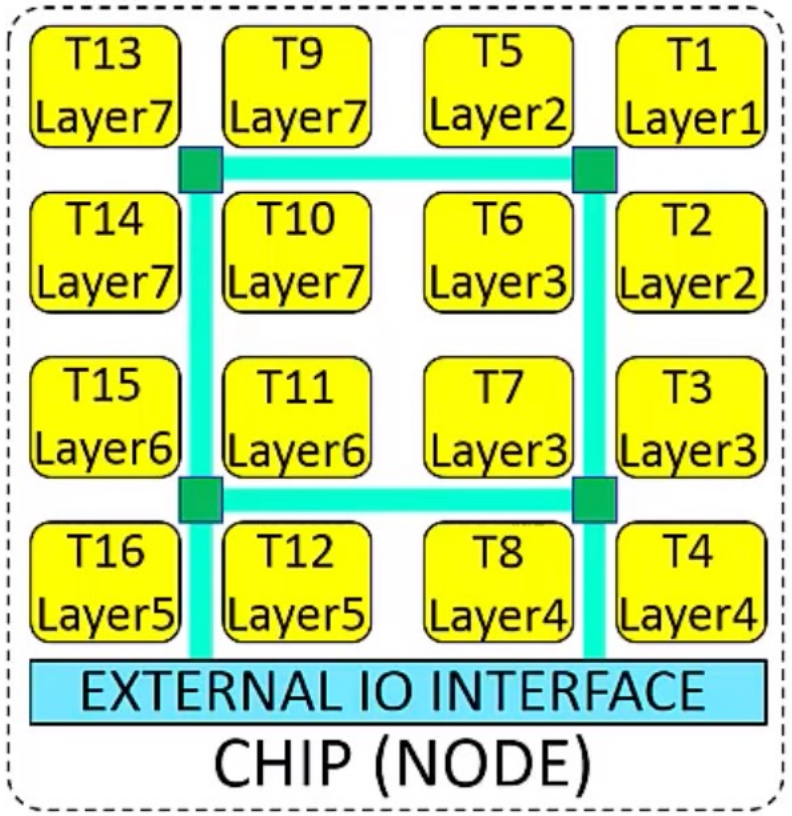
**Distribute the computation:**

- Across time
- Across multiple columns in a xbar
- Across multiple xbars
- Use an encoding trick

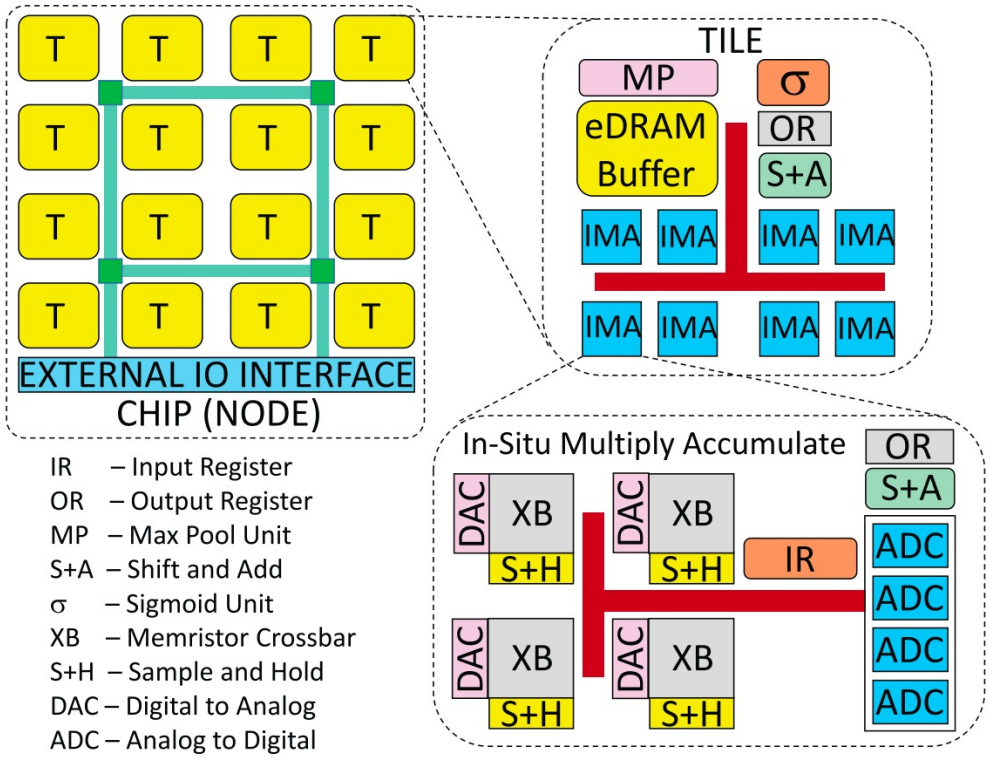


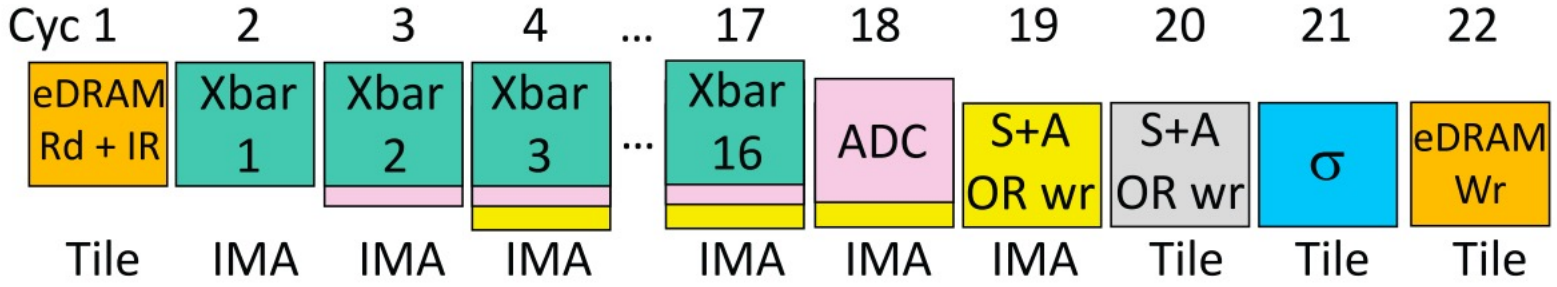

(b) Vector-Matrix Multiplier

# ISAAC


T13
Layer7
T9
Layer7
T5
Layer2
T1
Layer1
T14
Layer7
T10
Layer7
T6
Layer3
T2
Layer2
T15
Layer6
T11
Layer6
T7
Layer3
T3
Layer3
T16
Layer5
T12
Layer5
T8
Layer4
T4
Layer4
EXTERNAL IO INTERFACE
CHIP (NODE)

# ISAAC: Pipeline



# ISAAC Design space

Balanced use of xbars and ADCs;

Peak Thruput: 45 TOPs/s;

Storage: 63MB

**Optimize:**

- **CE: Computational Efficiency is represented by the num- ber of 16-bit operations performed per second per mm2 (GOPS/s $\times mm^2$).**
- **PE: Power Efficiency is represented by the number of 16-bit operations performed per watt (GOPS/W).**
- **SE: Storage Efficiency is the on-chip capacity for synaptic weights per unit area (MB/ $mm^2$).**

EIE(ISCA'16):

- Peak Thruput: 3TOPs/s
- Storage: 80MB (8MB)

Many large xbars; few ADCs;

Peak Thruput: 10 TOPs/s;

Storage: 1.7 GB

# Research on ReRAM-based PIM

**ReRAM-based ANN Architectures:**

- Mapping NN to ReRAM
- Architectures for NN inference
- Architectures for NN training
- MCA-aware pruning strategy
- Reconfigurable architectures
- Reducing overhead of analog implementation

**ReRAM-based PIM Techniques**

- Arithmetic and logical operations
- Data search operations
- Graph-processing operations
- Approximate computing approaches

# Thanks