Advanced Technology Opportunities

MICRO Tutorial (2016)

Website: http://eyeriss.mit.edu/tutorial.html

Joel Emer, Vivienne Sze, Yu-Hsin Chen

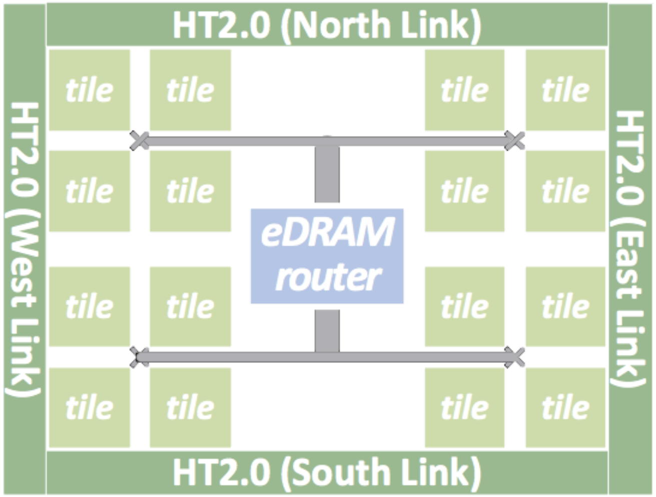
Advanced Storage Technology

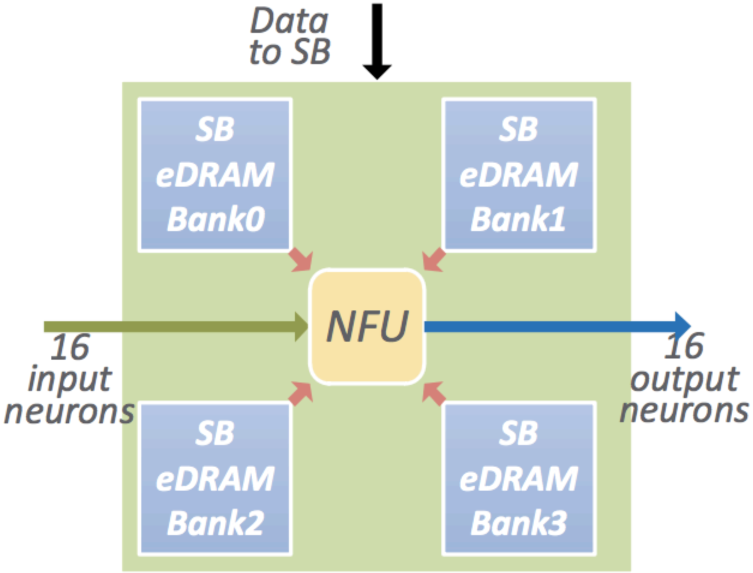
- Embedded DRAM (eDRAM)
  - Increase on-chip storage capacity
- 3D Stacked DRAM
  - e.g. Hybrid Memory Cube Memory (HMC), High Bandwidth Memory (HBM)
  - Increase memory bandwidth

eDRAM (DaDianNao)

- Advantages of eDRAM
  - 2.85x higher density than SRAM
  - 321x more energy-efficient than DRAM (DDR3)
- Store weights in eDRAM (36MB)
  - Target fully connected layers since dominated by weights

16 Parallel Tiles

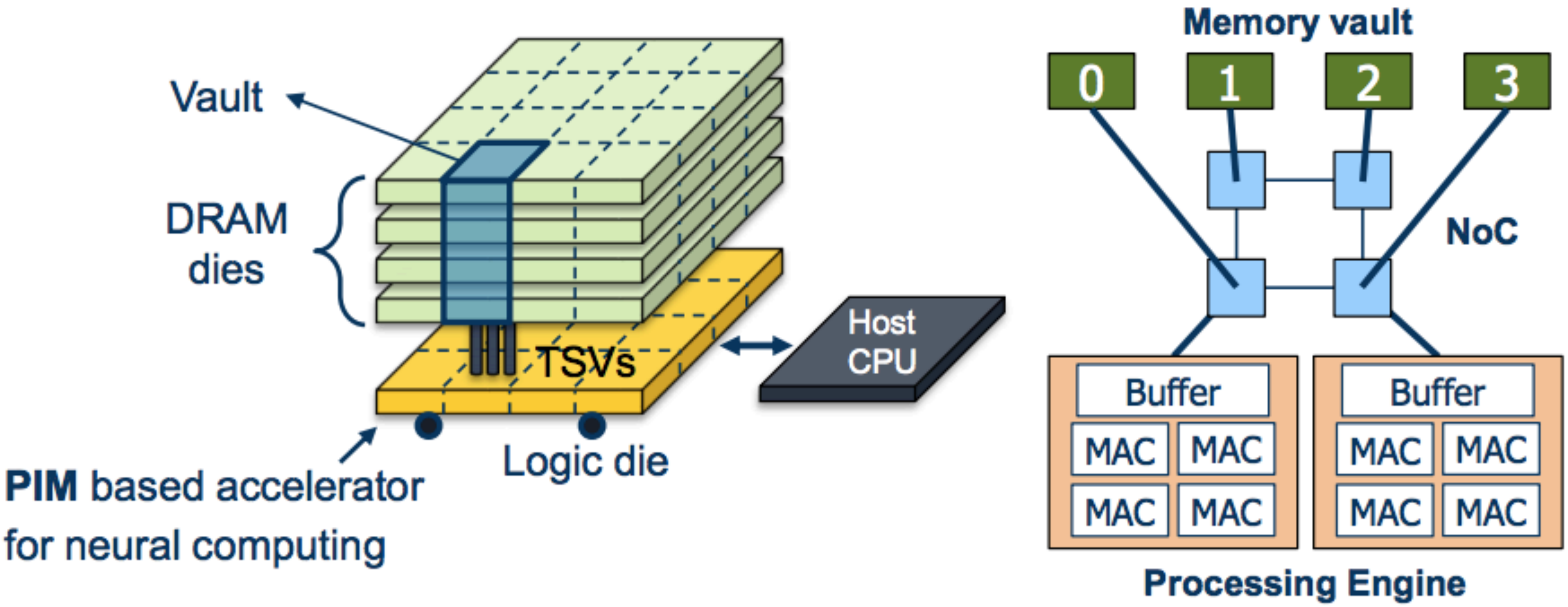




[Chen et al., DaDianNao, MICRO 2014]

Stacked DRAM (NeuroCube)

- NeuroCube on Hyper Memory Cube Logic Die
  - 6.25x higher BW than DDR3
    - HMC (16 ch x 10GB/s) > DDR3 BW (2 ch x 12.8GB/s)
  - Computation closer to memory (reduce energy)



[Kim et al., NeuroCube, ISCA 2016]

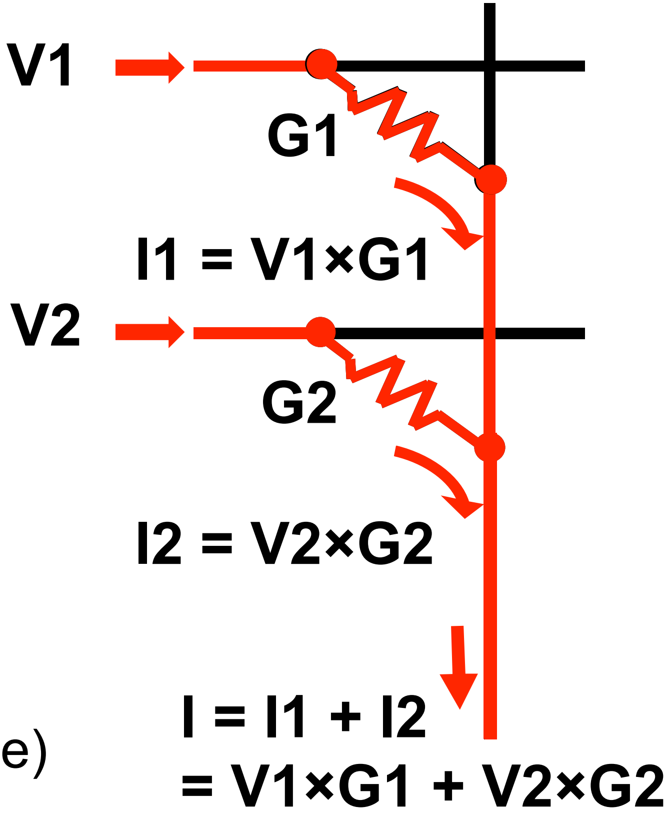
Analog Computation

- **Conductance = Weight**
- **Voltage = Input**
- **Current = Voltage × Conductance**
- **Sum currents for addition**

$$Output = \sum Weight \times Input$$

Input = V1, V2, …

Filter Weights = G1, G2, … (conductance)



Weight Stationary Dataflow

Figure Source: ISAAC, ISCA 2016

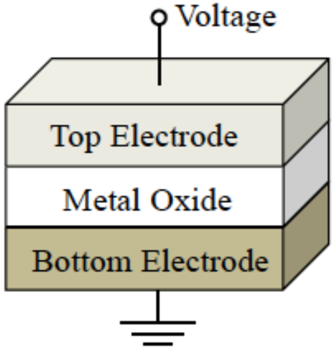
Memristor Computation

Use memristors as programmable weights (resistance)

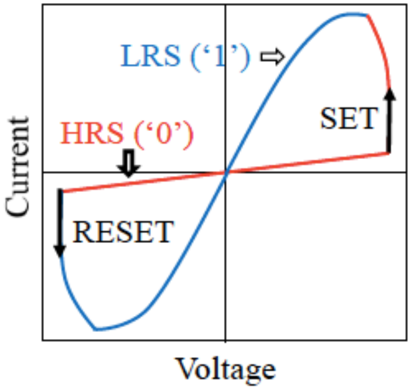
- Advantages
  - High Density (< 10nm x 10nm size*)
    - ~30x smaller than SRAM**
    - 1.5x smaller than DRAM**
  - Non-Volatile
  - Operates at low voltage
  - Computation within memory (in situ)
    - Reduce data movement

*[Govoreanu et al., IEDM 2011], **ITRS 2013

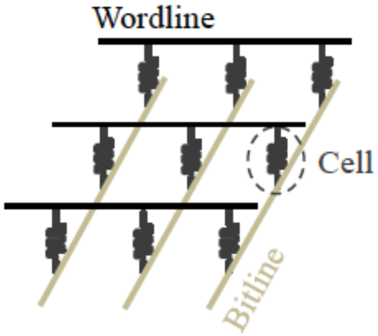
Memristor



(a) Conceptual view of a ReRAM cell

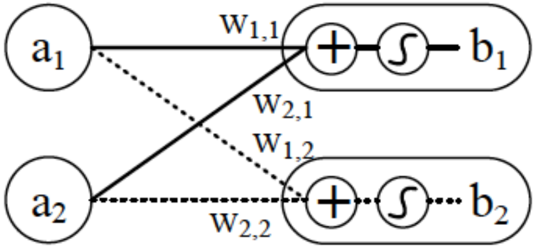


(b) I-V curve of bipolar switching

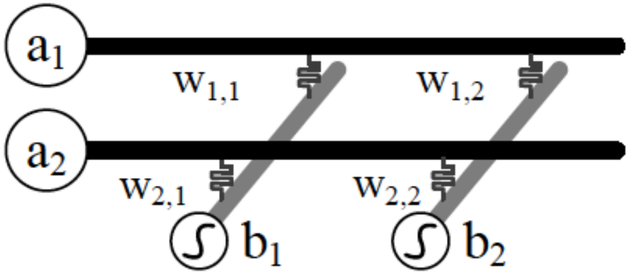


(c) schematic view of a crossbar architecture

$$b_j = \sigma\left(\sum_{\forall i} a_i \cdot w_{i,j}\right)$$



(a) An ANN with one input and one output layer

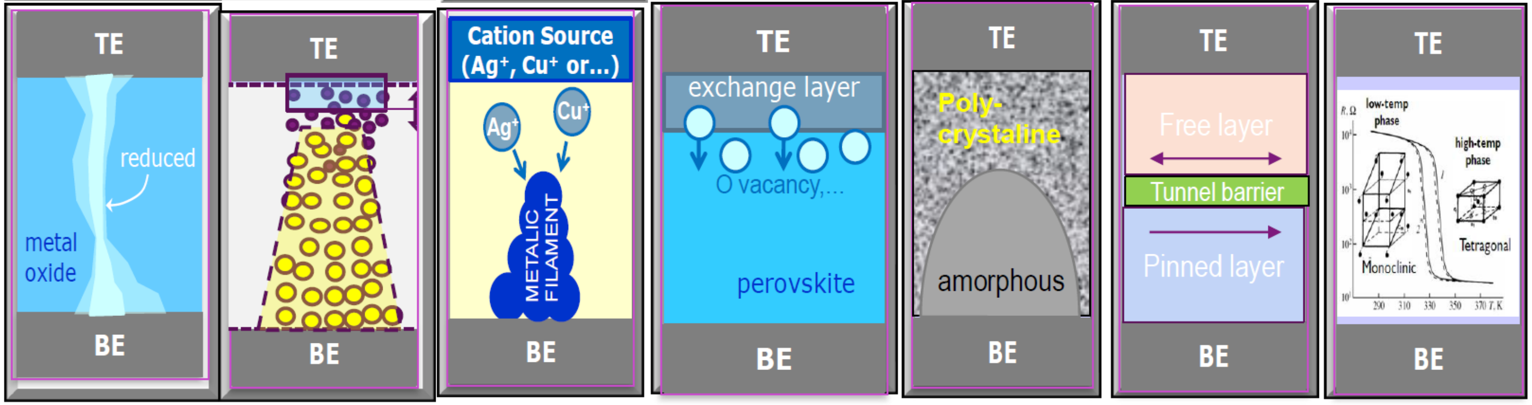


(b) using a ReRAM crossbar array for neural computation

[Chi et al., ISCA 2016]

Resistive Memory Devices

1D Filamentary			2D Interfacial	3D Bulk Transition		
Thermo-Chemical Fuse/ antifuse	Oxygen vacancy migration	Electro-chemical	Schottky barrier	Phase change	Tunnel Magneto resistance	Electronic MIT (Mott)



UNIPOLAR	BIPOLAR	BIPOLAR	BIPOLAR	UNIPOLAR	BIPOLAR	UNIPOLAR
RRAM				PCRAM	MRAM	?

Figure Source: Han Wang, USC

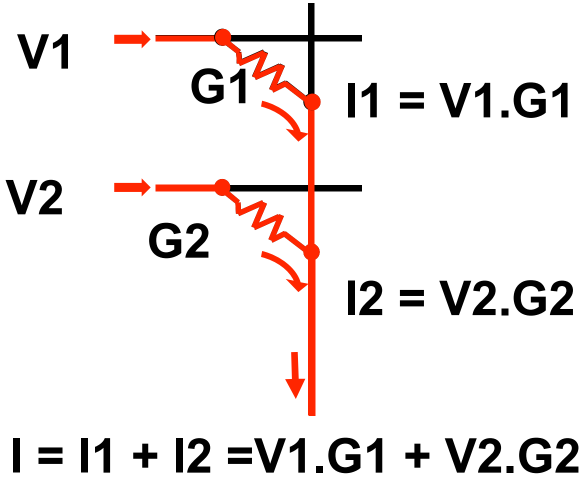
Challenges with Memristors

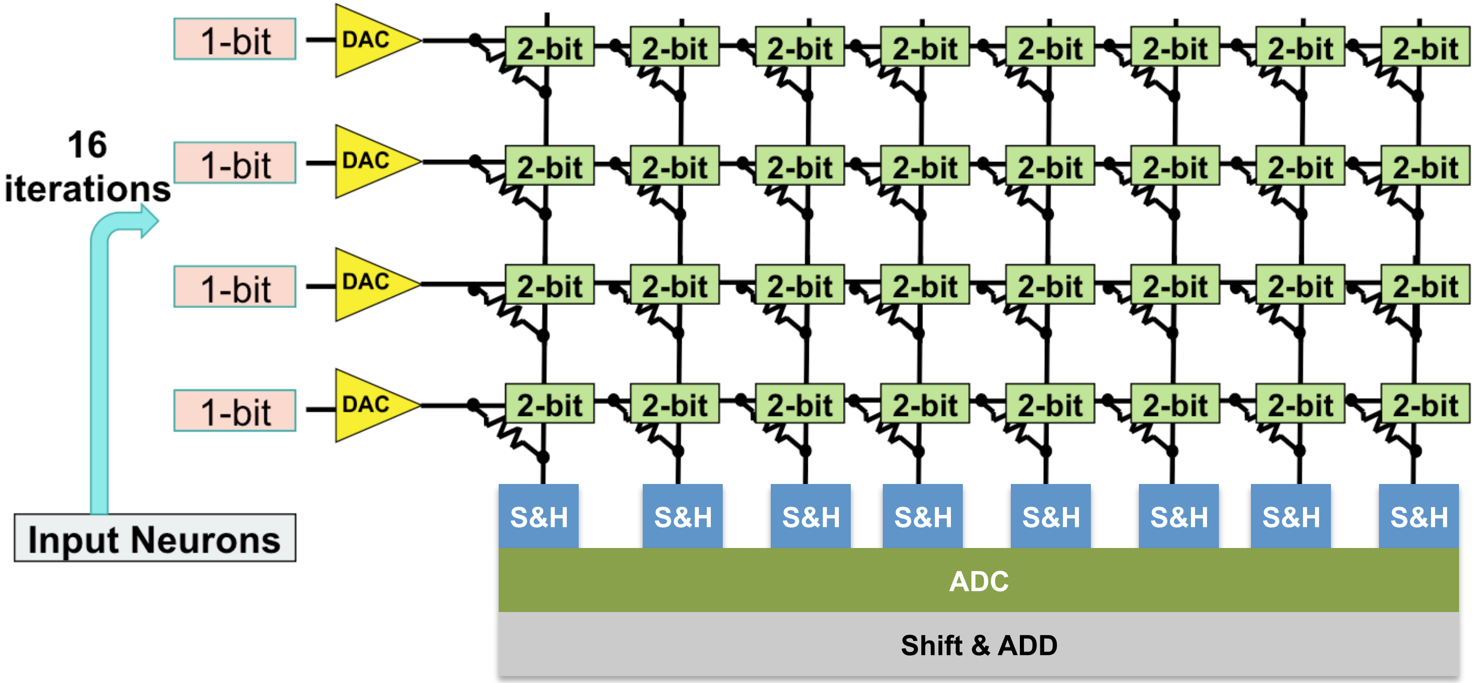
- Limited Precision
- A/D and D/A Conversion
- Array Size and Routing
  - Wire dominates energy for array size of 1k × 1k
  - IR drop along wire can degrade read accuracy
- Write/programming energy
  - Multiple pulses can be costly
- Variations & Yield
  - Device-to-device, cycle-to-cycle
  - Non-linear conductance across range

[Eryilmaz et al., ISQED 2016]

ISAAC

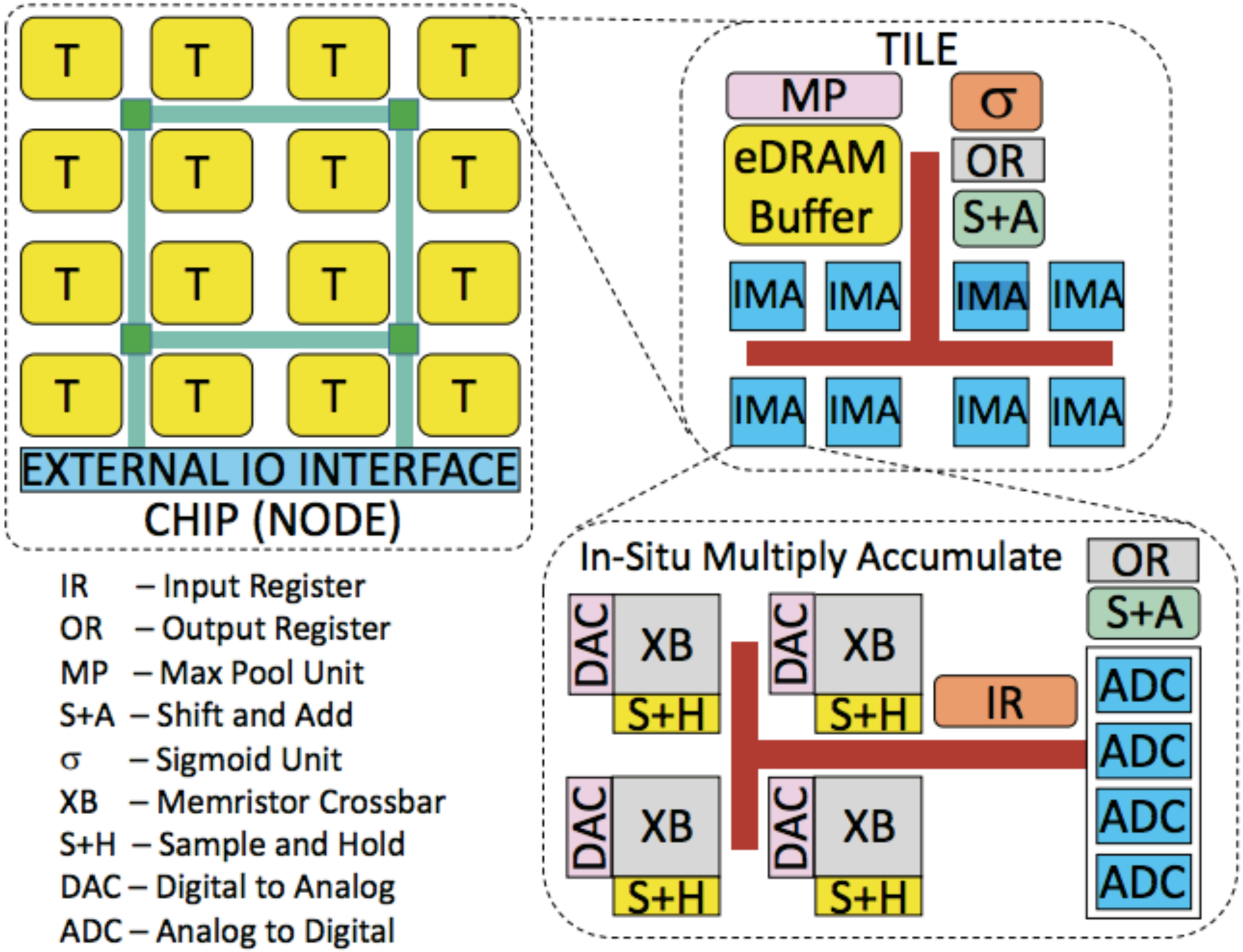
- eDRAM using memristors
- 16-bit dot-product operation
  - 8 x 2-bits per memristors
  - 1-bit per cycle computation





[Shafiee et al., ISCA 2016]

ISAAC



IR – Input Register
OR – Output Register
MP – Max Pool Unit
S+A – Shift and Add
σ – Sigmoid Unit
XB – Memristor Crossbar
S+H – Sample and Hold
DAC – Digital to Analog
ADC – Analog to Digital

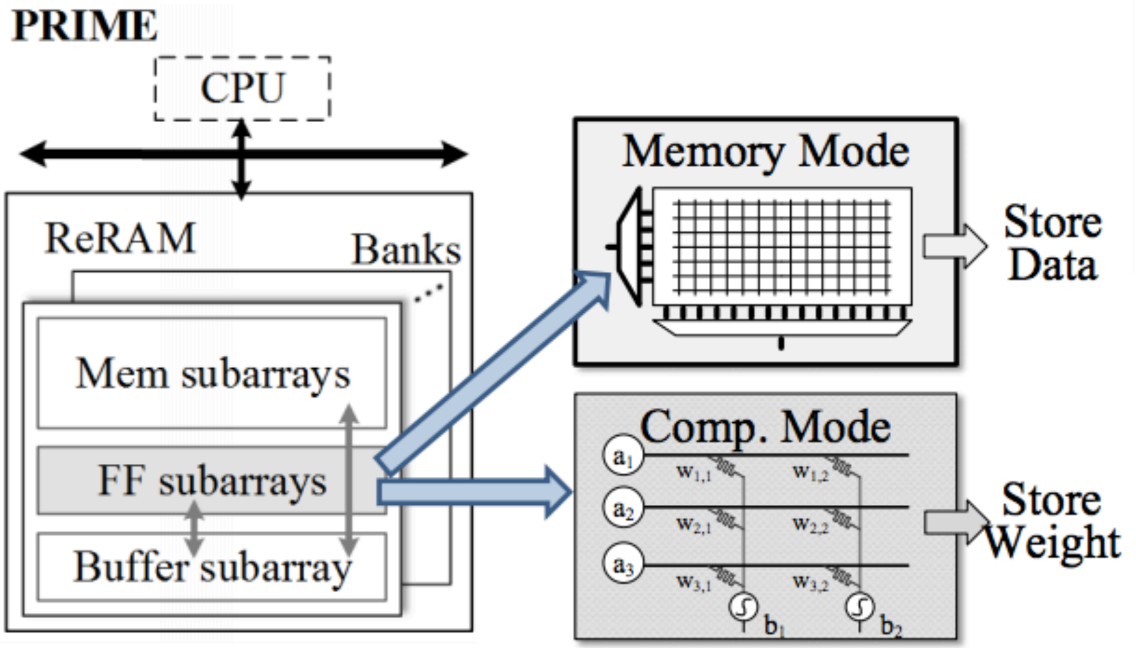
Eight 128x128 arrays per IMA

12 IMAs per Tile

[Shafiee et al., ISCA 2016]

PRIME

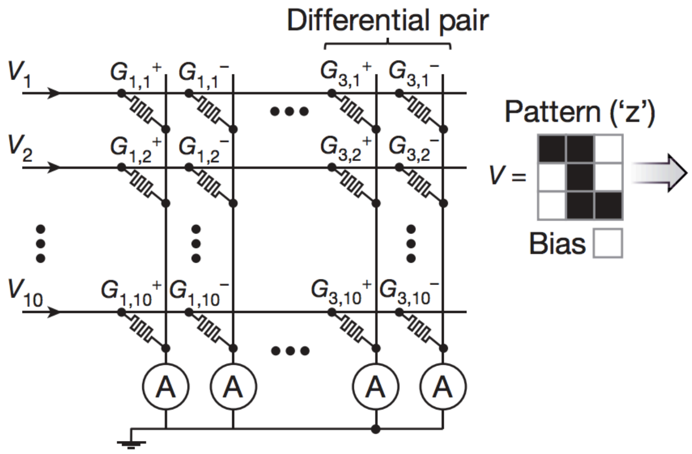
- Bit precision for each 256x256 ReRAM array
  - 3-bit input, 4-bit weight (2x for 6-bit input and 8-bit weight)
  - Dynamic fixed point (6-bit output)
- Reconfigurable to be main memory or accelerator
  - 4-bit MLC computation; 1-bit SLC for storage

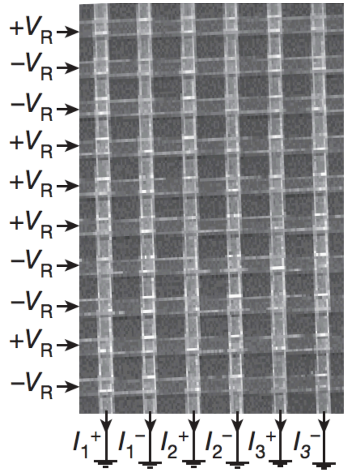


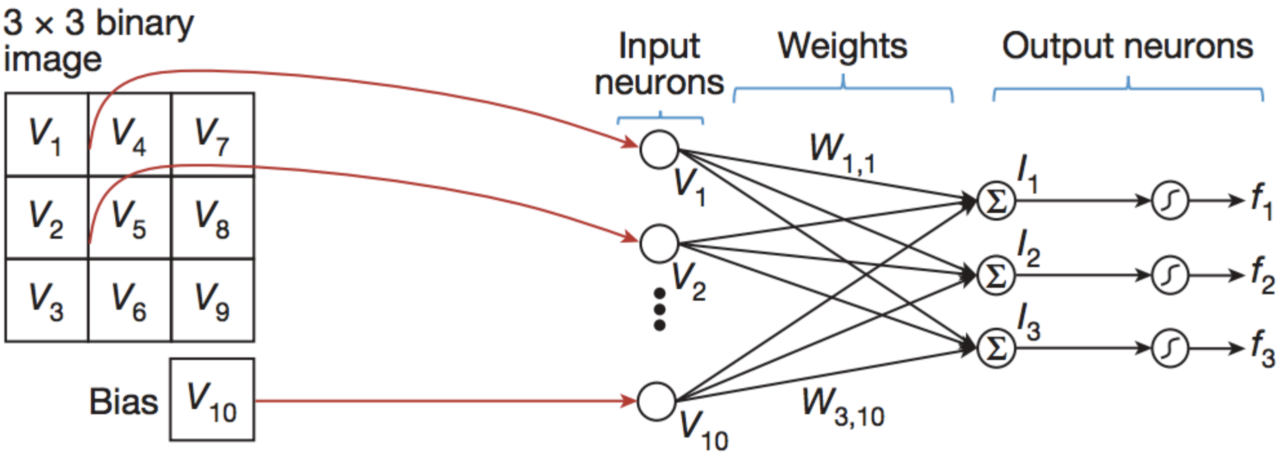
[Chi et al., ISCA 2016]

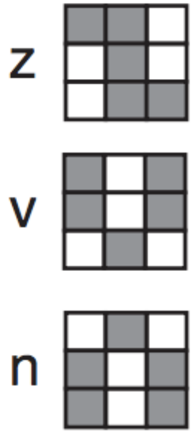
Fabricated Memristor Crossbar

- **Transistor-free metal-oxide 12x12 crossbar**
  - **A single-layer perceptron (linear classification)**
  - **3x3 binary image**
  - **10 inputs x 3 outputs x 2 differential weights = 60 memristors**


Differential pair
V_1
V_2
V_{10}
$G_{1,1}^+$ $G_{1,1}^-$ $G_{3,1}^+$ $G_{3,1}^-$
$G_{1,2}^+$ $G_{1,2}^-$ $G_{3,2}^+$ $G_{3,2}^-$
$G_{1,10}^+$ $G_{1,10}^-$ $G_{3,10}^+$ $G_{3,10}^-$
A
A
A
A
Pattern (‘z’)
V =
Bias







[Prezioso et al., Nature 2015]