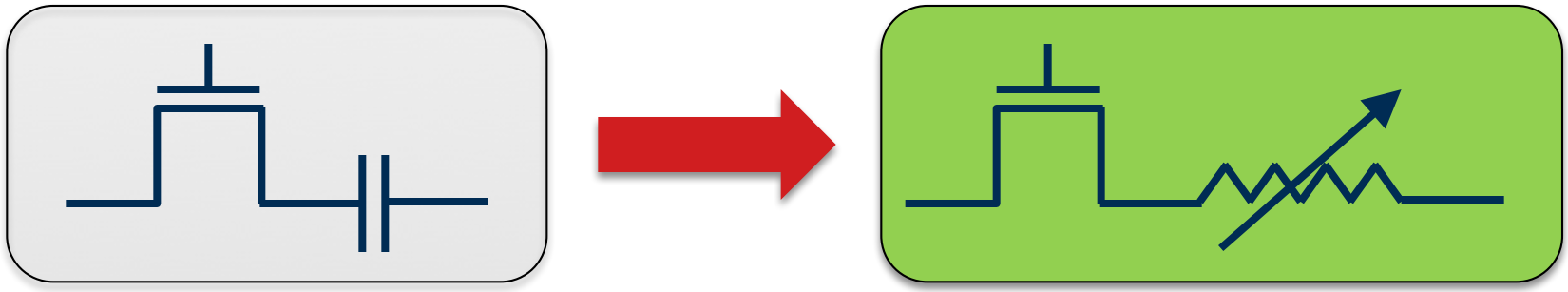
Emerging Non Volatile Memory

Resistive Memory Technologies

- Key concept: replace DRAM cell capacitor with a programmable resistor

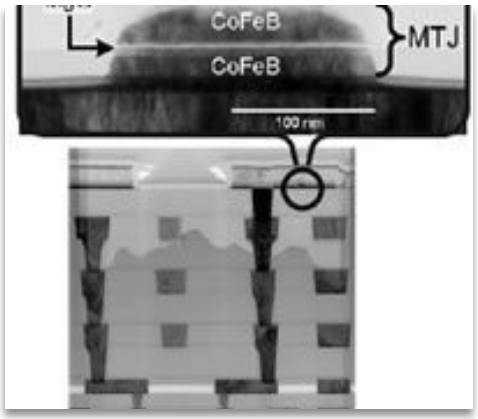


- 1T-1C DRAM
- Charge based sensing
- Volatile

- 1T-1R STT-MRAM, PCM, RRAM
- Resistance based sensing
- Non-volatile

Leading Contenders

STT-MRAM



[Halupka, et al. ISSCC'10]

- Limited to single-level cell
- 3D un-stackable
+ High endurance (~10^{15})
+ ~4ns switching time
+ ~50uW switching power

PCM-RAM

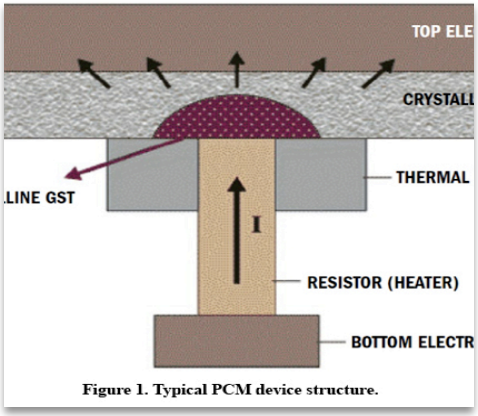
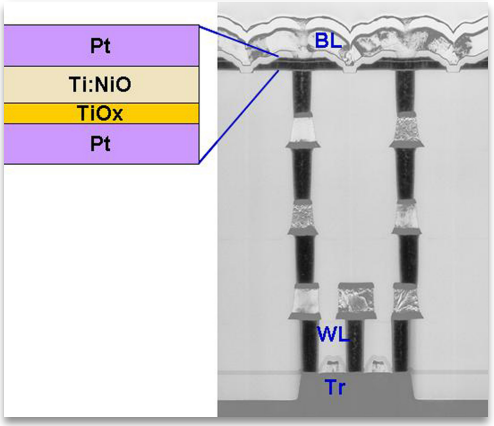


Figure 1. Typical PCM device structure.

[Pronin. EETime'13]

+ Multi-level cell capable
+ $4F^2$ 3D-stackable cell
- Endurance: ~10^9 writes
- ~100ns switching time
- ~300uW switching power

R-RAM

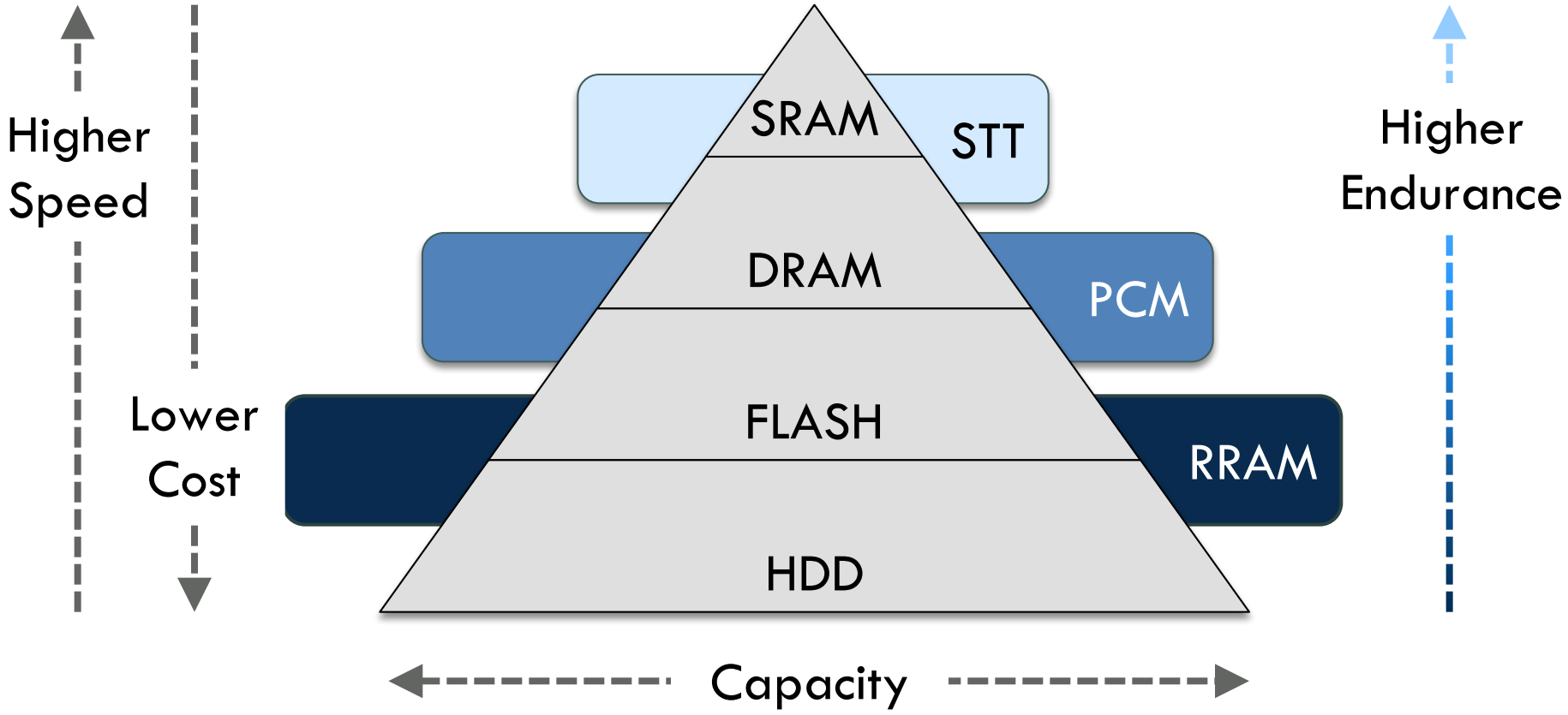


[Henderson. InfoTracks'11]

+ Multi-level cell capable
+ $4F^2$ 3D-stackable cell
- Endurance: 10^6~10^{12} writes
+ ~5ns switching time
+ ~50uW switching power

[ITRS'13]

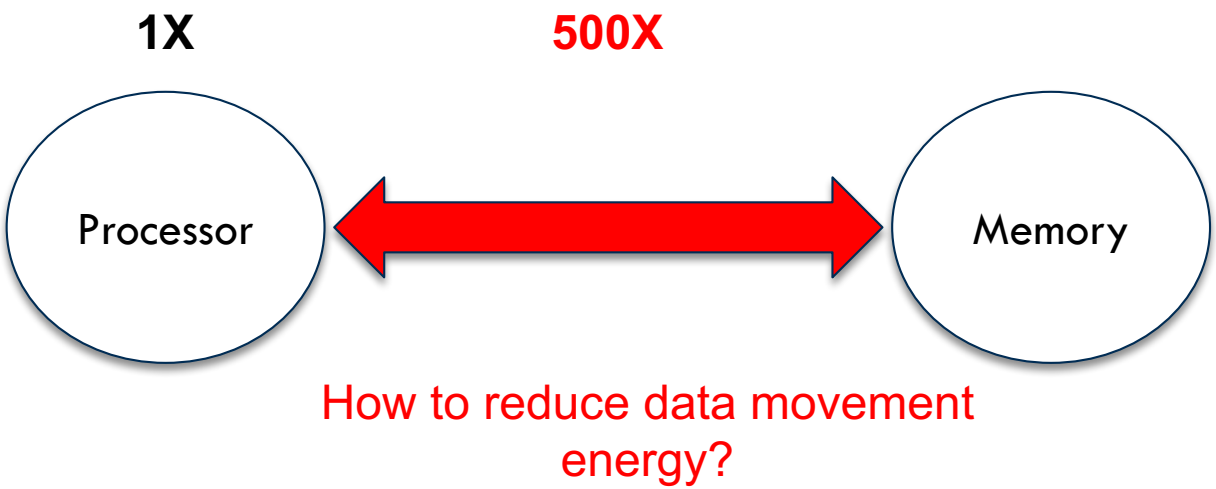
Positioning of Resistive Memories



In-Memory Processing

Example Research Question

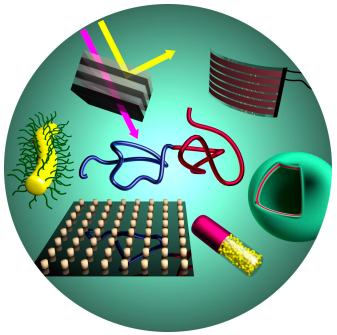
- Can we reduce the cost of data movement between memory and processor core?

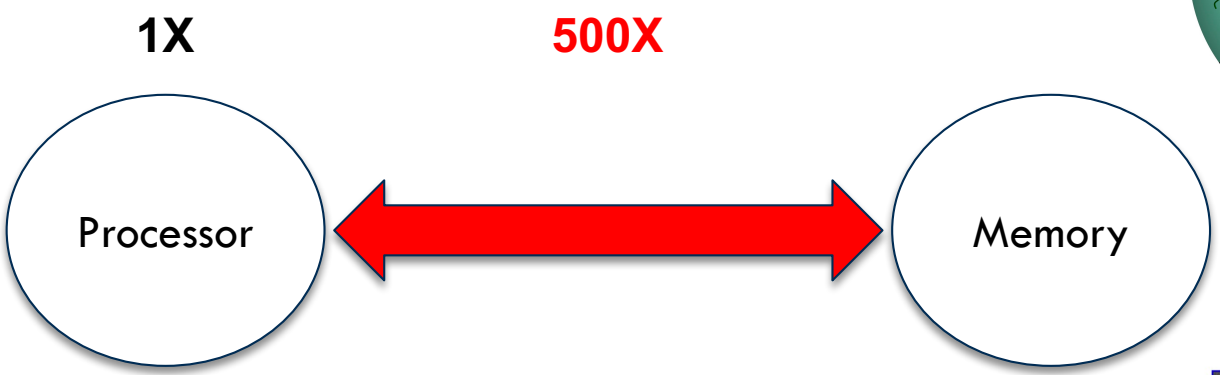


data sources: Nvidia

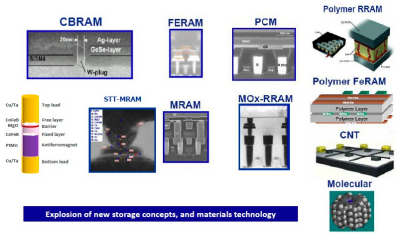
Example Research Question

- Can we reduce the cost of data movement between memory and processor core?





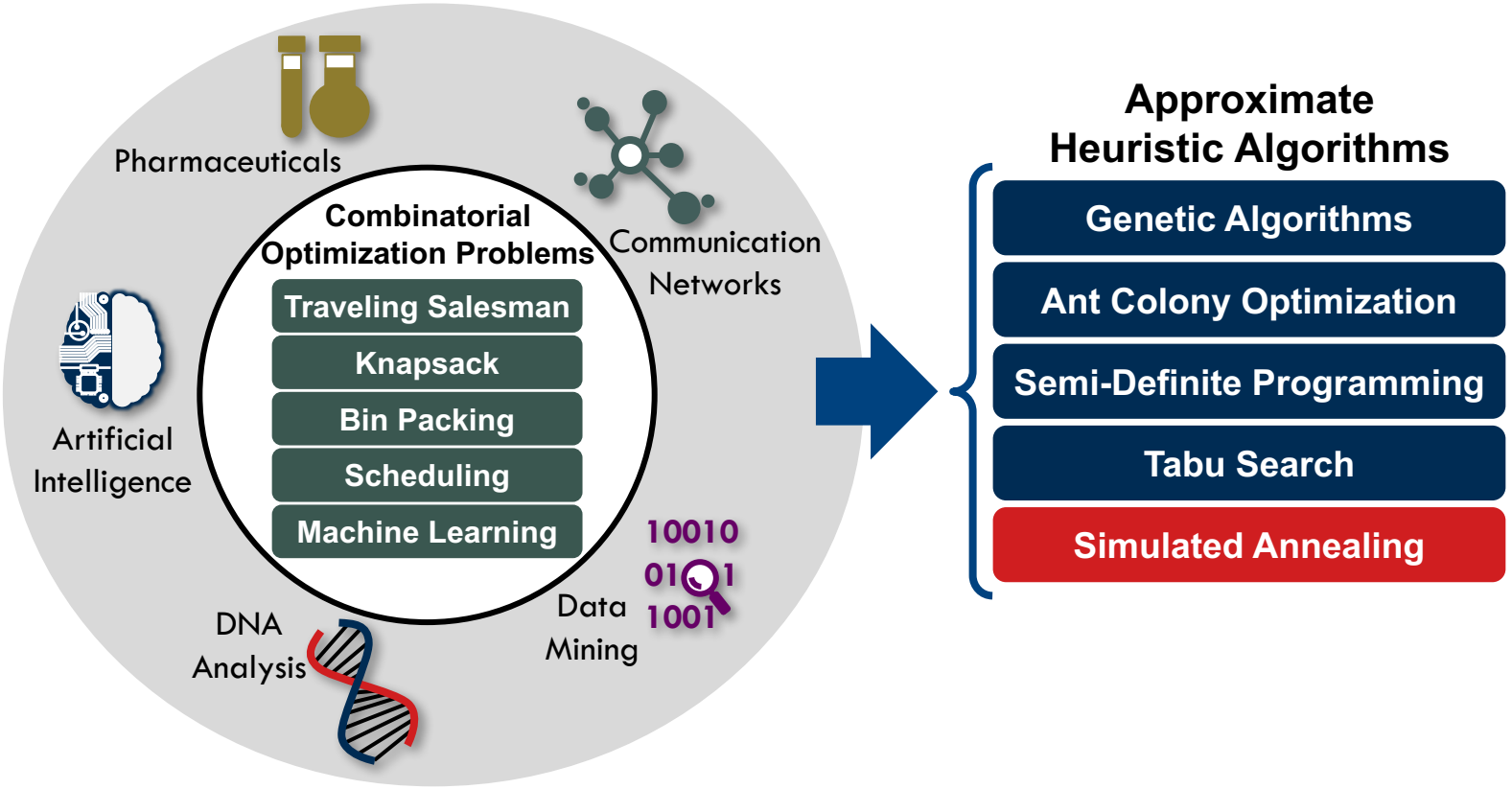
How to reduce data movement energy?



data sources: Nvidia

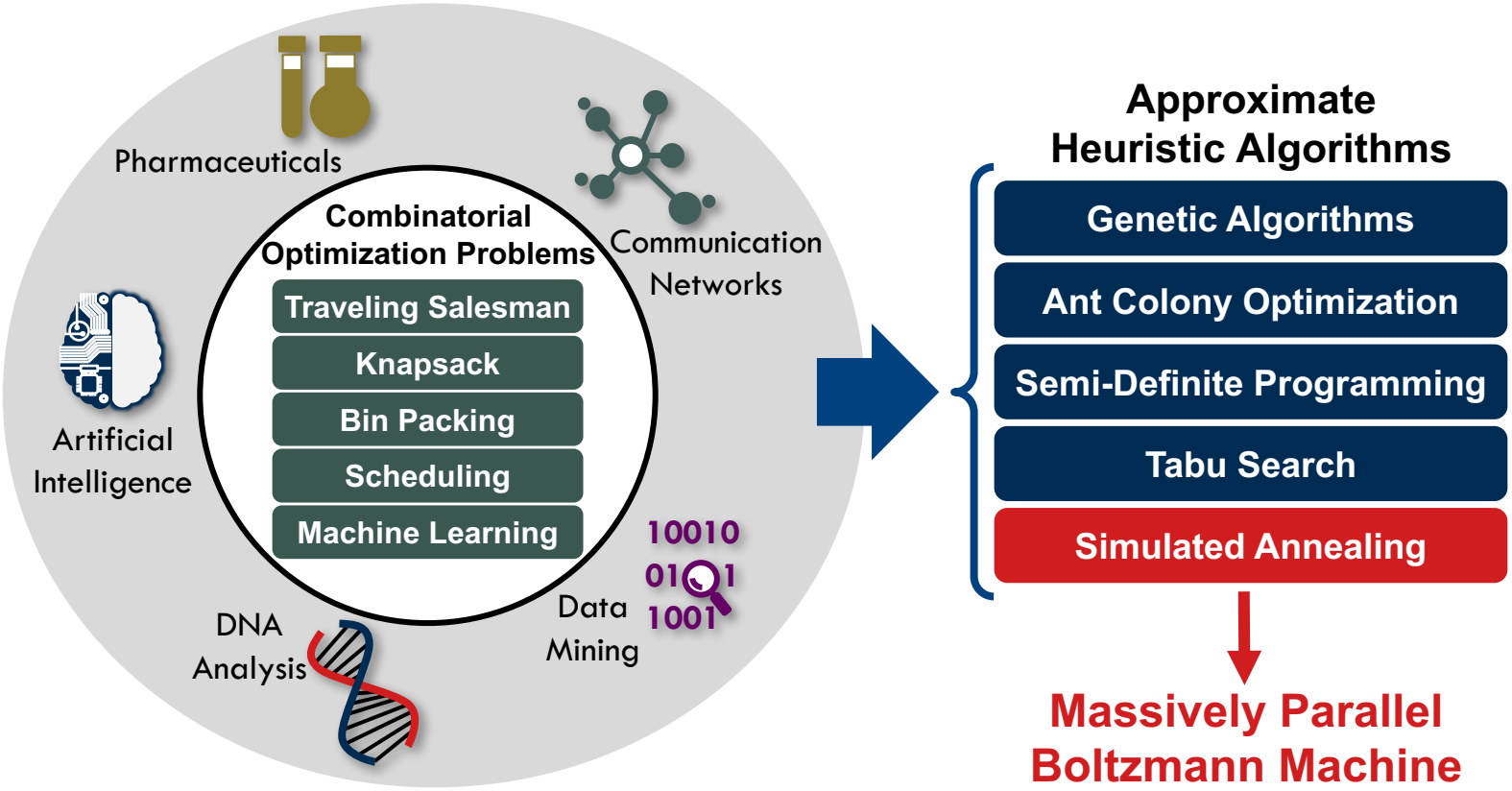
Combinatorial Optimization

- Numerous critical problems in science and engineering can be cast within the combinatorial optimization framework.



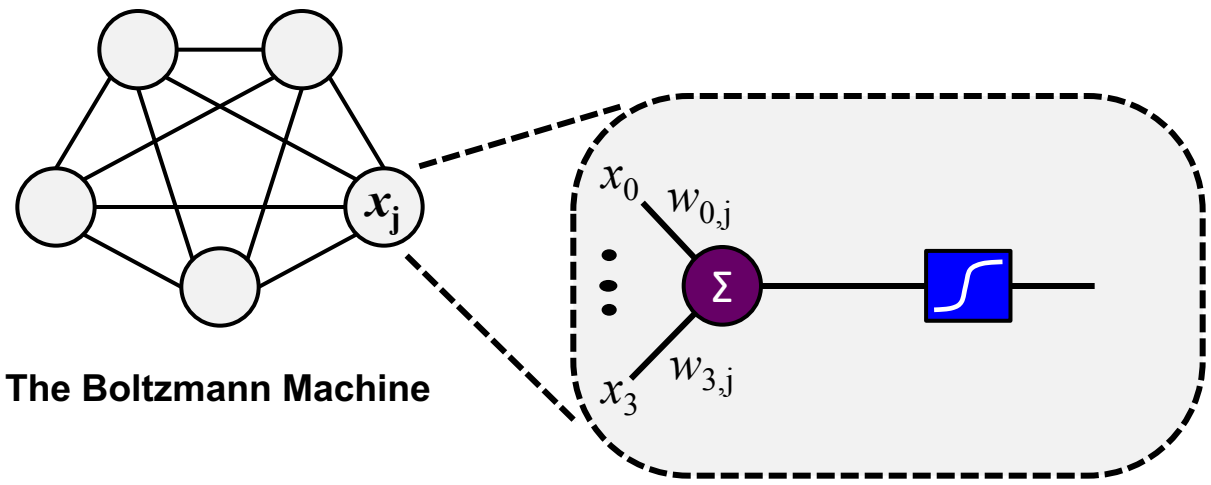
Combinatorial Optimization

- Numerous critical problems in science and engineering can be cast within the combinatorial optimization framework.



The Boltzmann Machine

- Two-state units connected with real-valued edge weights form a stochastic neural network.
- Goal: iteratively update the state or weight variables to minimize the network energy (E).

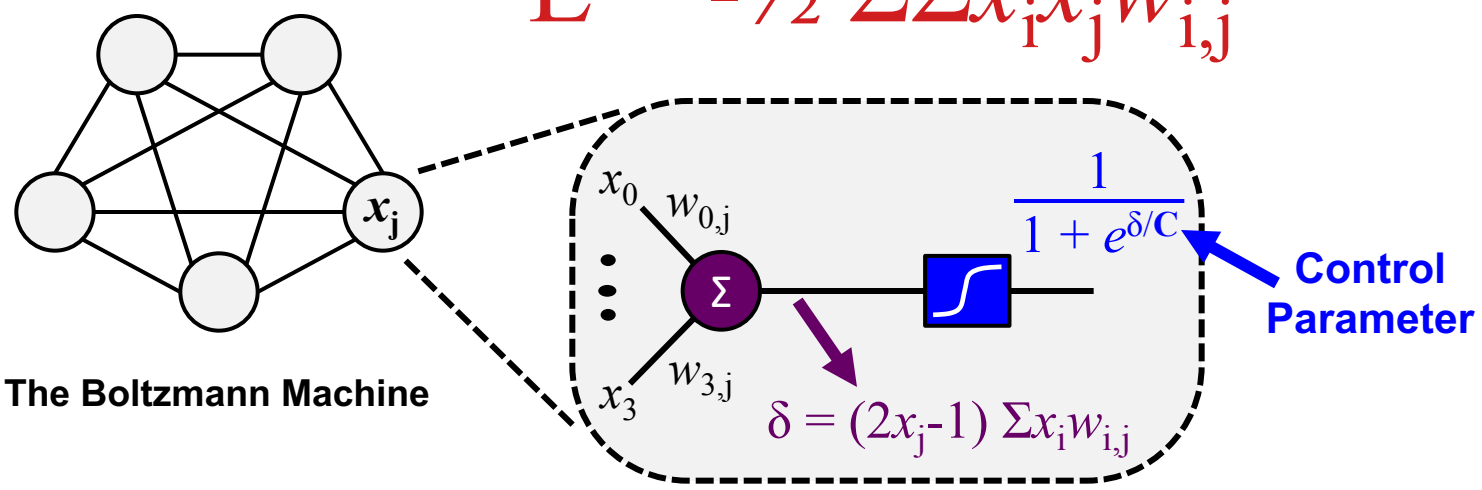


The Boltzmann Machine

The Boltzmann Machine

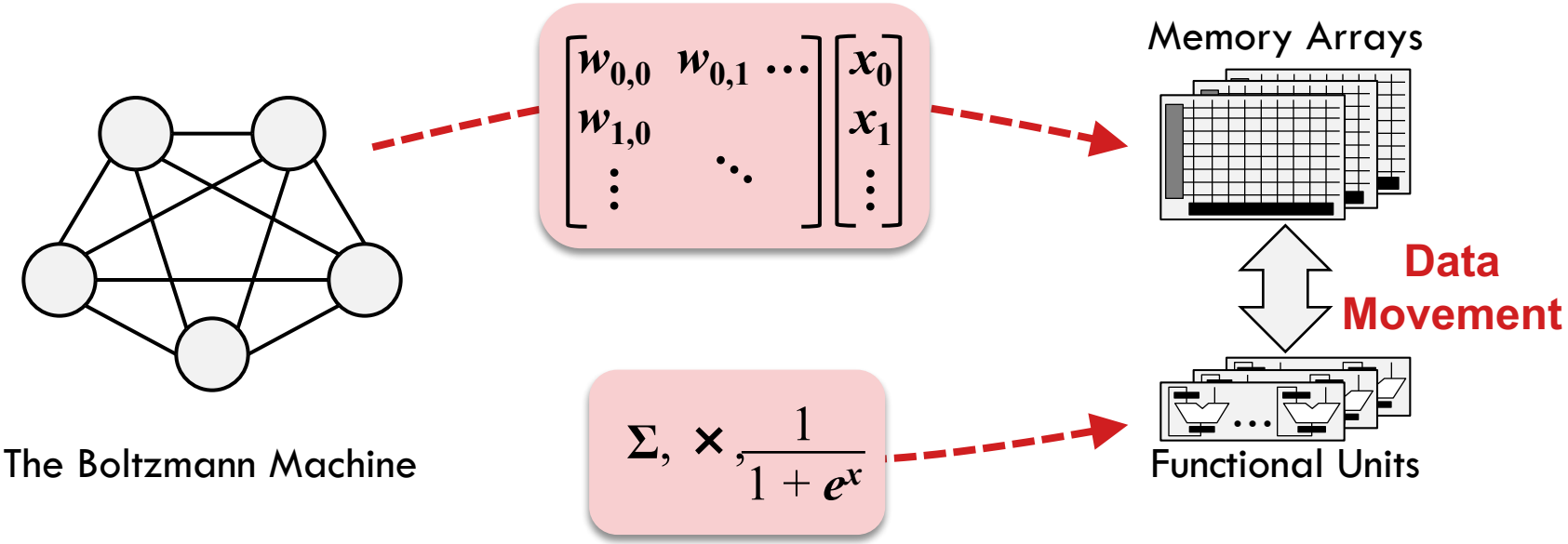
- Two-state units connected with real-valued edge weights form a stochastic neural network.
- Goal: iteratively update the state or weight variables to minimize the network energy (E).

$$E = -\tfrac{1}{2} \Sigma\Sigma x_i x_j w_{i,j}$$



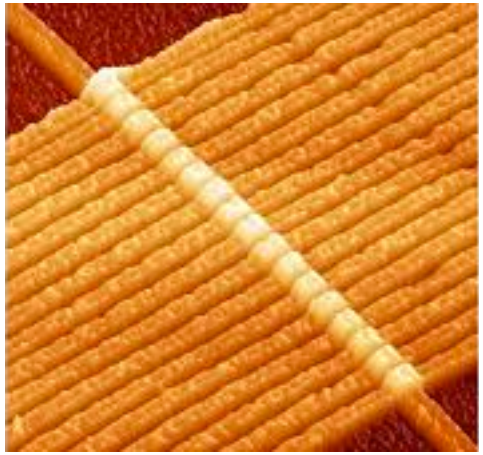
Computational Model

- Network energy is minimized by adjusting either the edge weights or recomputing the states.
- Iterative matrix-vector multiplication between weights and states is critical to finding minimal network energy.

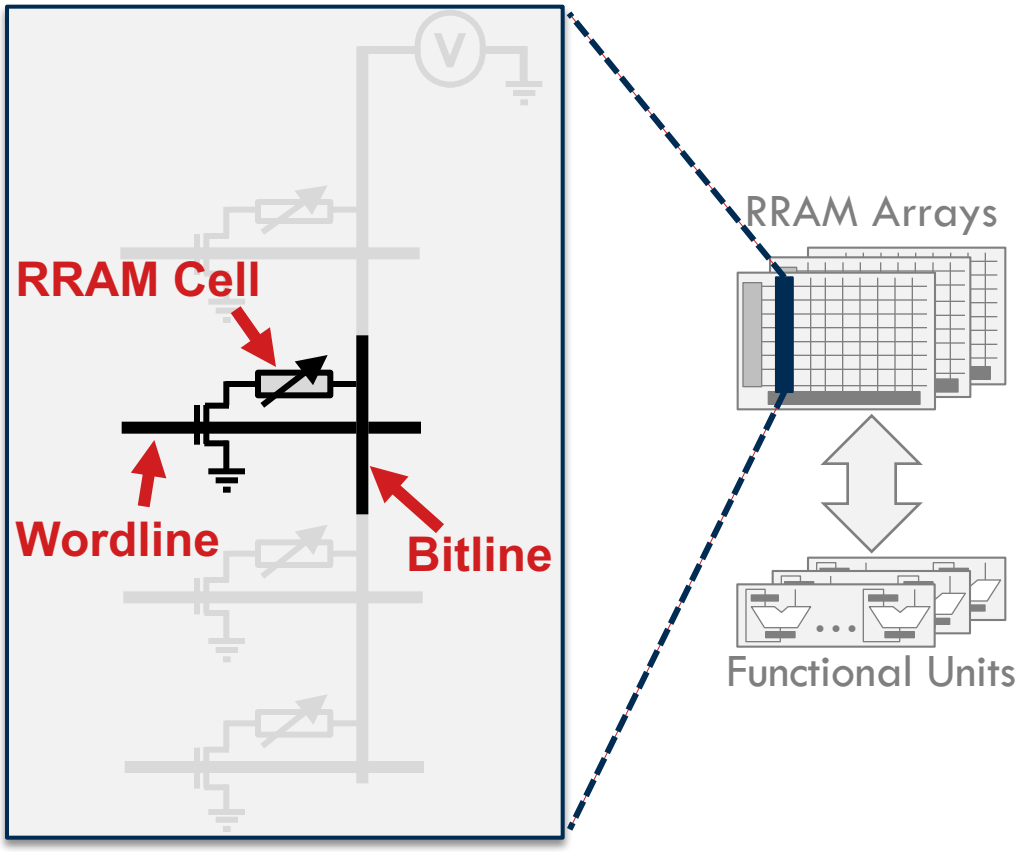


Resistive Random Access Memory

- An RRAM cell comprises an access transistor and a resistive switching medium.



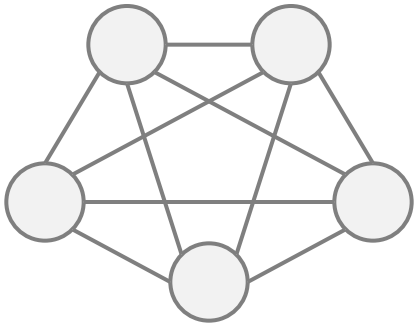
RRAM: Resistive RAM
(source: HP, 2009)



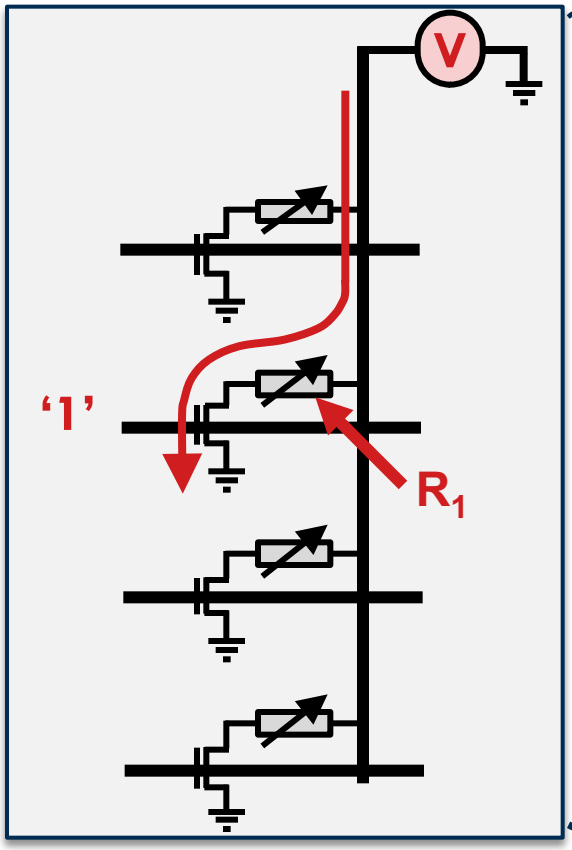
Resistive Random Access Memory

- A read is performed by activating a wordline and measuring the bitline current (I).

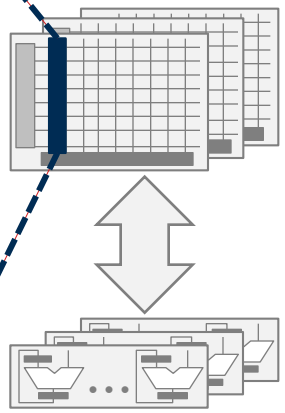
$I = V/R_1$



The Boltzmann Machine



RRAM Arrays

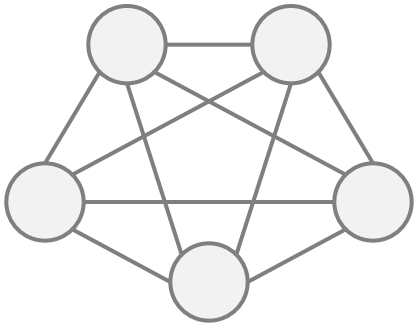


Functional Units

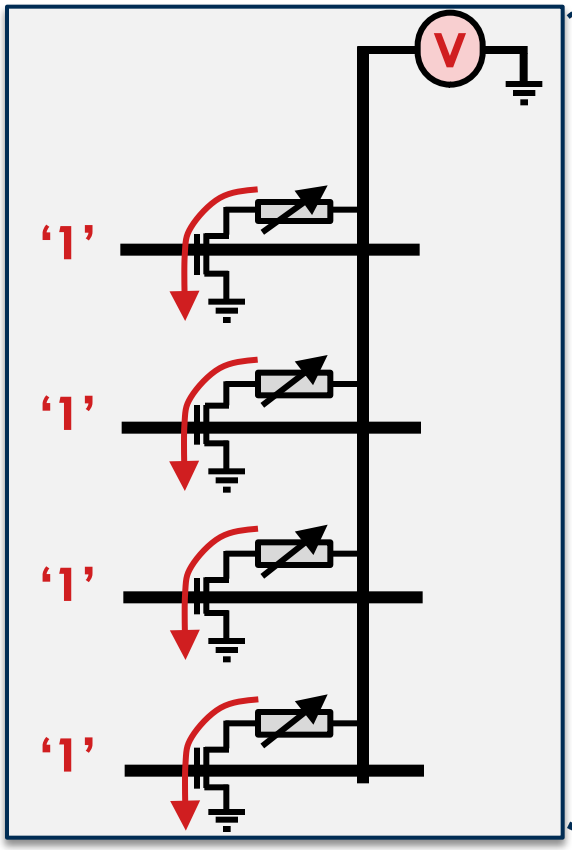
Memristive Boltzmann Machine

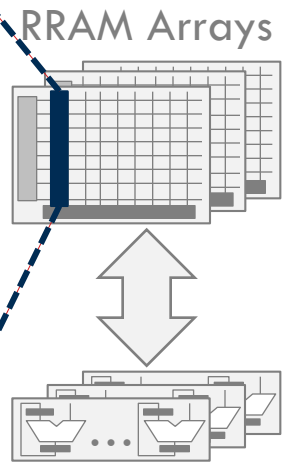
- Key Idea: exploit current summation on the RRAM bitlines to compute dot product.

$I = \Sigma V/R_i$



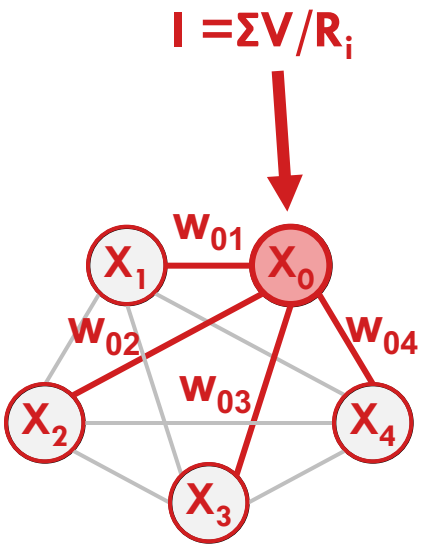
The Boltzmann Machine



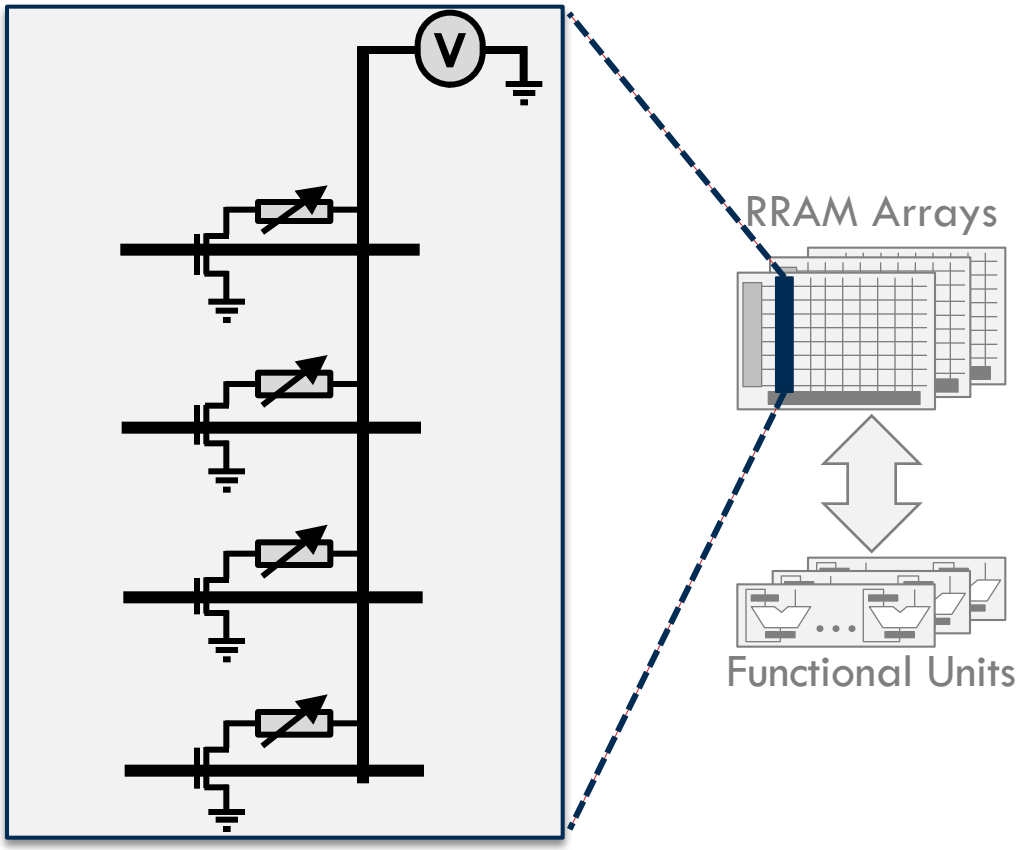


Memristive Boltzmann Machine

- Memory cells represent the weights and state variables are used to control the bitline and wordlines.

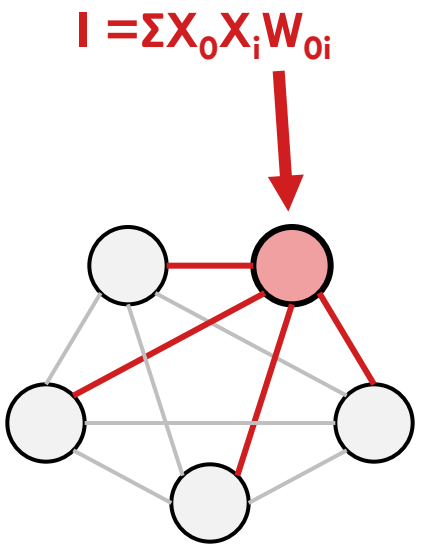


The Boltzmann Machine

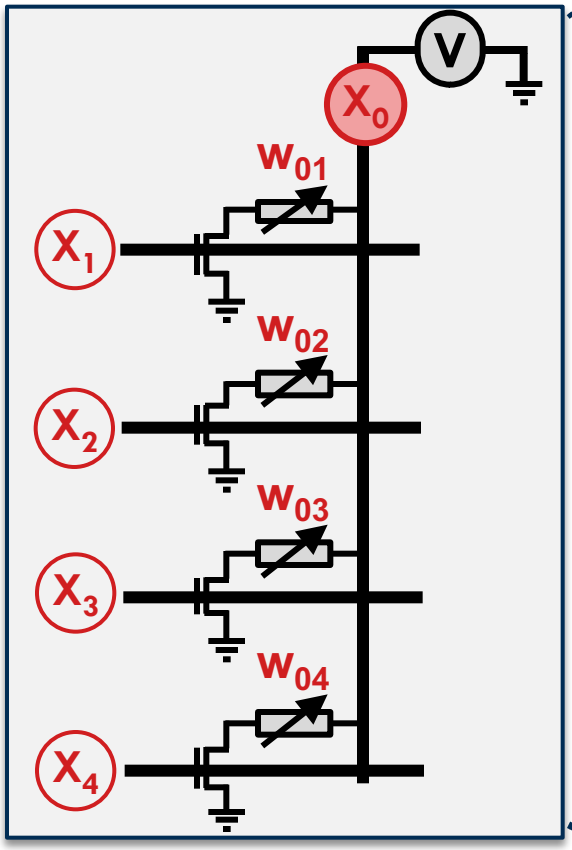


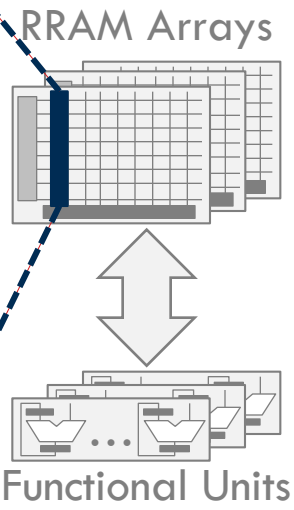
Memristive Boltzmann Machine

- Memory cells represent the weights and state variables are used to control the bitline and wordlines.



The Boltzmann Machine





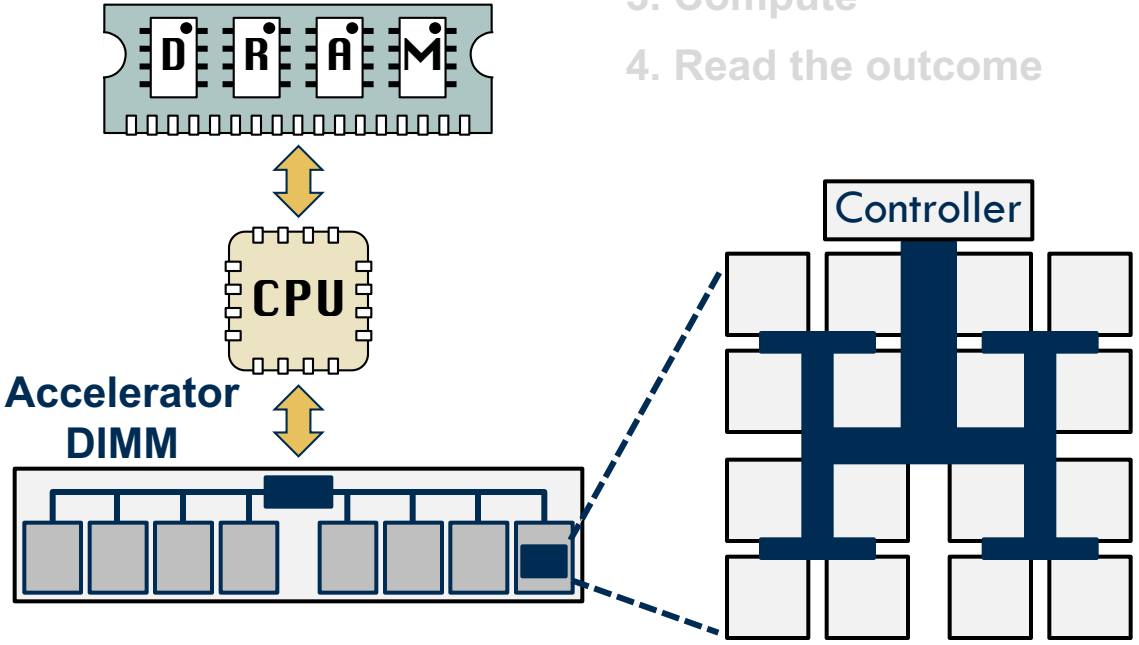
System Integration

Software configures the on-chip data layout and initiates the optimization by writing to a memory mapped control register.

To maintain ordering, accesses to the accelerator are made uncacheable by the processor.

DDR3 reads and writes are used for configuration and data transfer.

1. Configure the DIMM
2. Write weights and states
3. Compute
4. Read the outcome



Summary of Results