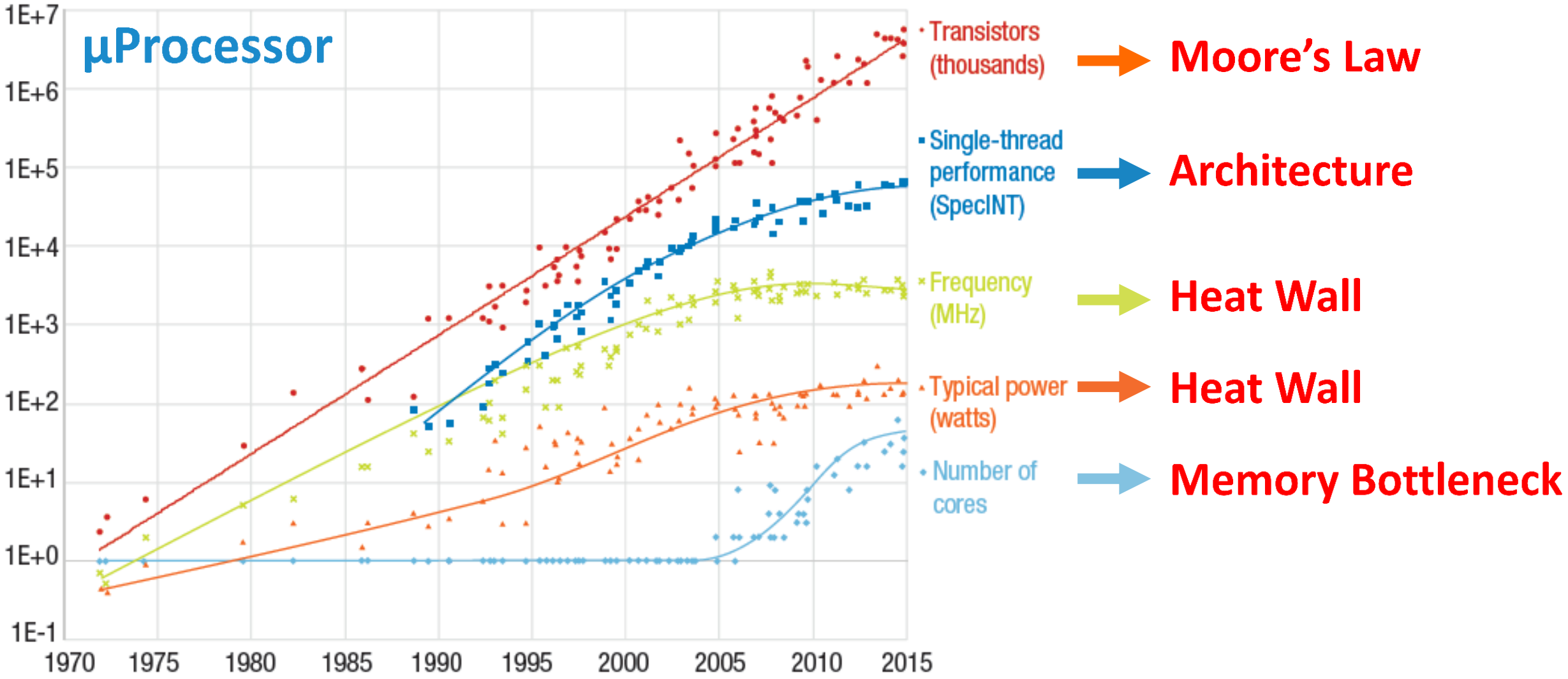
RRAM-Based Reconfigurable Computing

Mohammed Zidan and Wei D. Lu

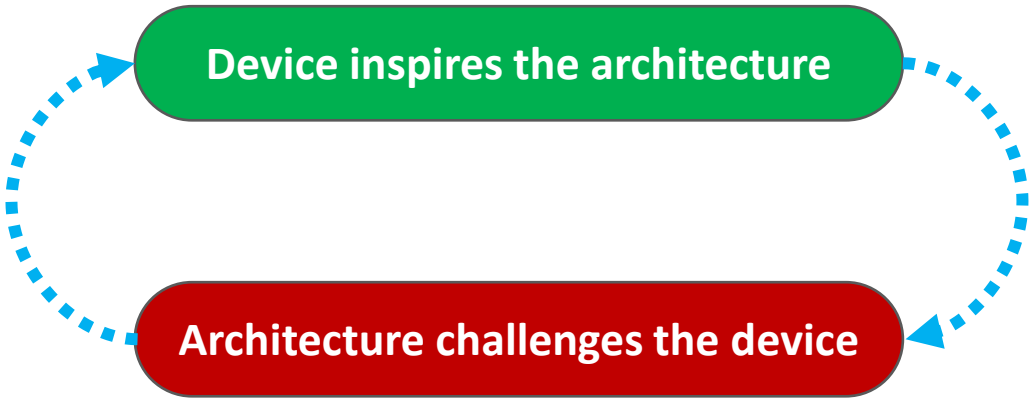
Sept 12, Tysons Corner, VA

Ageing Technology

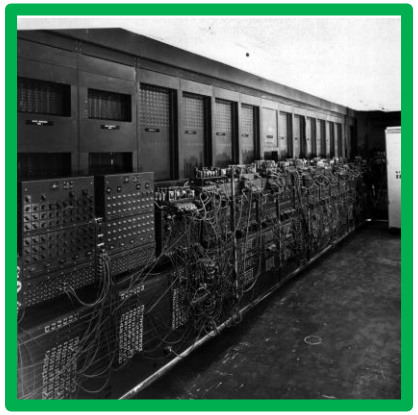
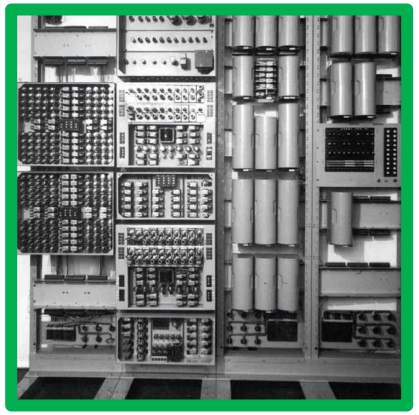
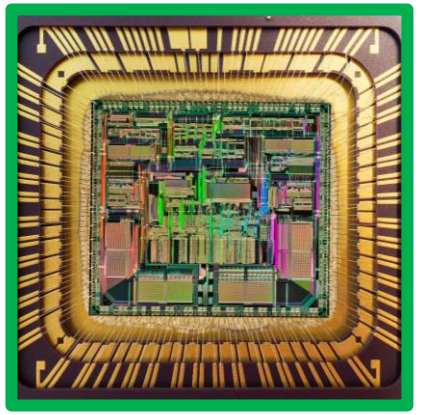


- K. Rupp, "40 Years of Microprocessor Trend Data," blog
- K. Bresniker et al., "Adapting to Thrive in a New Economy of Memory Abundance," *Computer* 2015.

Innovation Cycle



Generations

First	Second	Third	Fourth	Fifth
				?
CRTs	Transistors	ICs	Microprocessor (Moor's Low)	

Modern Applications

Big Data

Cognitive

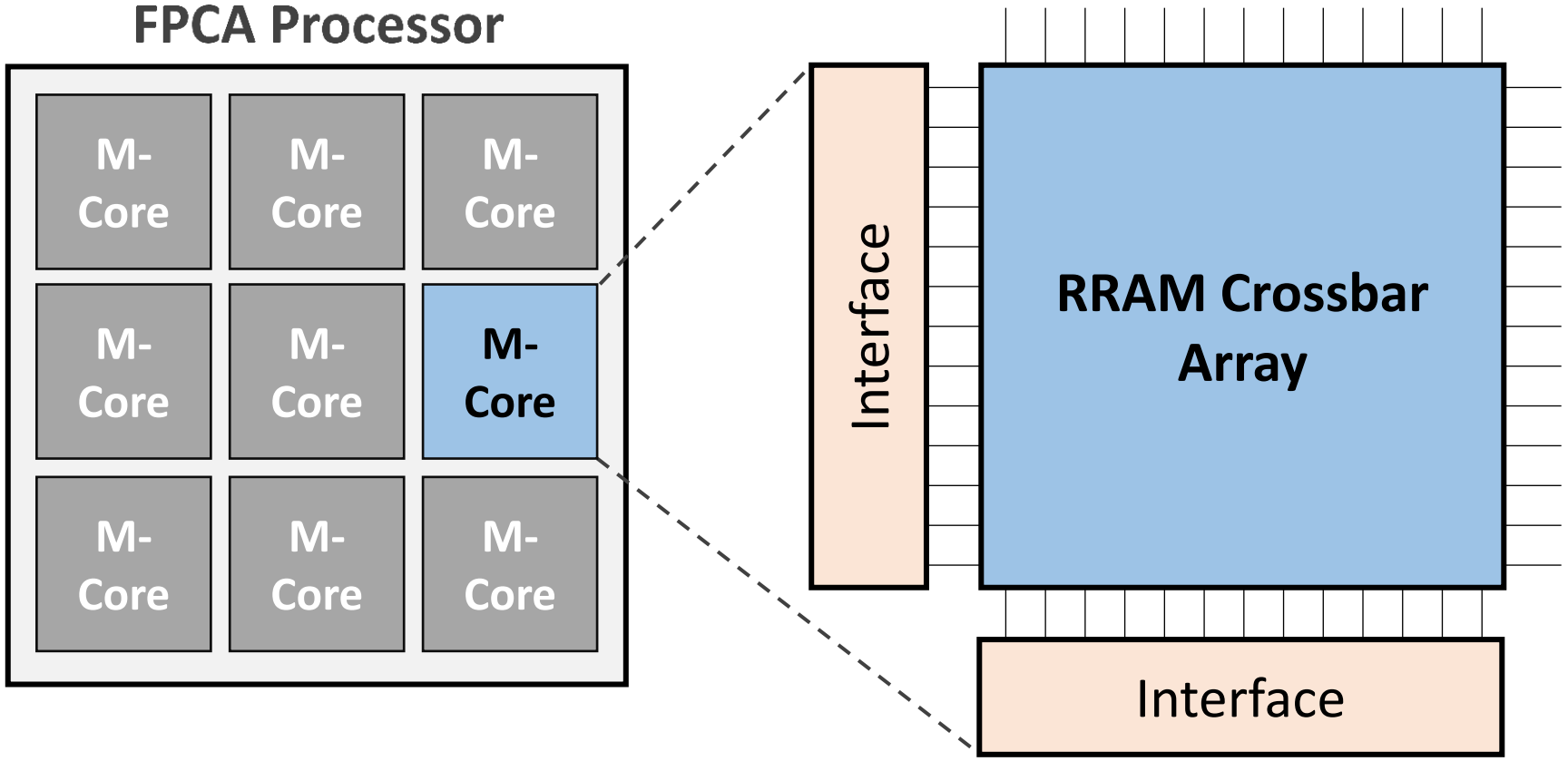
IoT

Massivly Prallel

New Computing Devices	No Memory Bottleneck	Classical Processing	Cognitive Processing	Energy Efficient

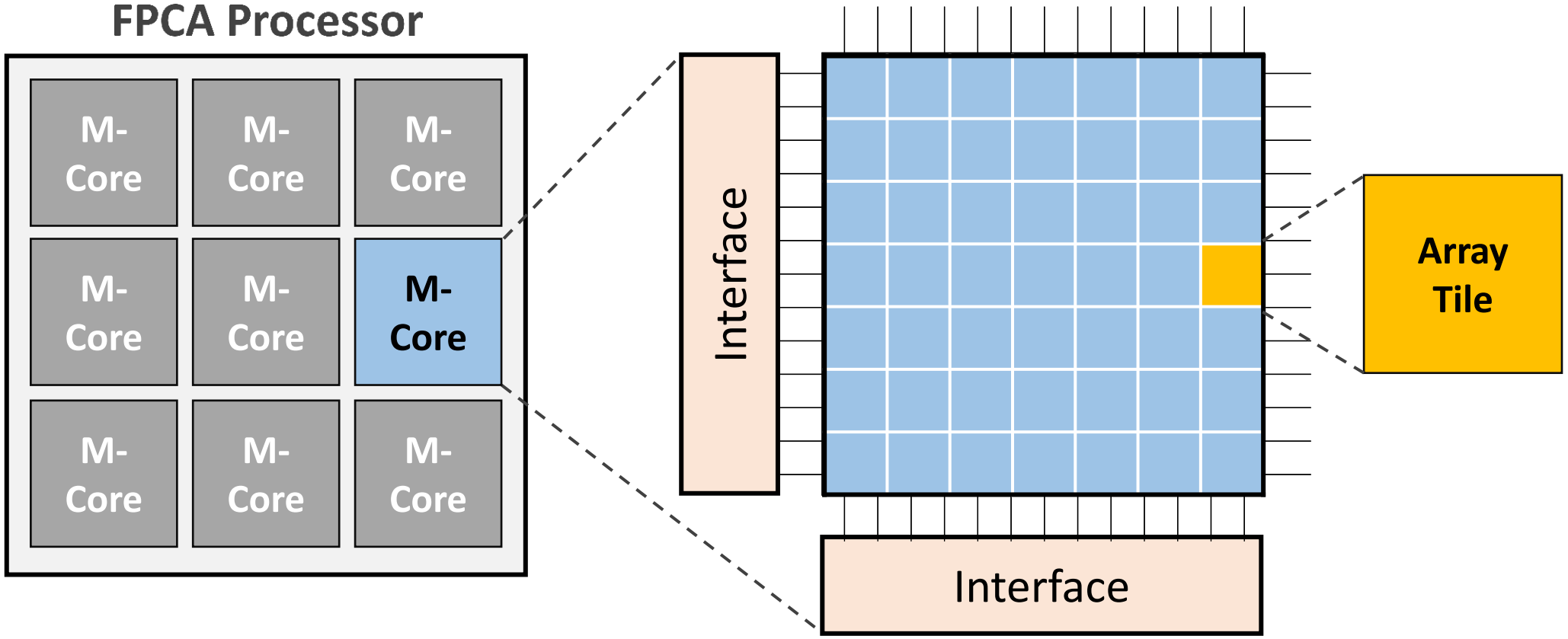
Field Programmable Crossbar Array

» In-Memory Reconfigurable Computing



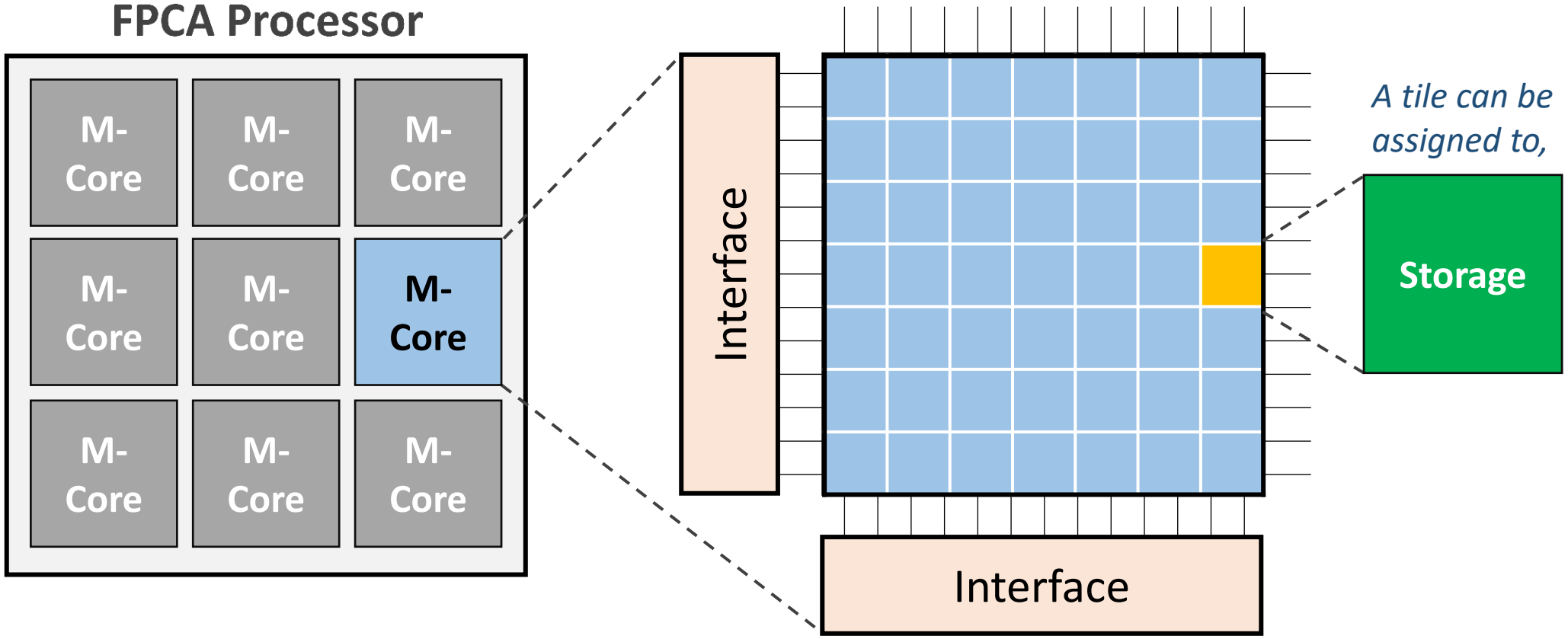
Field Programmable Crossbar Array

» In-Memory Reconfigurable Computing



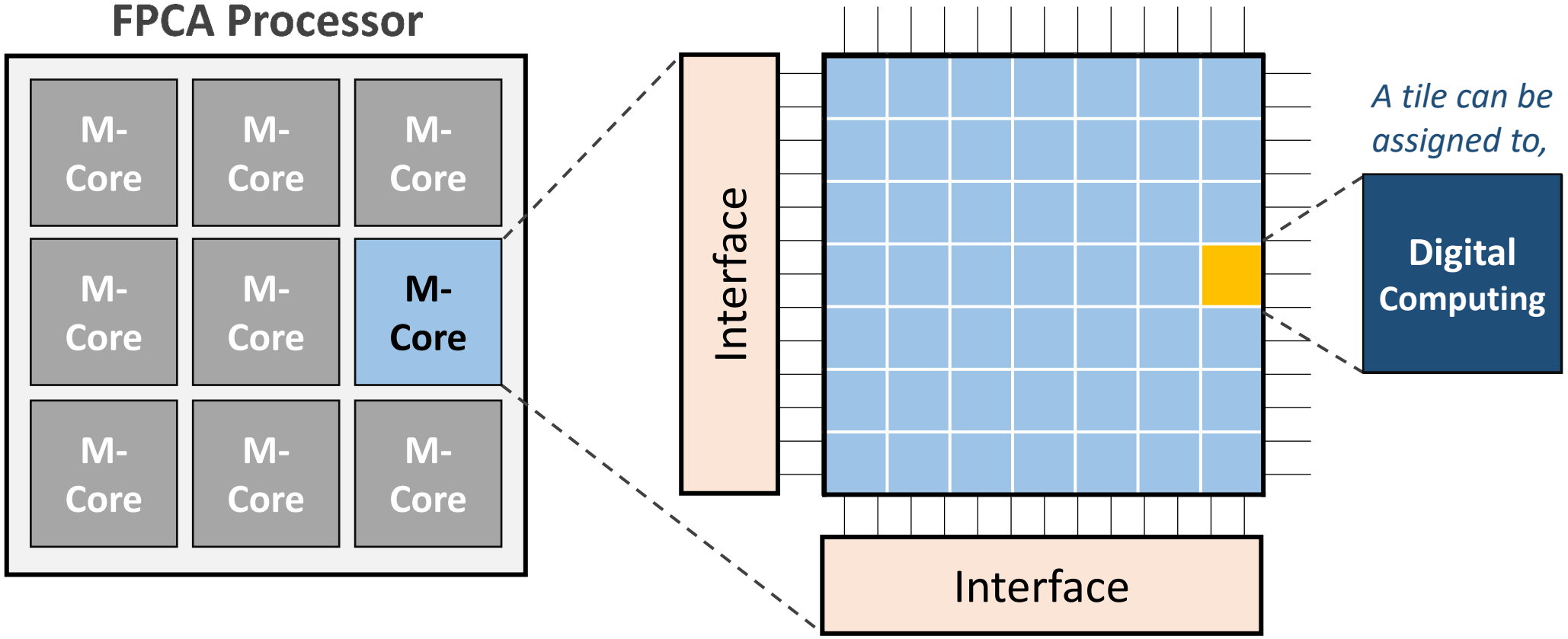
Field Programmable Crossbar Array

» In-Memory Reconfigurable Computing



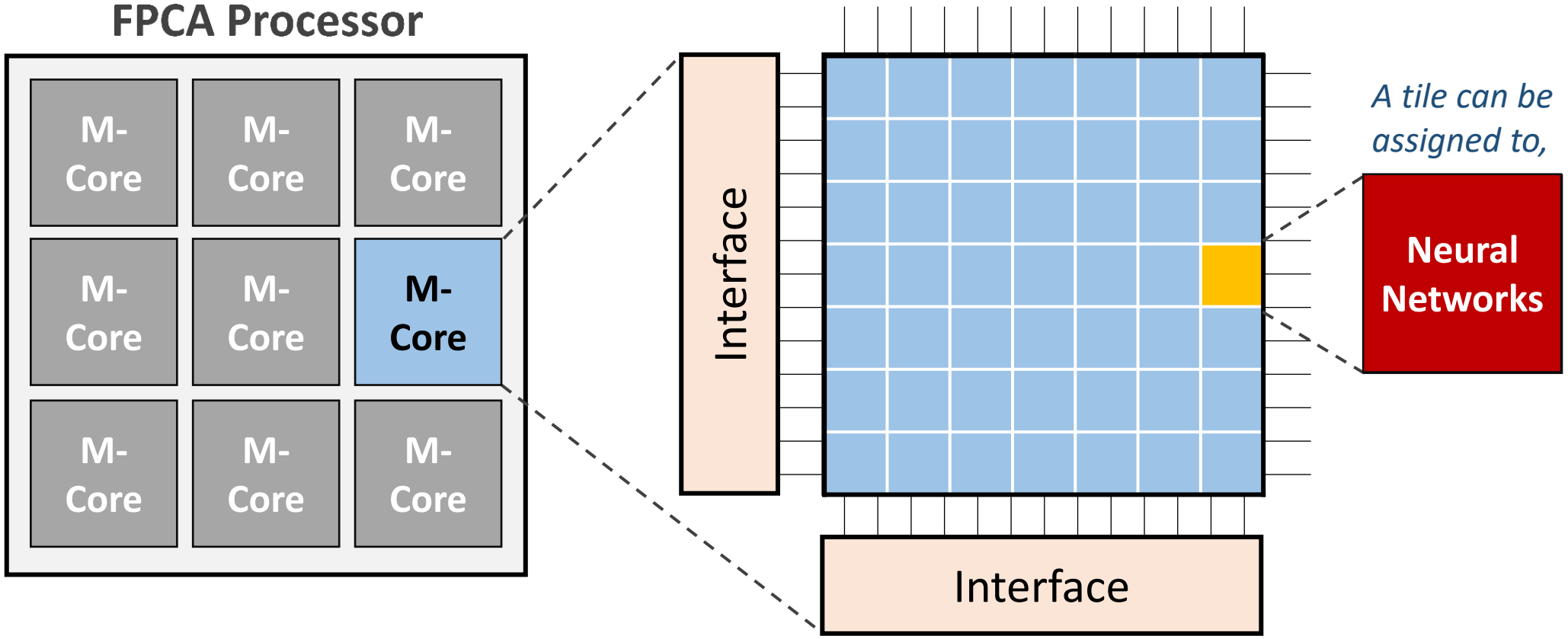
Field Programmable Crossbar Array

» In-Memory Reconfigurable Computing



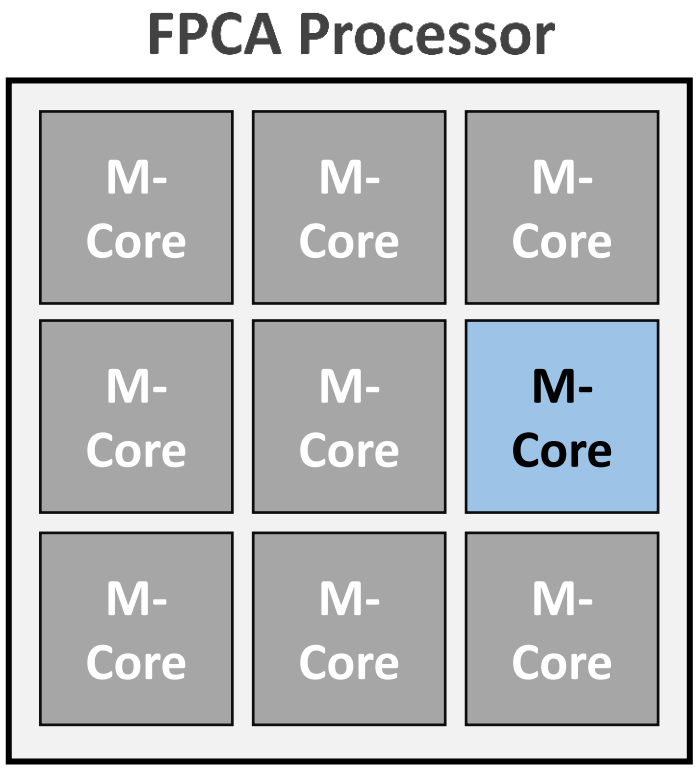
Field Programmable Crossbar Array

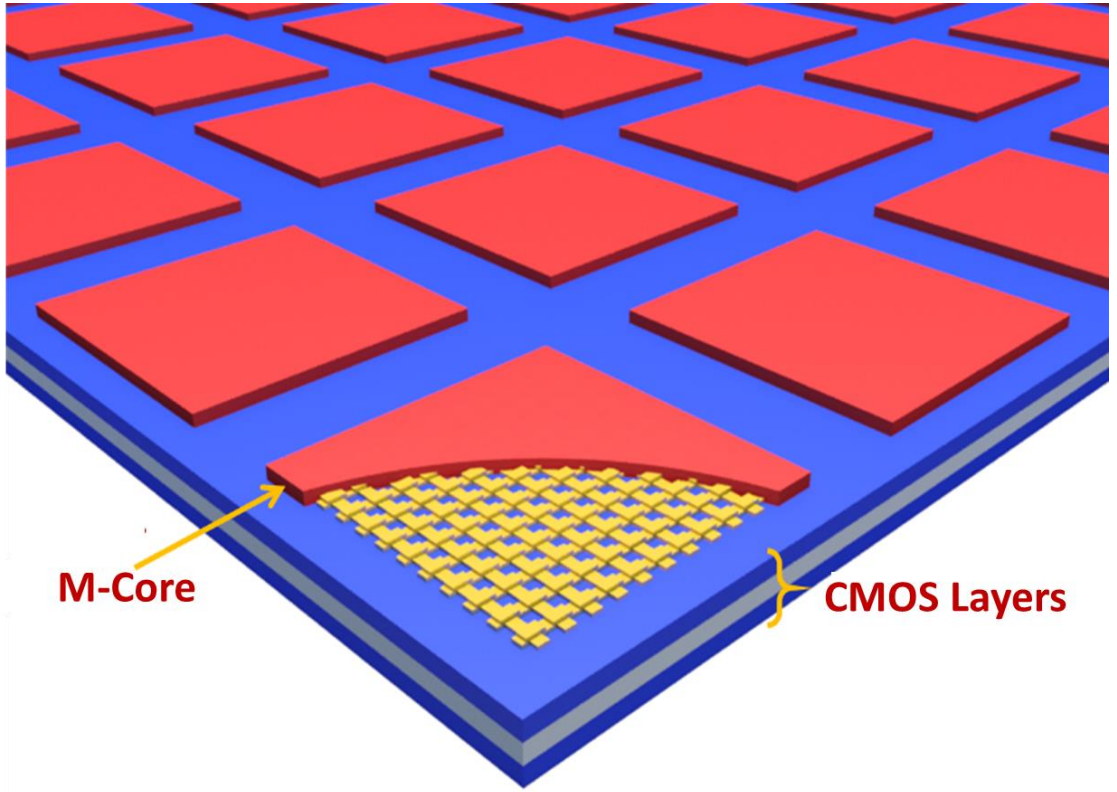
» In-Memory Reconfigurable Computing



Field Programmable Crossbar Array

» Monolithic Integration

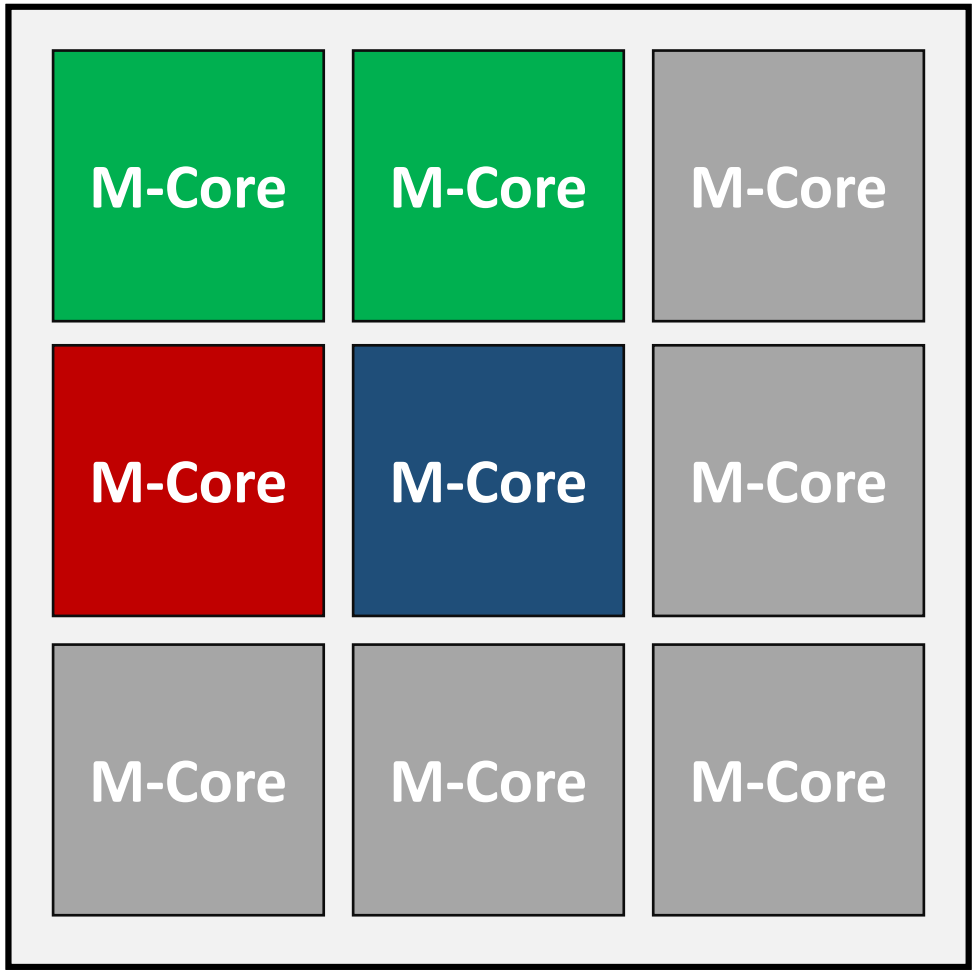




Field Programmable Crossbar

» Reconfigurable Cores

Workload "A"

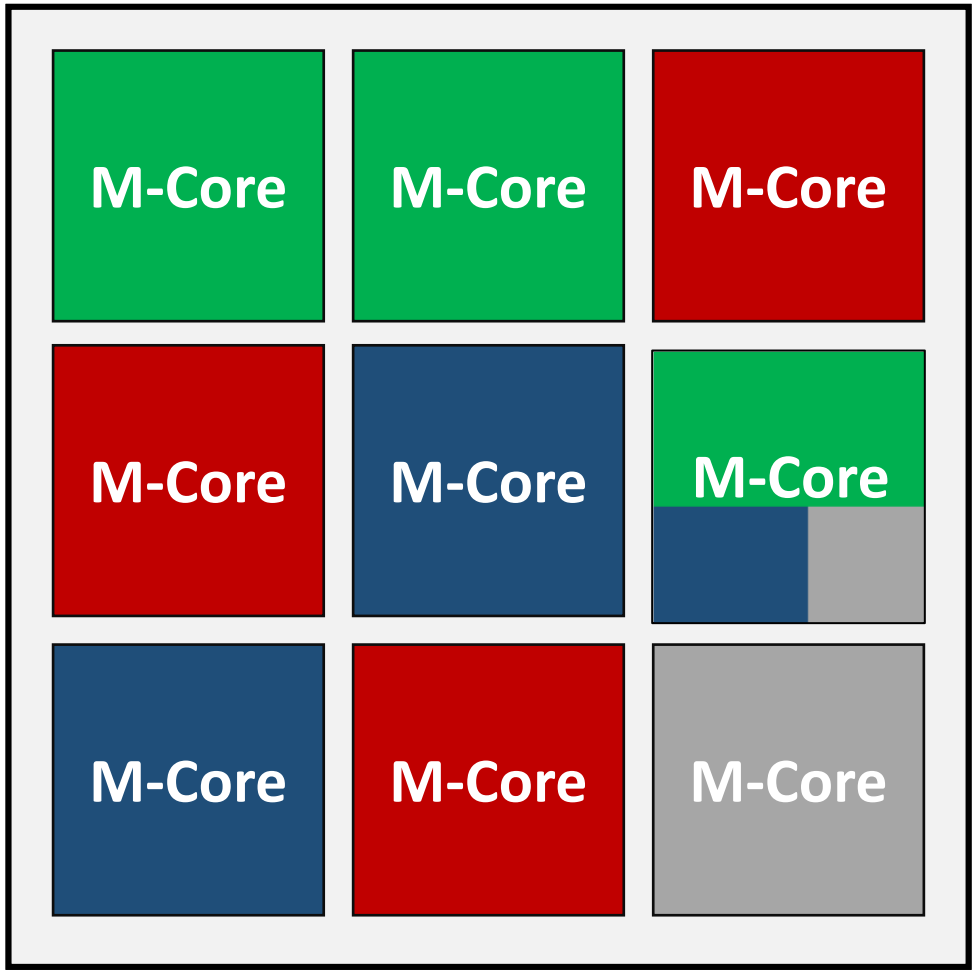


- Unused
- Storage
- Digital Computing
- Analog Computing

Field Programmable Crossbar

» Reconfigurable Cores

Workload "B"

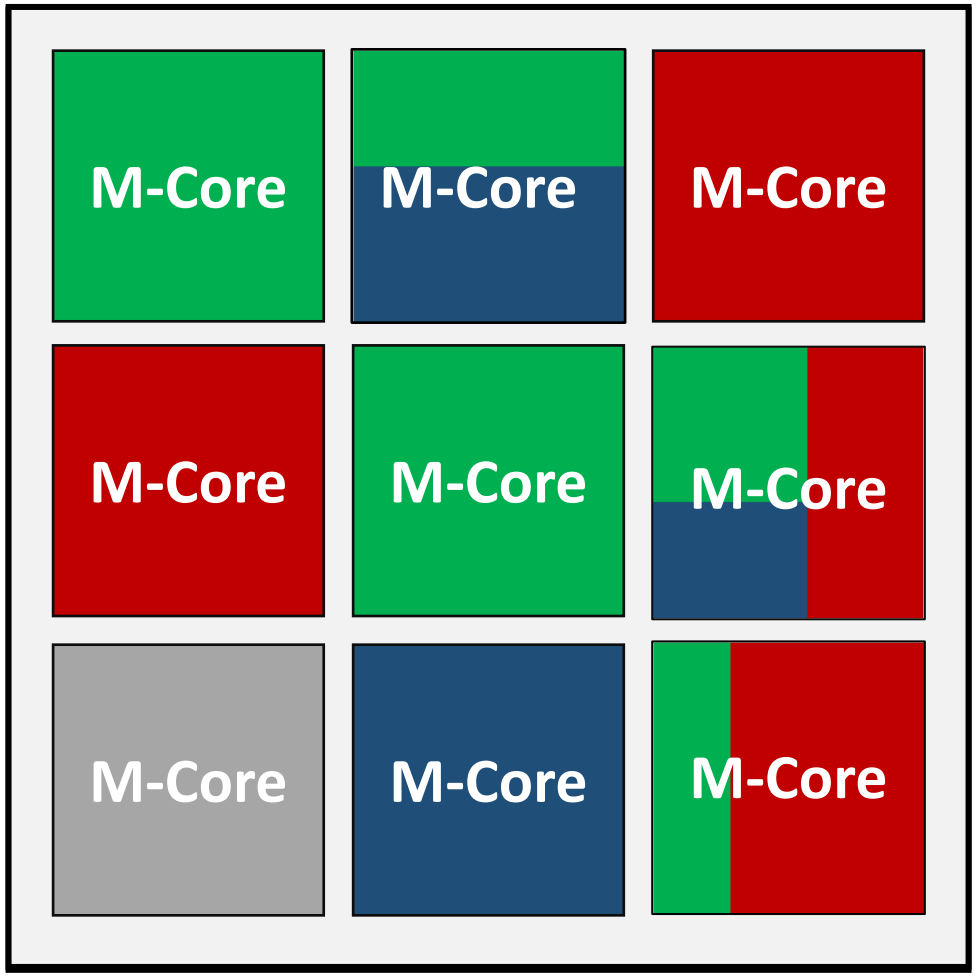


- Unused
- Storage
- Digital Computing
- Analog Computing

Field Programmable Crossbar

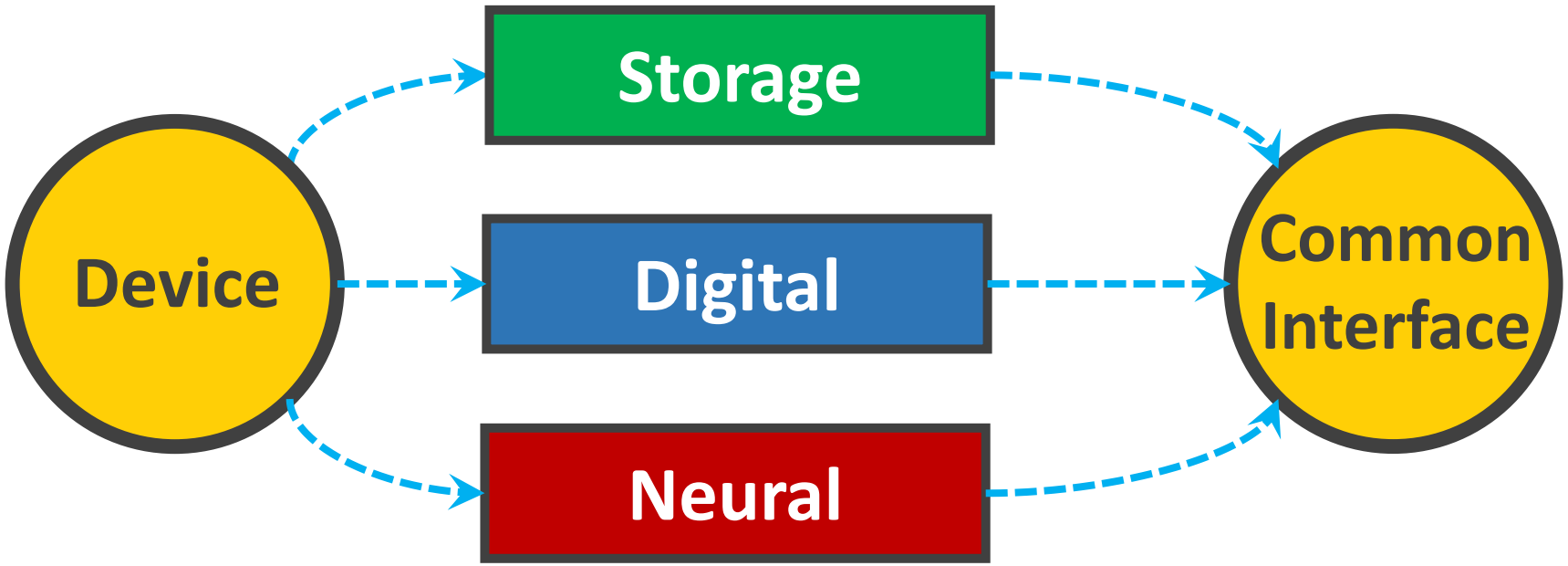
» Reconfigurable Cores

Workload "C"



- Unused
- Storage
- Digital Computing
- Analog Computing

Enabling Technology



Reconfigurable Computing

» Binary vs. Analog Devices

RRAM Type	Analog	Binary
Device Levels	**<100**	2
ON/OFF Ratio	Average	**High**
Endurance	Average	**High**
Programing	Slow	**Fast**

Reconfigurable Computing

» Binary vs. Analog Devices

RRAM Type	Analog	Binary
Device Levels	<100	2
ON/OFF Ratio	Average	High
Endurance	Average	High
Programing	Slow	Fast
Data Storage	✓	✓
Digital Computing		✓
Neural Networks	✓	
Internal Data Movement		

Reconfigurable Computing

» Binary vs. Analog Devices

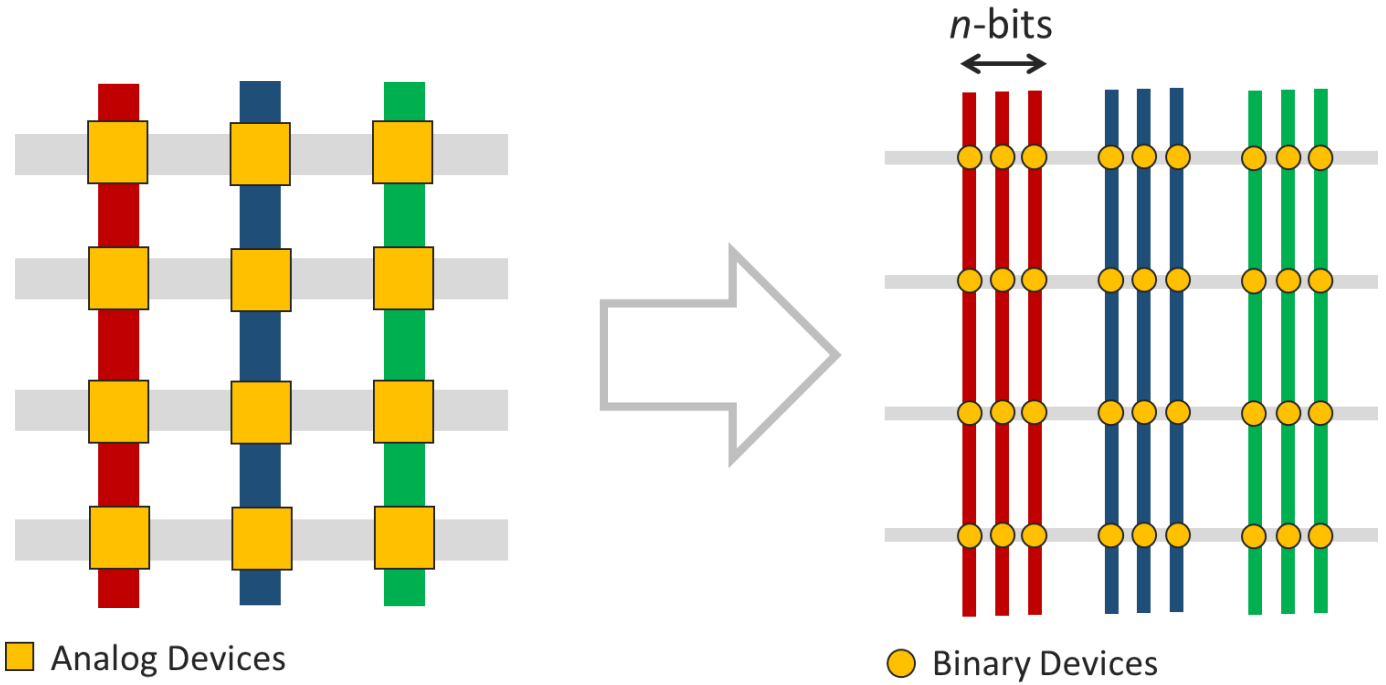
RRAM Type	Analog	Binary
Device Levels	<100	2
ON/OFF Ratio	Average	High
Endurance	Average	High
Programing	Slow	Fast
Data Storage	✓	✓
Digital Computing		TR & PDE
Neural Networks	✓	BCNN
Internal Data Movement		In-Situ

Reconfigurable Computing

A. Binary Coded Neural Networks

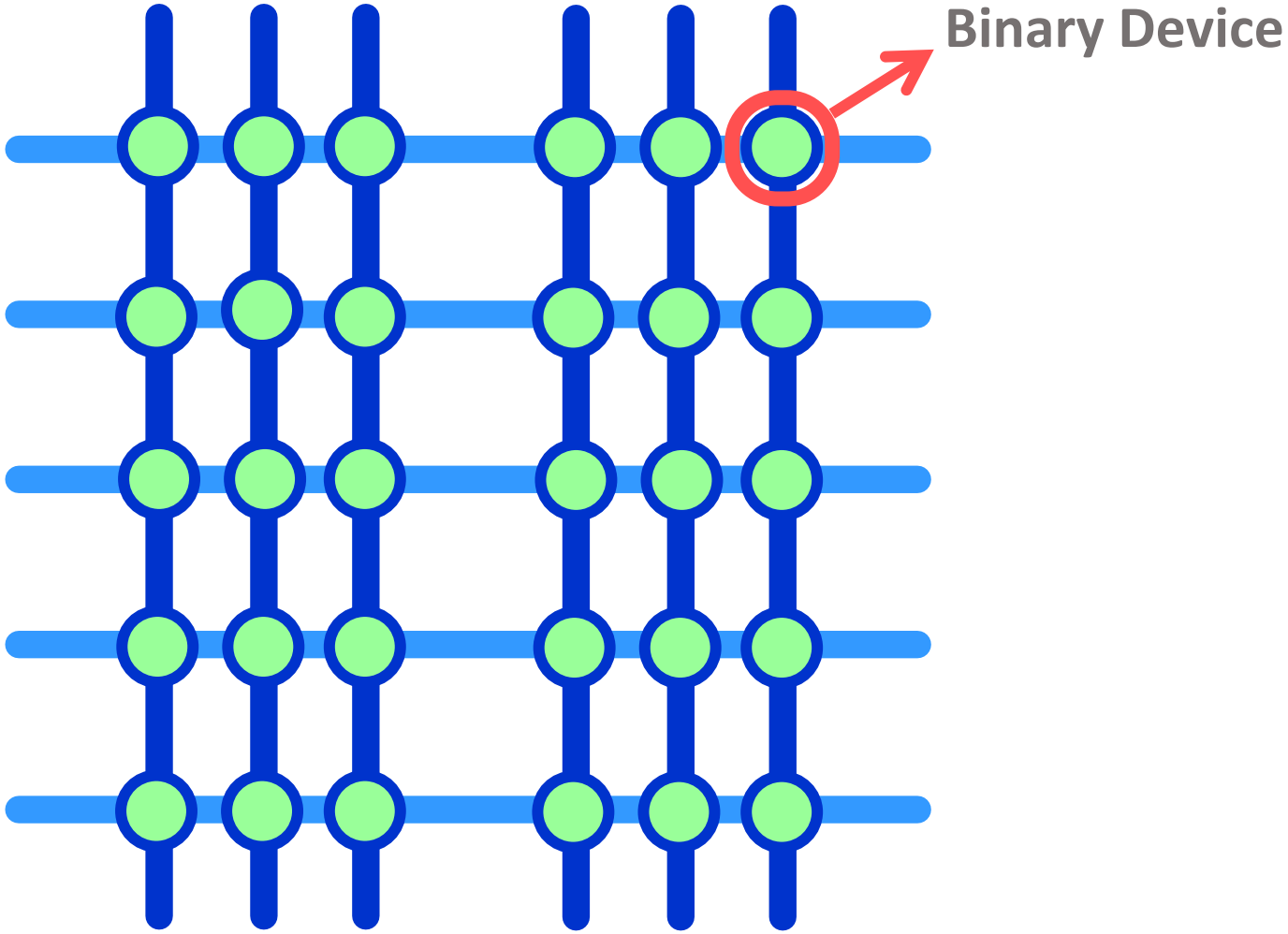
» Binary Synaptic weights

- Our approach is utilized binary RRAM devices to implement a semi-analog (hybrid) neural network.
- Each classical analog device is replaced with “n” binary devices in the new network.
- The number of synaptic-weight bits can be dynamically configured when needed.



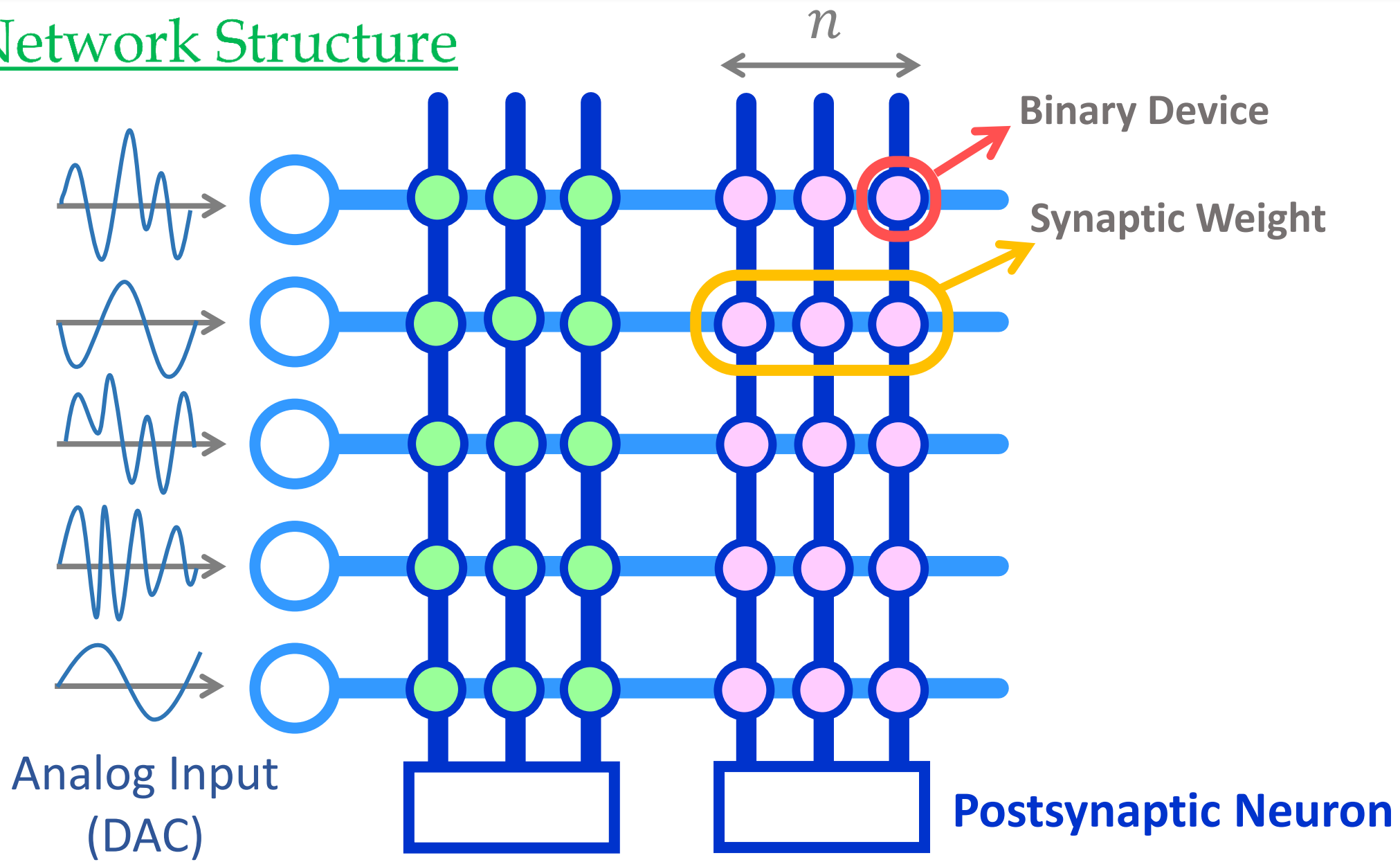
Hybrid Neural Network

» Network Structure



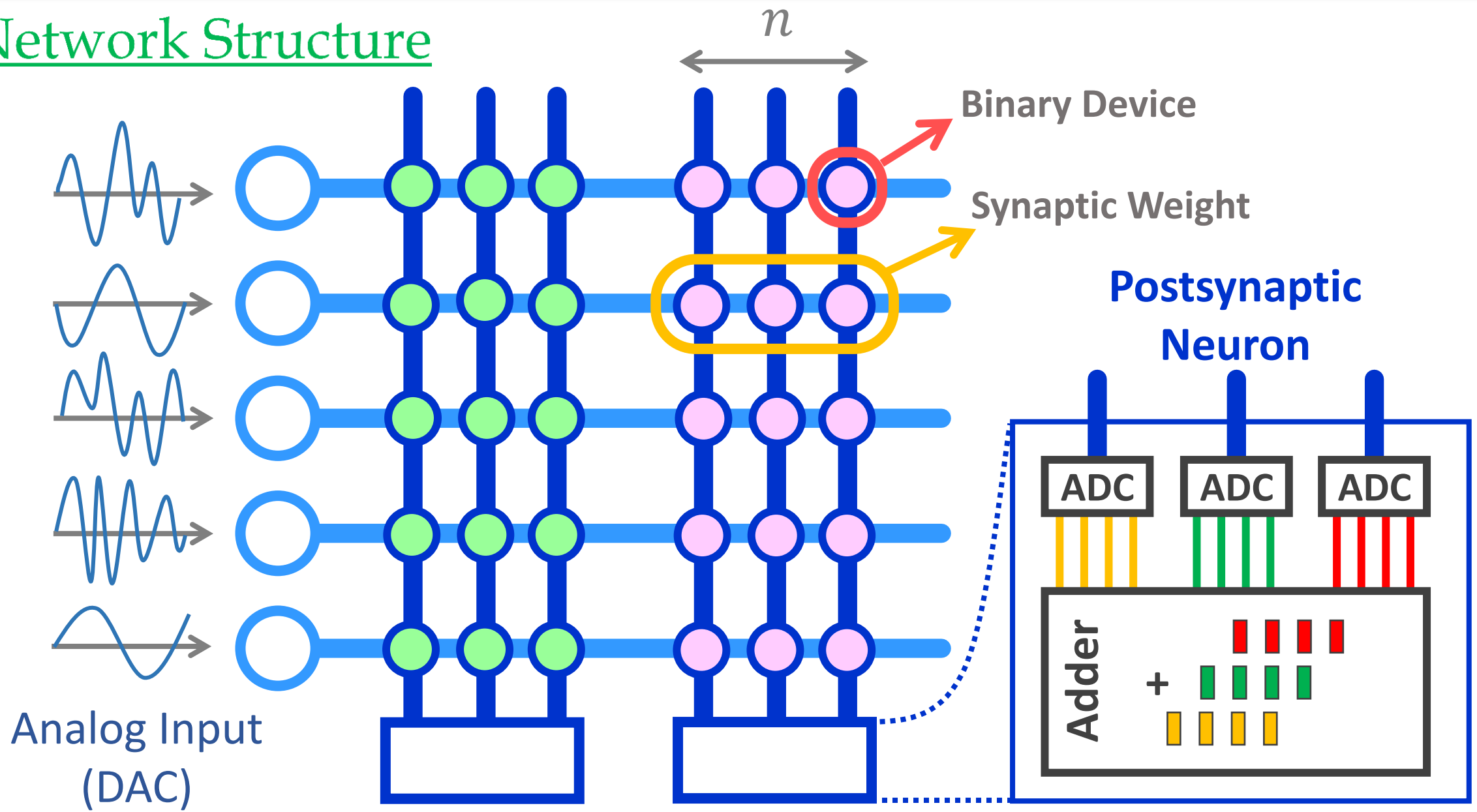
Hybrid Neural Network

» Network Structure



Hybrid Neural Network

» Network Structure

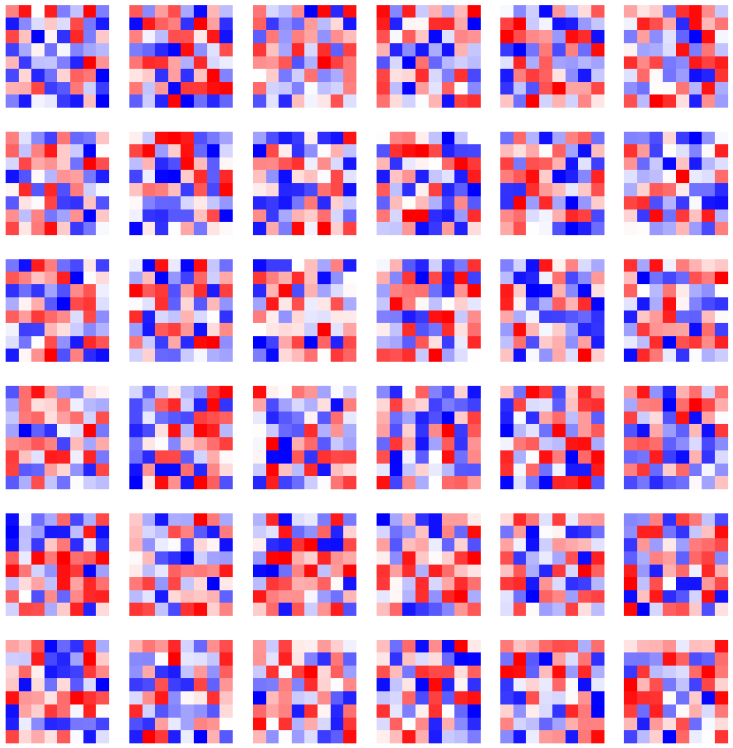


Network Training

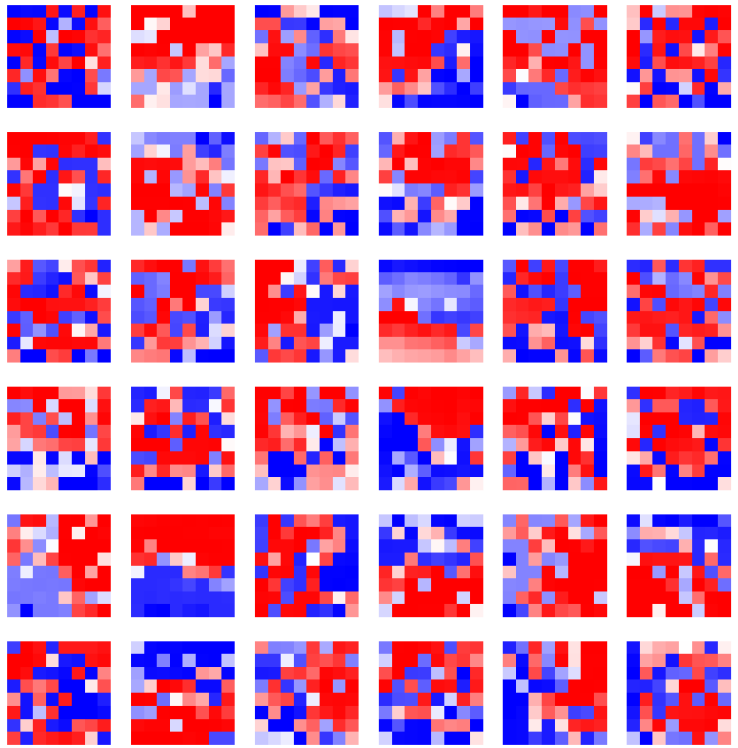
» Training Precision Effect

- Low training rate is required to train the network receptive fields (dictionaries) properly.
- This is translated into a larger number of bits to allow small "Δ*w*" values.

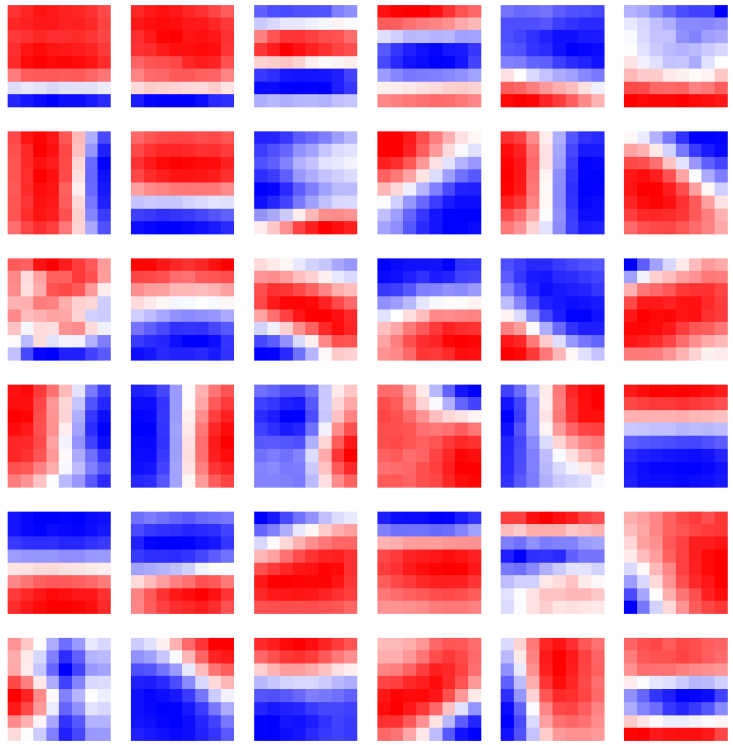
"n = 4"



"n = 8"



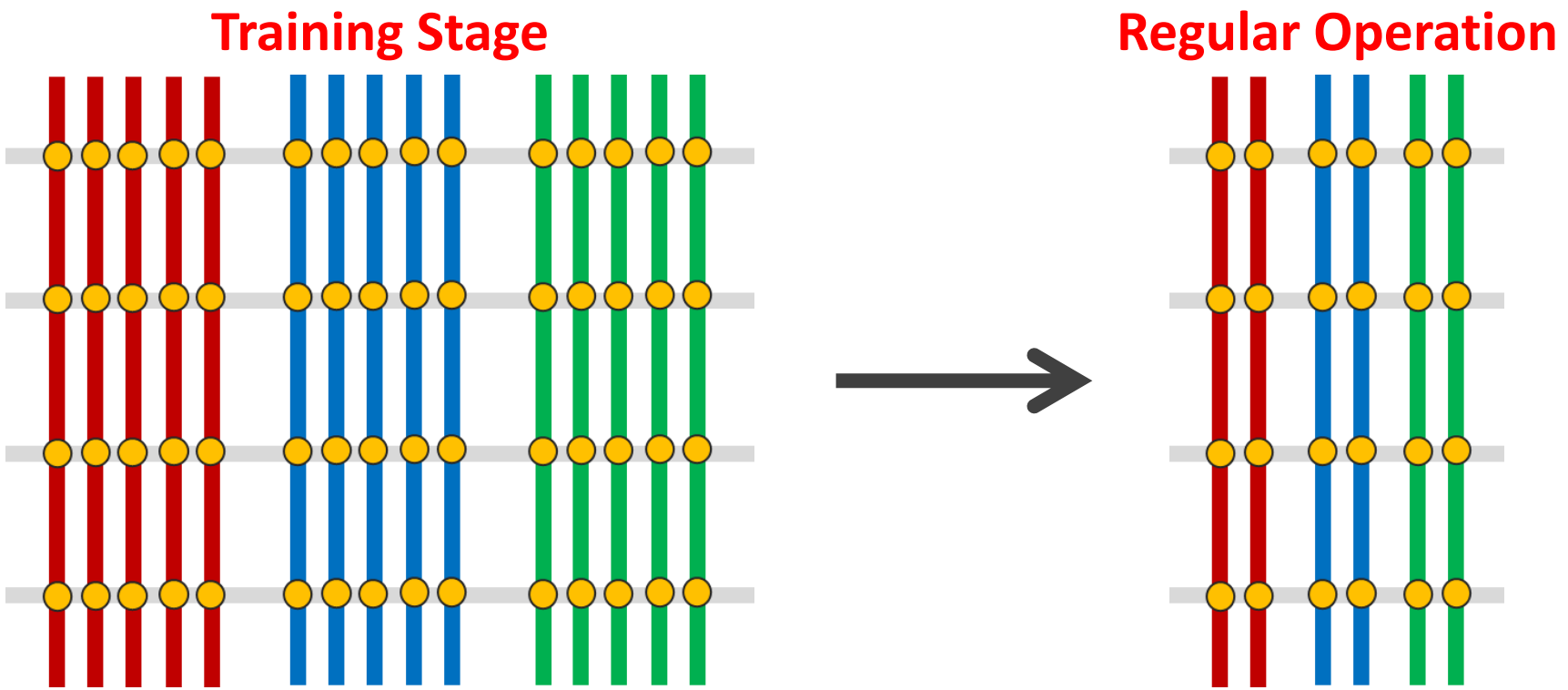
"n = 16"



Network Training

» Training Precision Effect

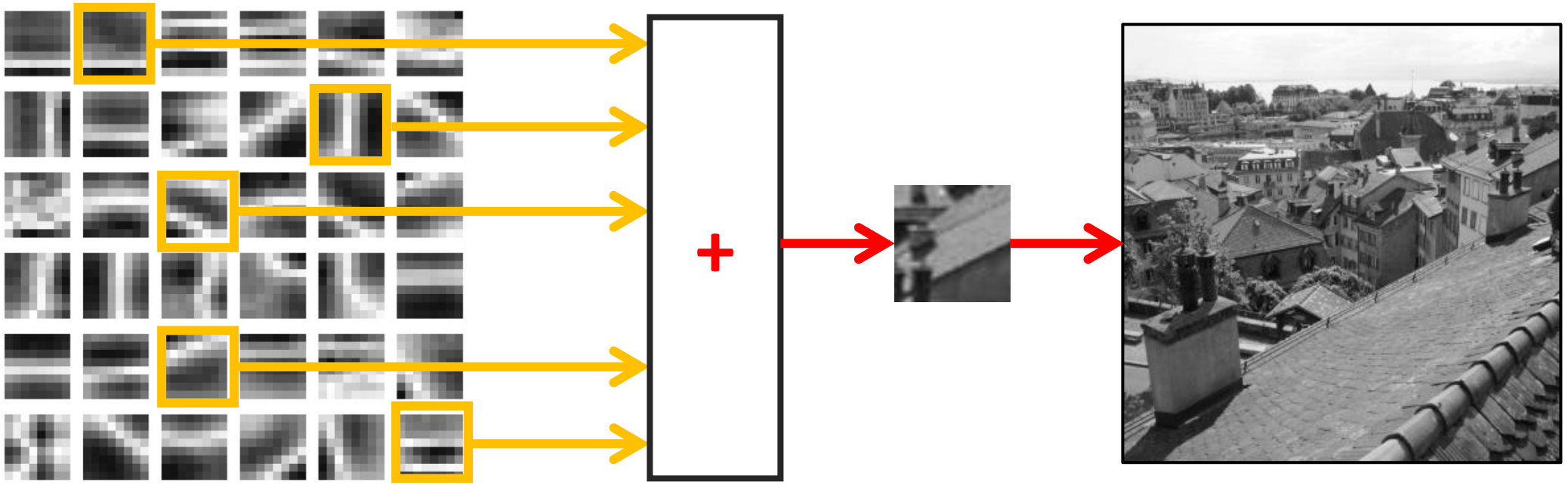
- Typically, training is infrequent or is performed offline.
- Hence, after training the number bits per synaptic weights can be significantly reduced by assigning fewer columns per neuron.



Analog Image Compression

» Sparse Coding

- In analog image compression (sparse coding) each piece of a picture is represented as weighted combination of the network dictionary.
- We adopted locally competitive algorithm (LCA) to perform the analog image compression.



Analog Image Compression

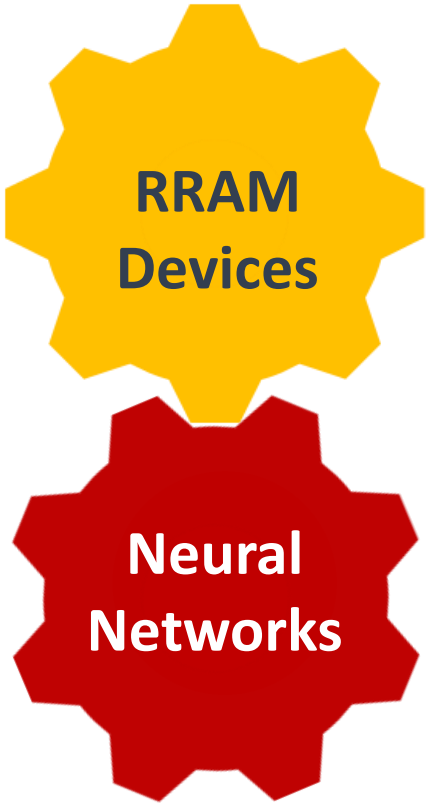
» Results

“Original”

“Reconstructed”

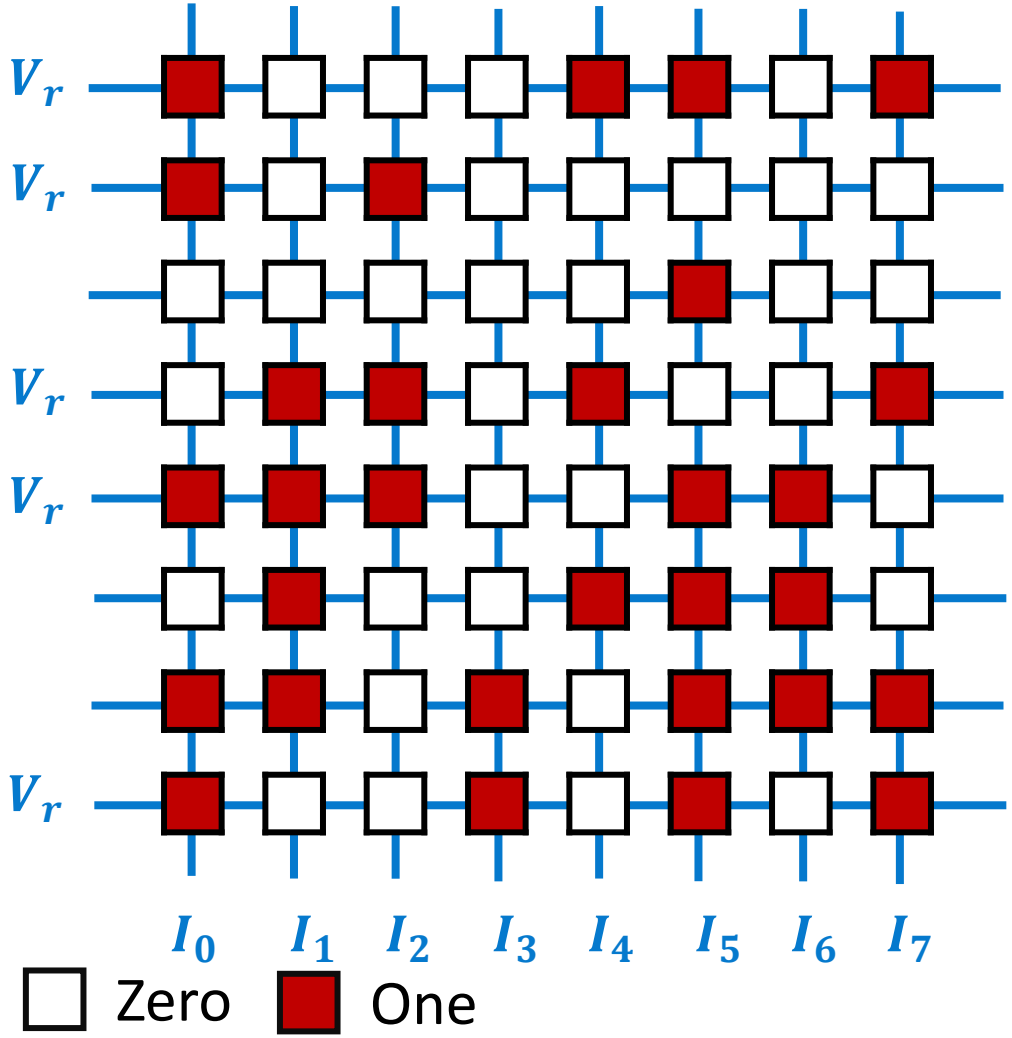
Reconfigurable RRAM Computing

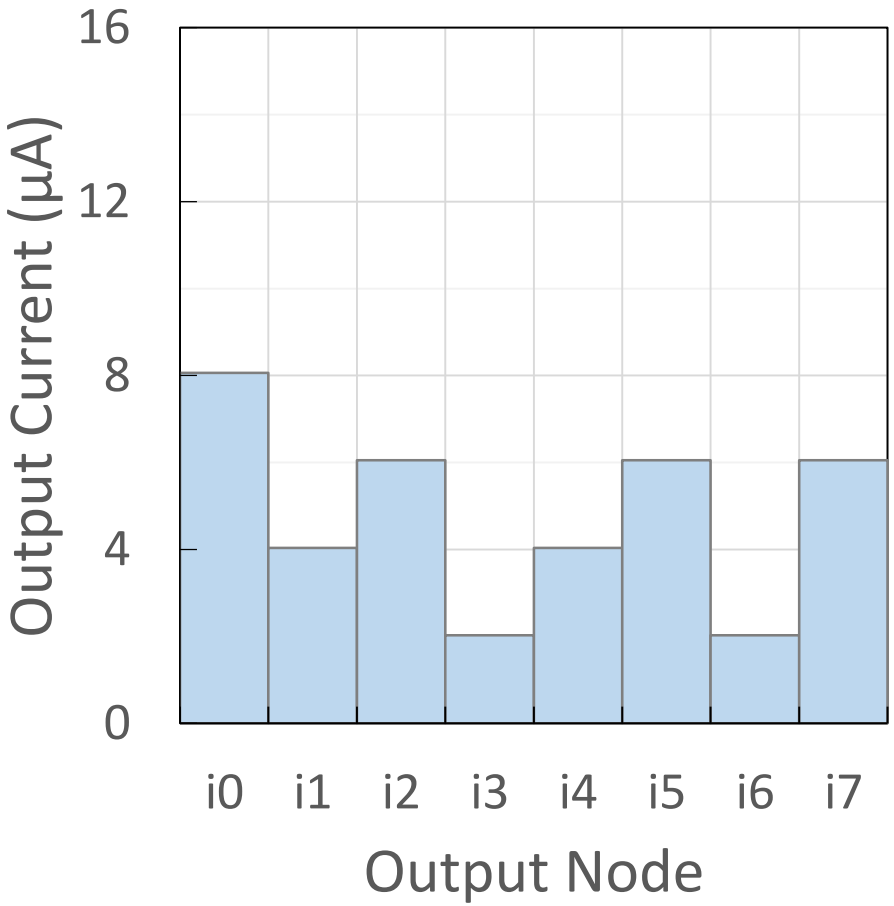
Reconfigurable RRAM Computing



B. Digital Computing

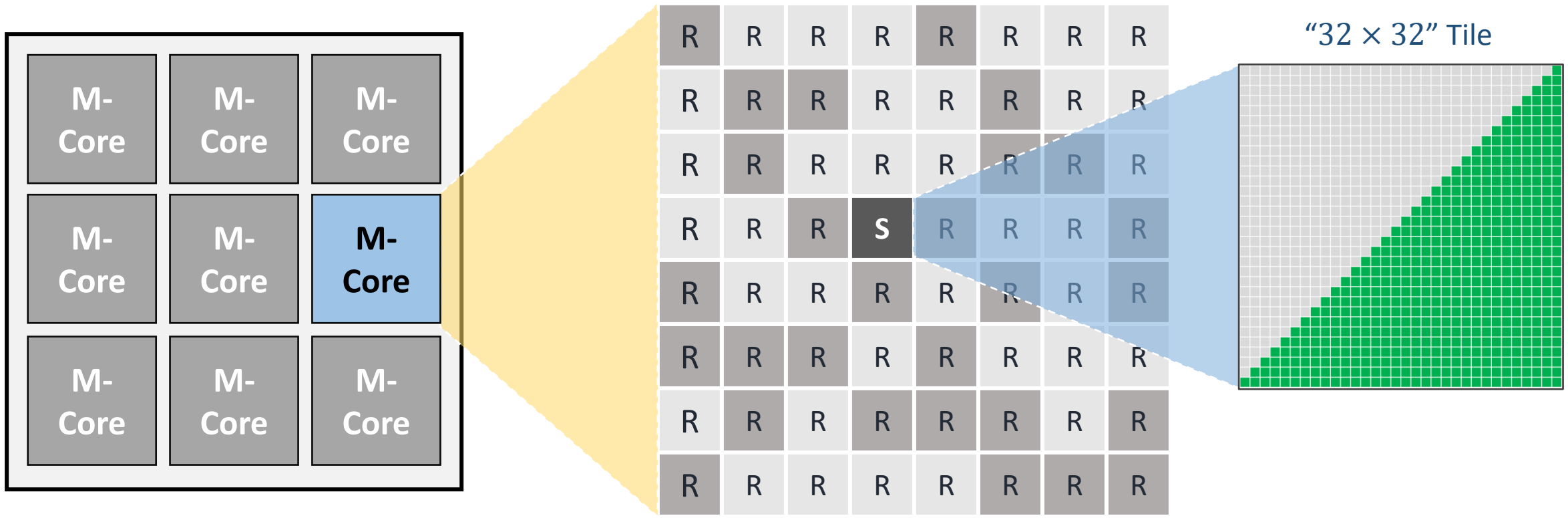
» Crossbar Tree Reduction





B. Digital Computing

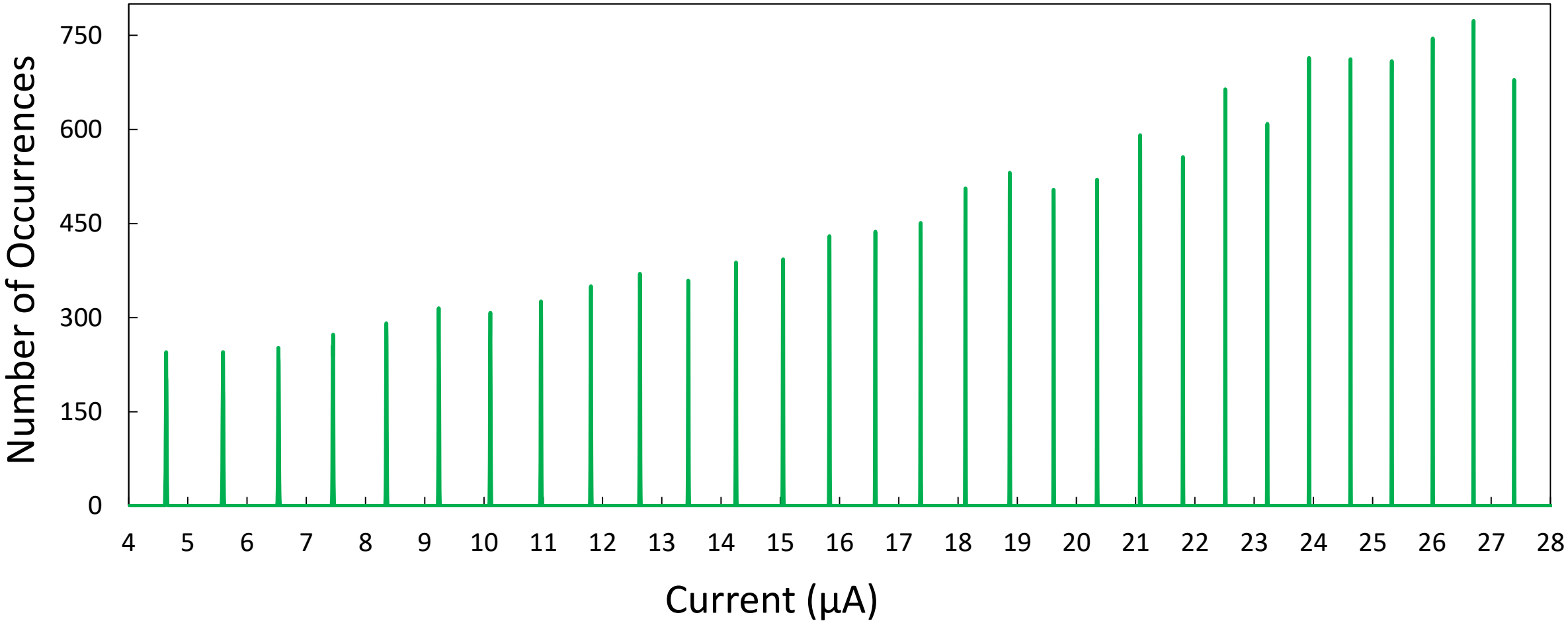
» Crossbar Tree Reduction



B. Digital Computing

» Crossbar Tree Reduction

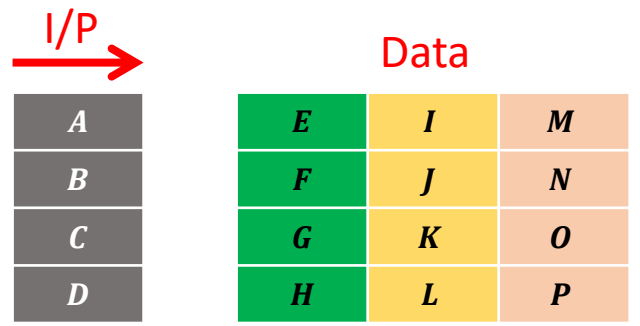
44,800 Simulation points

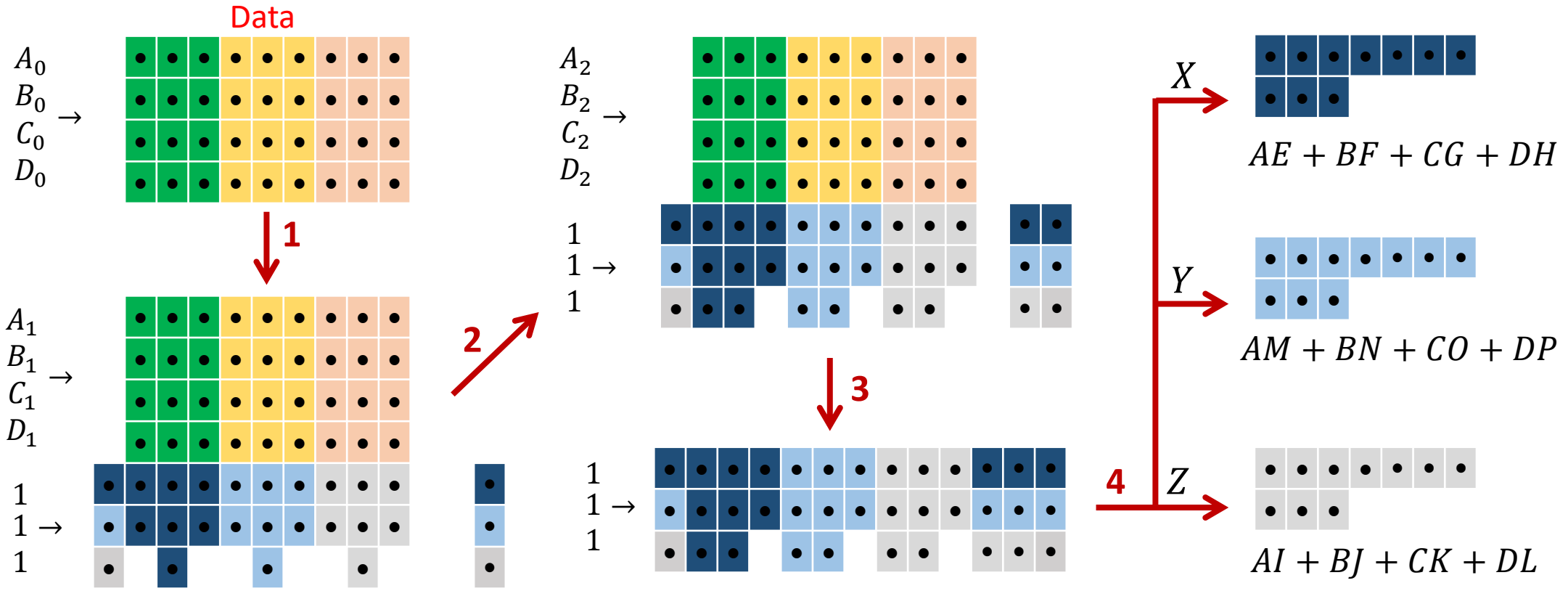


B. Digital Computing

» Crossbar Tree Reduction

$$[A \quad B \quad C \quad D] \cdot \begin{bmatrix} E & I & M \\ F & J & N \\ G & K & O \\ H & L & P \end{bmatrix} = [X \quad Y \quad Z]$$

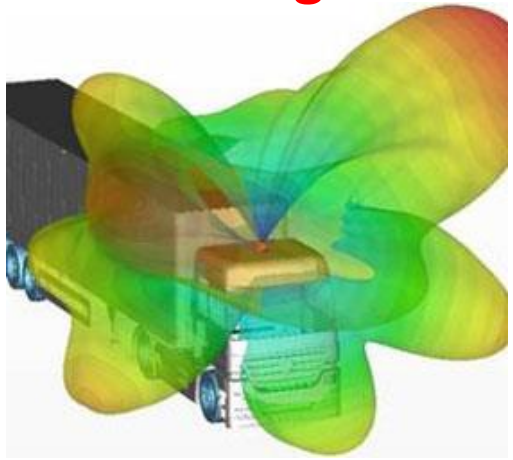




B. Digital Computing

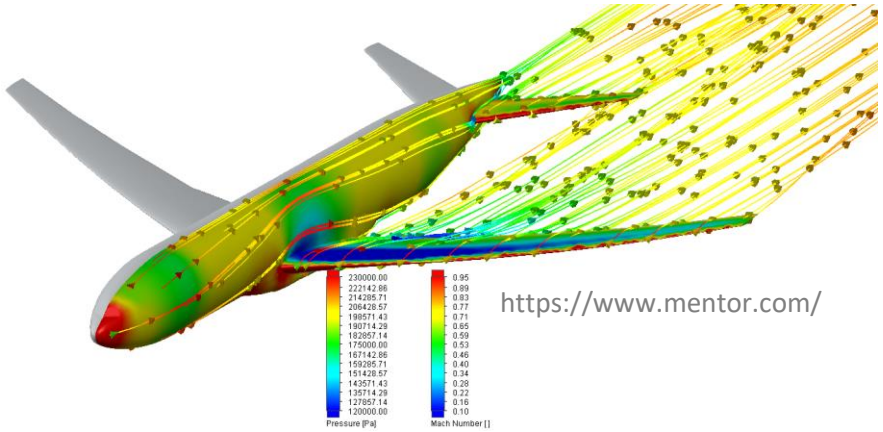
» PDE Solver

Electromagnetics



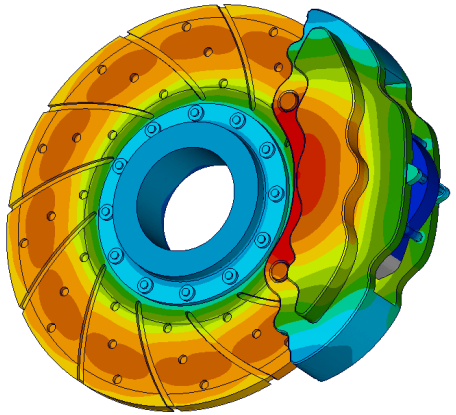
https://www.horiba-mira.com/

Aerodynamics



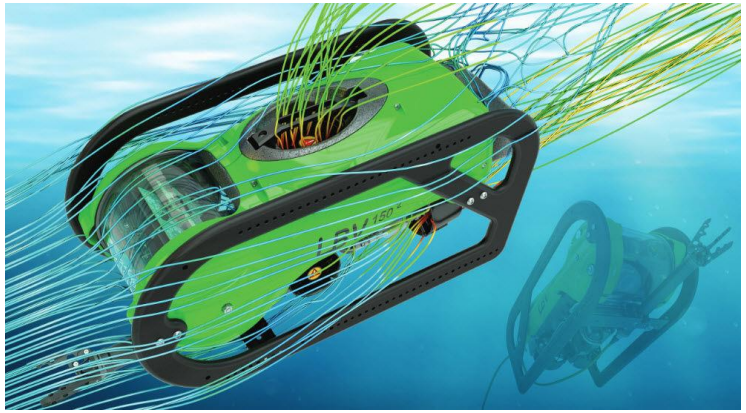
https://www.mentor.com/

Heat Transfer



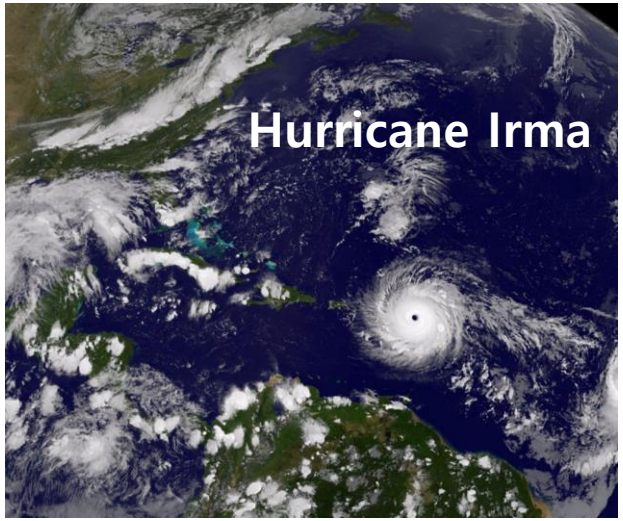
http://www.theseus-fe.com

Fluids Flow



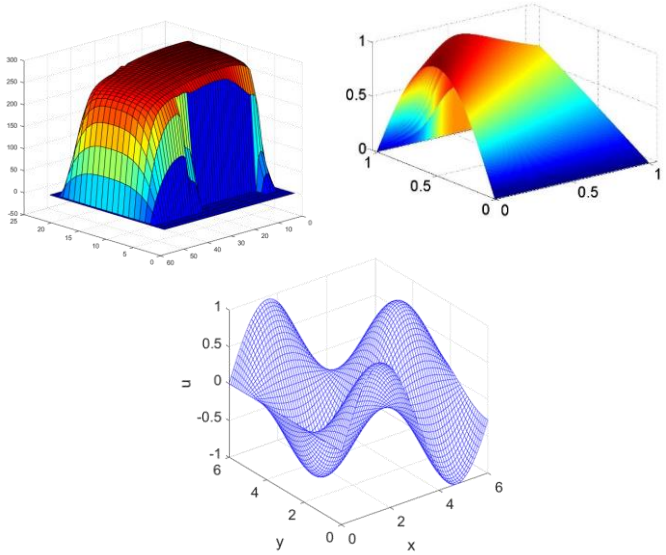
http://www.metrosystems-des.com

Weather Forecasting



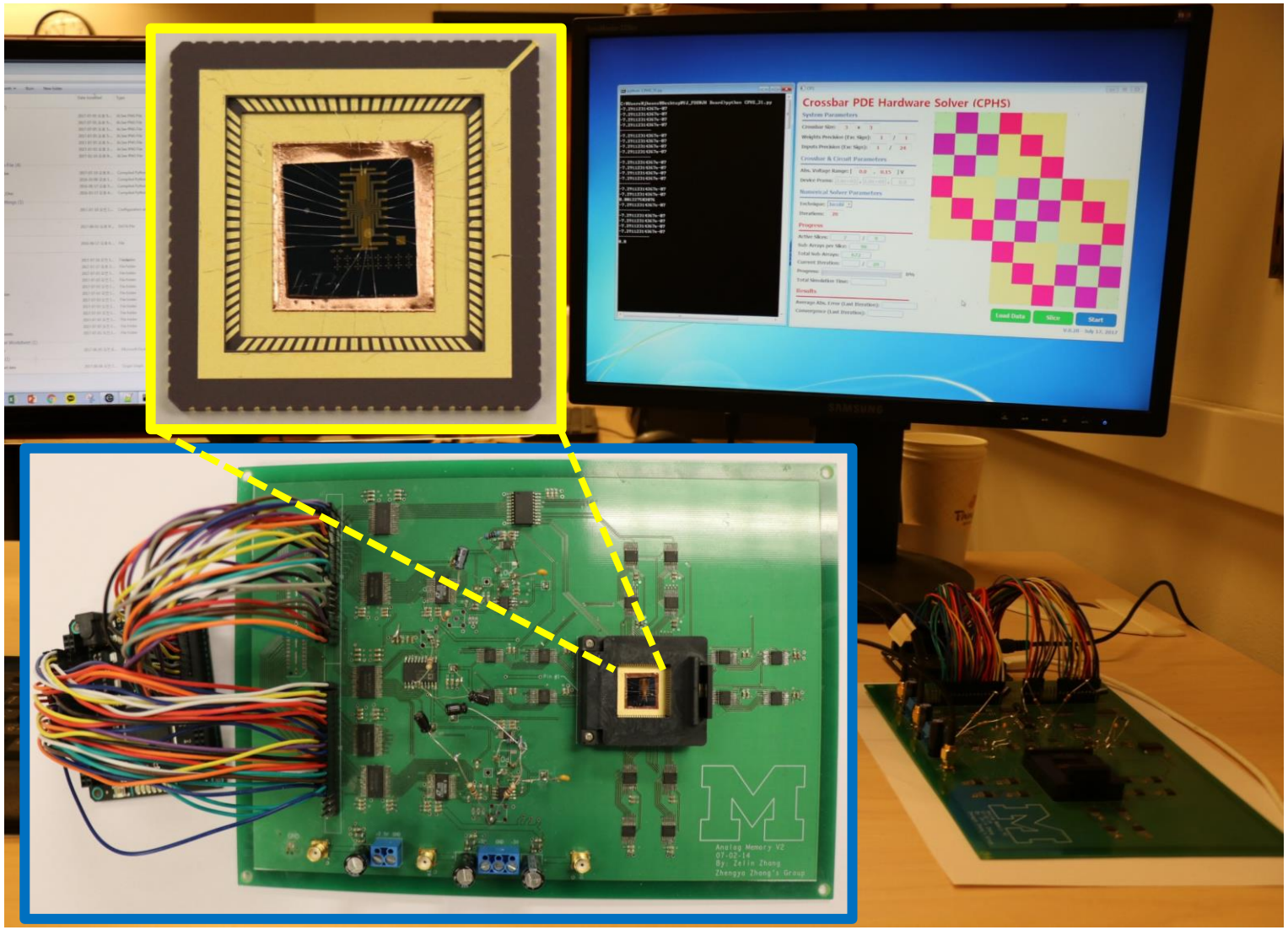
NASA

Many Others



B. Digital Computing

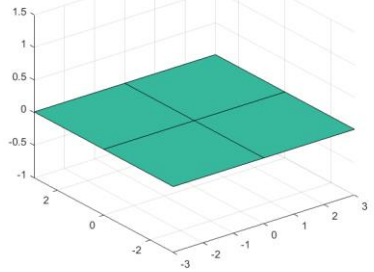
» PDE Solver



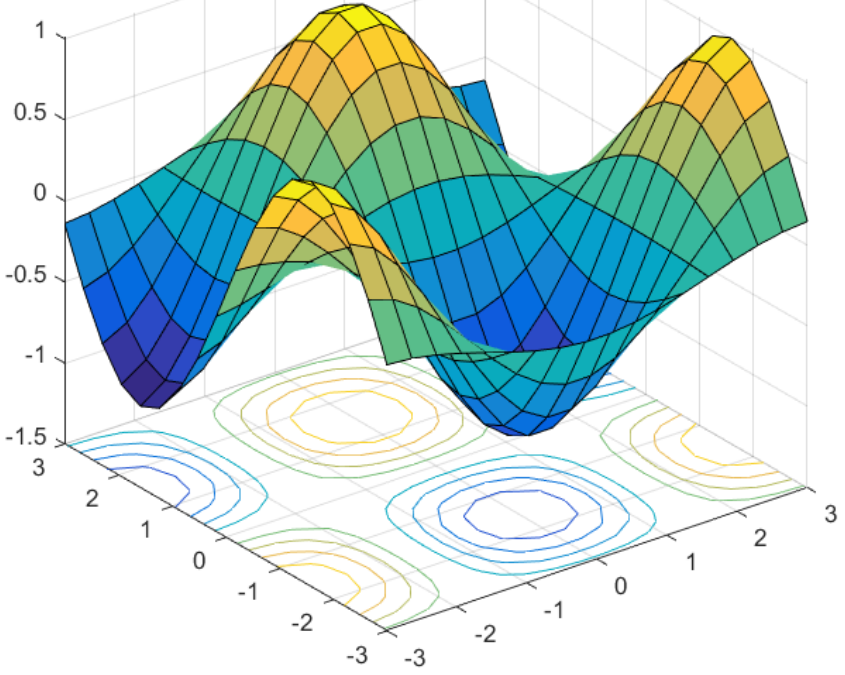
B. Digital Computing

» PDE Solver

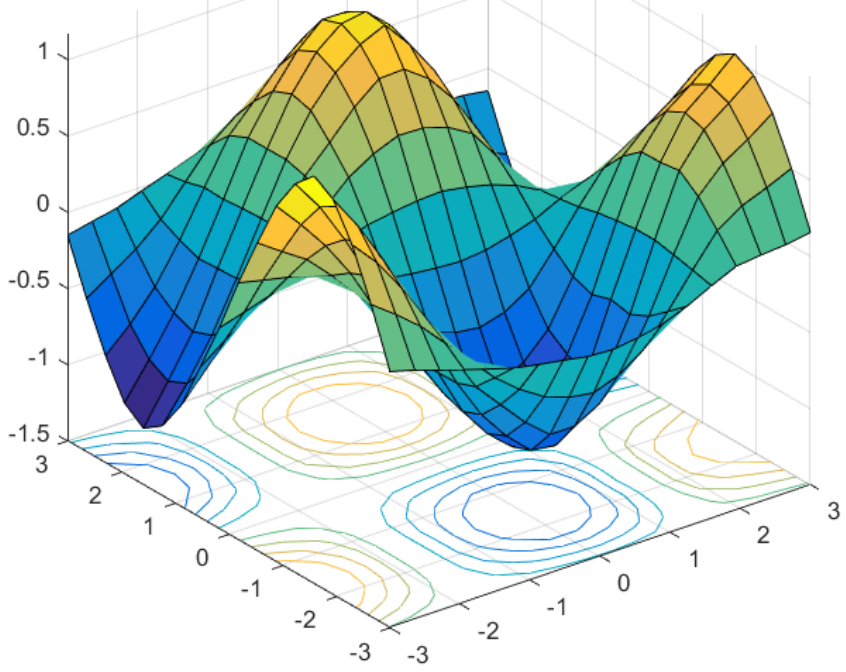
Initial Condition



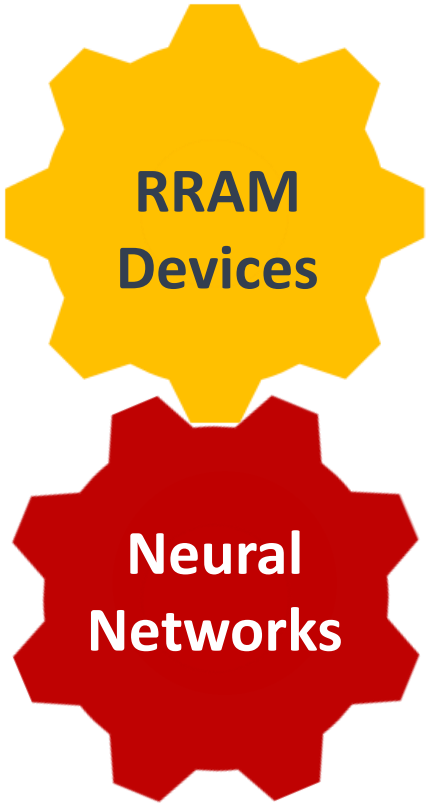
Floating Point Solver



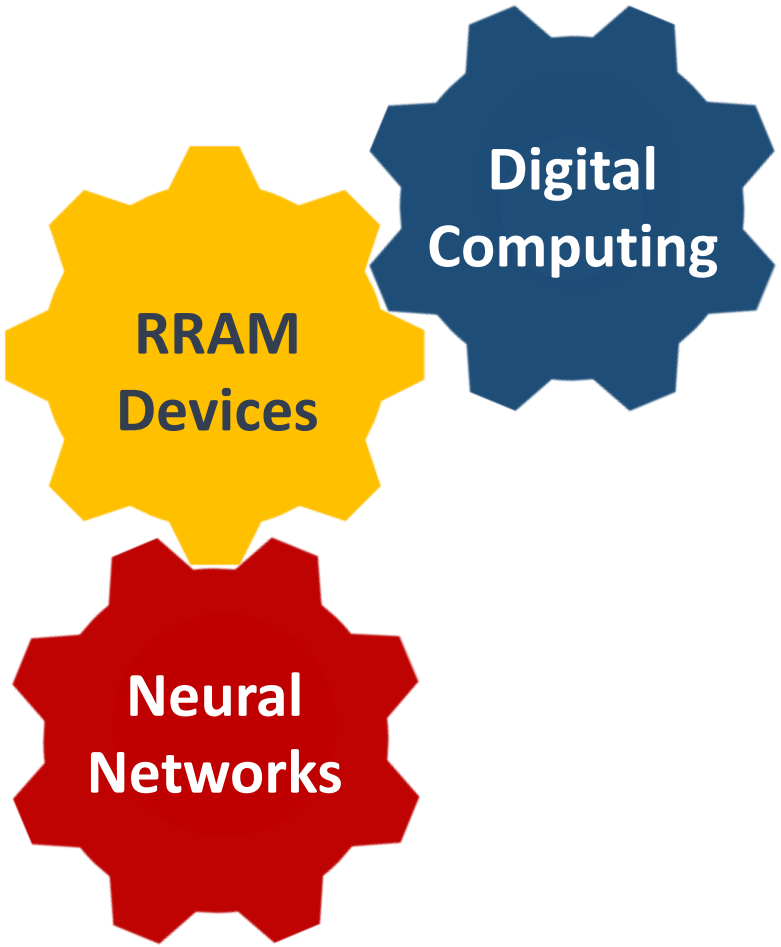
Measured Results



Reconfigurable RRAM Computing

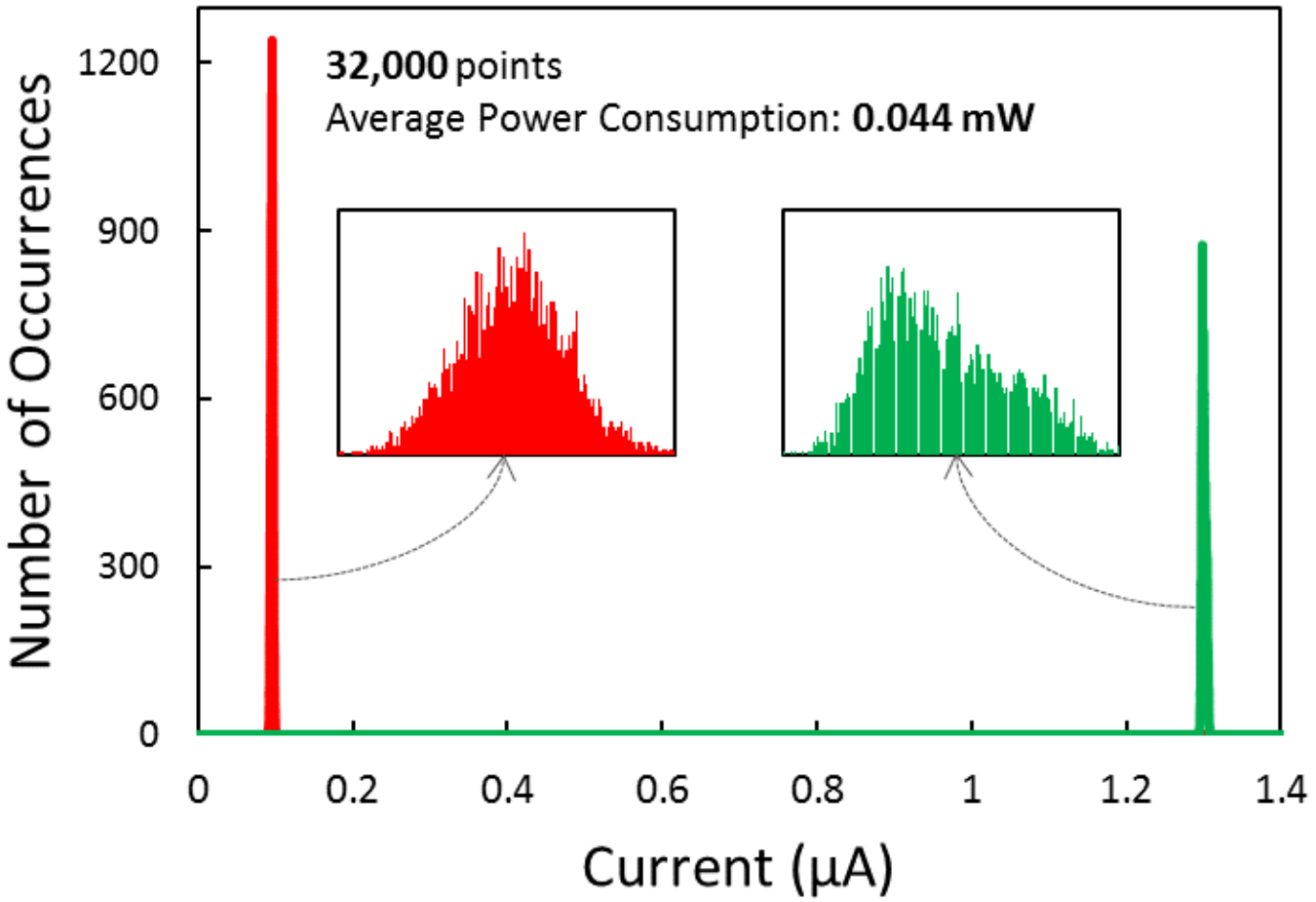


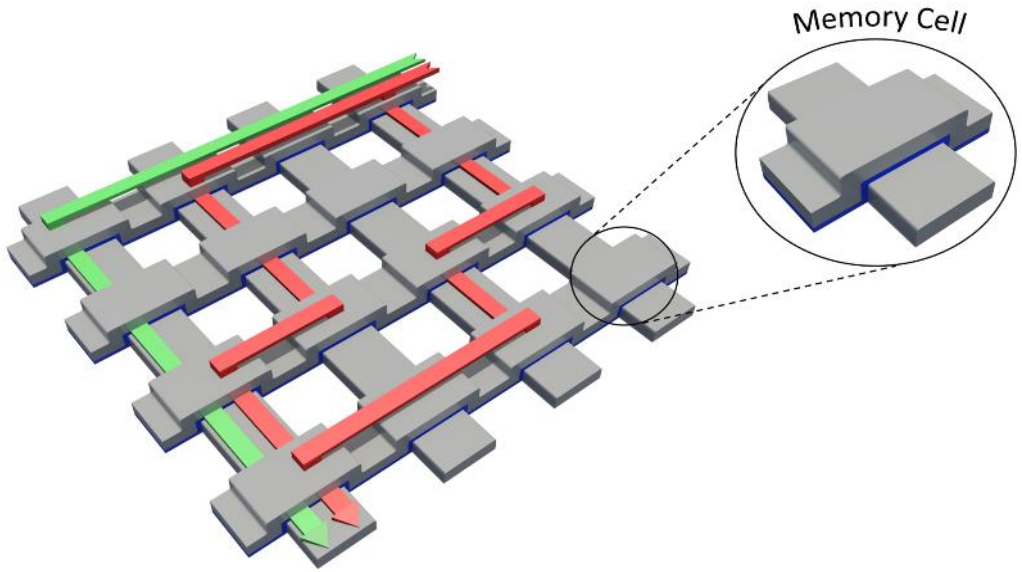
Reconfigurable RRAM Computing



Memory / Data Storage

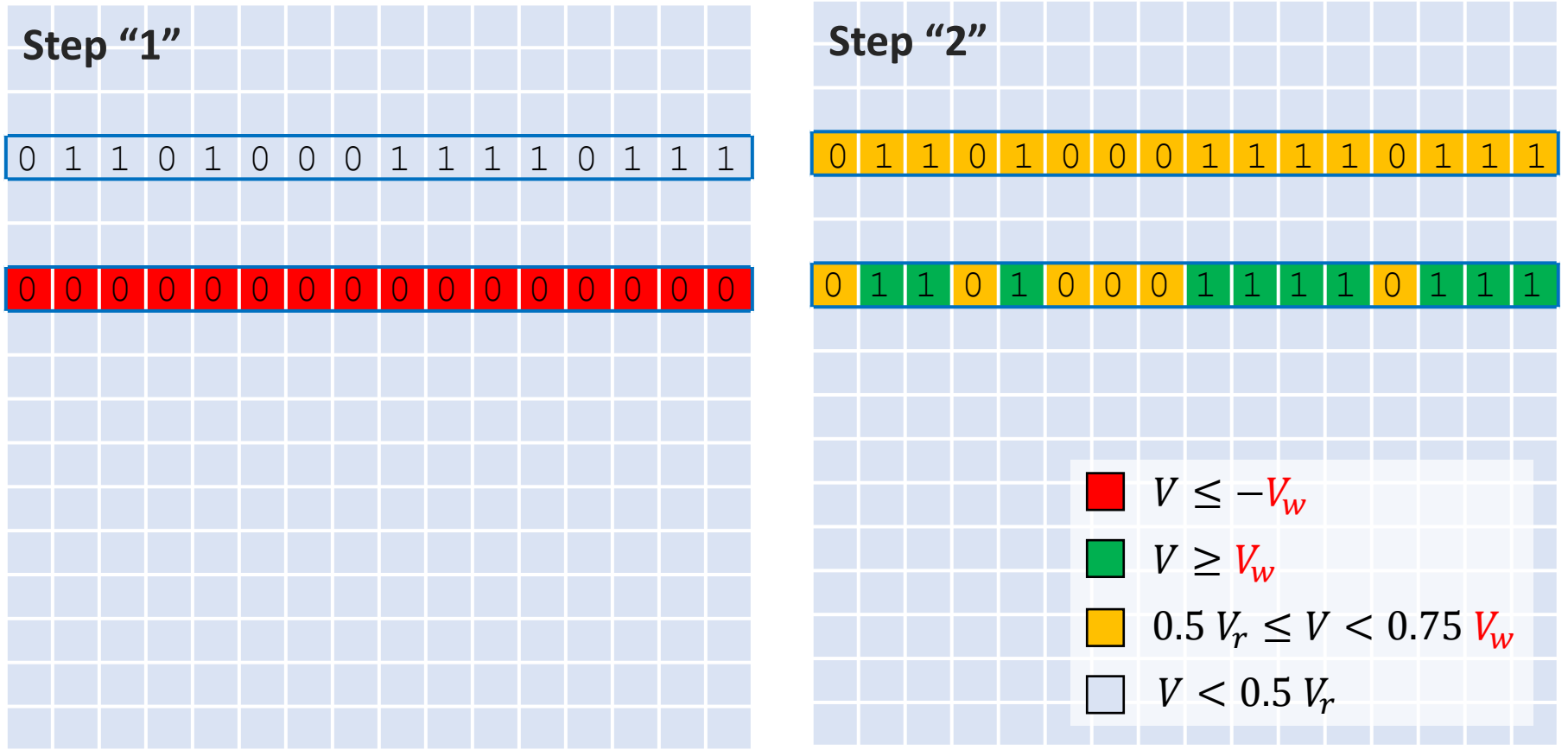
» Reliable Storage



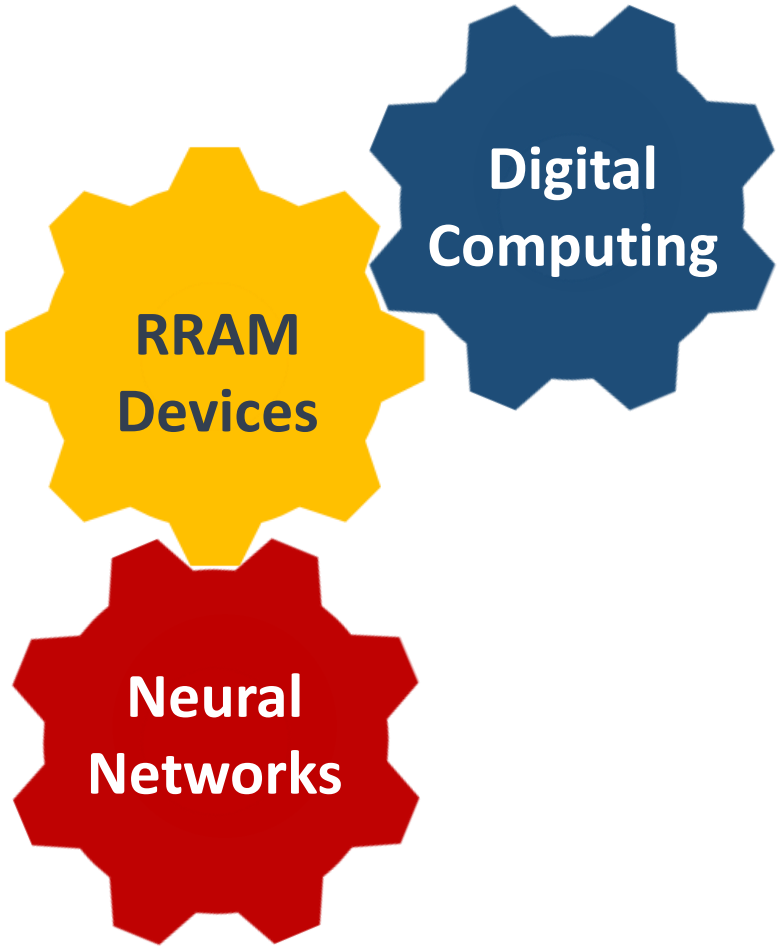


In-Situ Data Migration

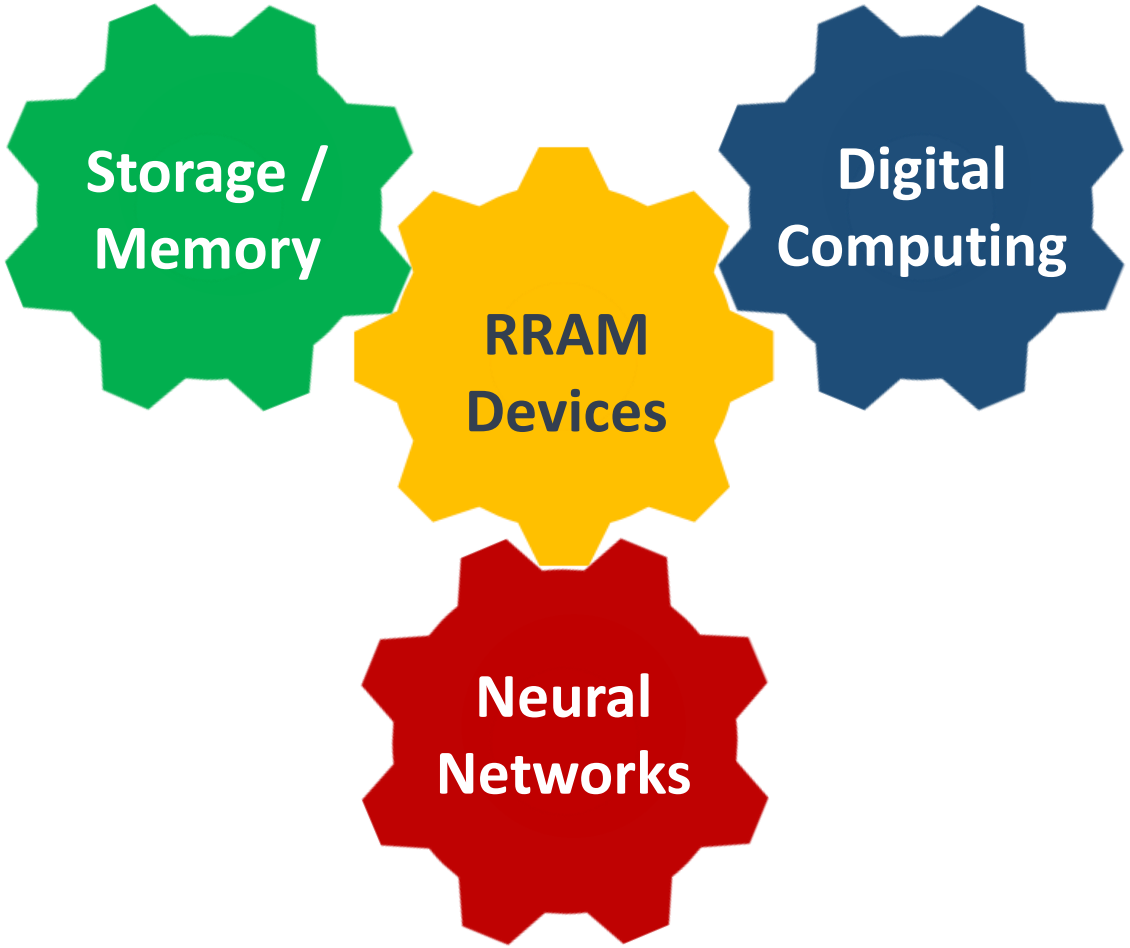
» In-Situ Data Migration



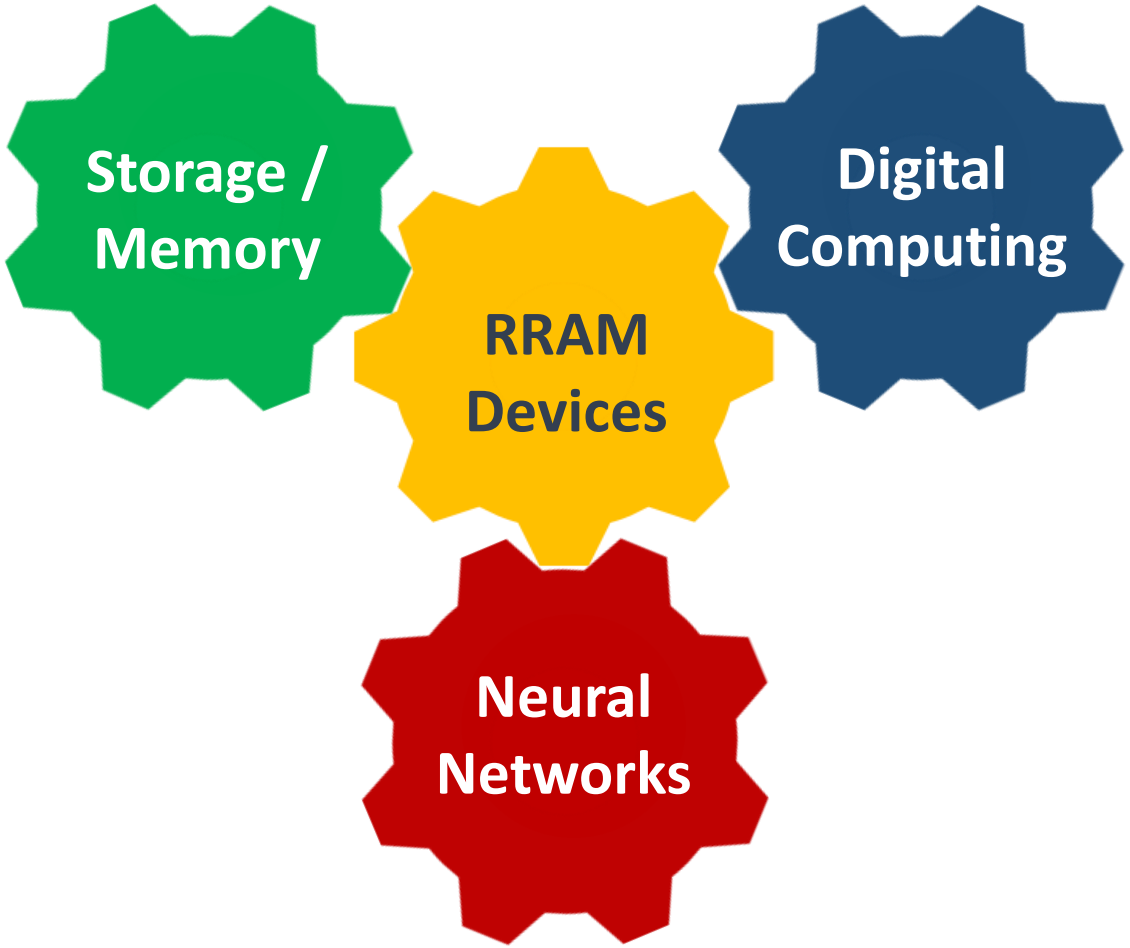
Reconfigurable RRAM Computing



Summary

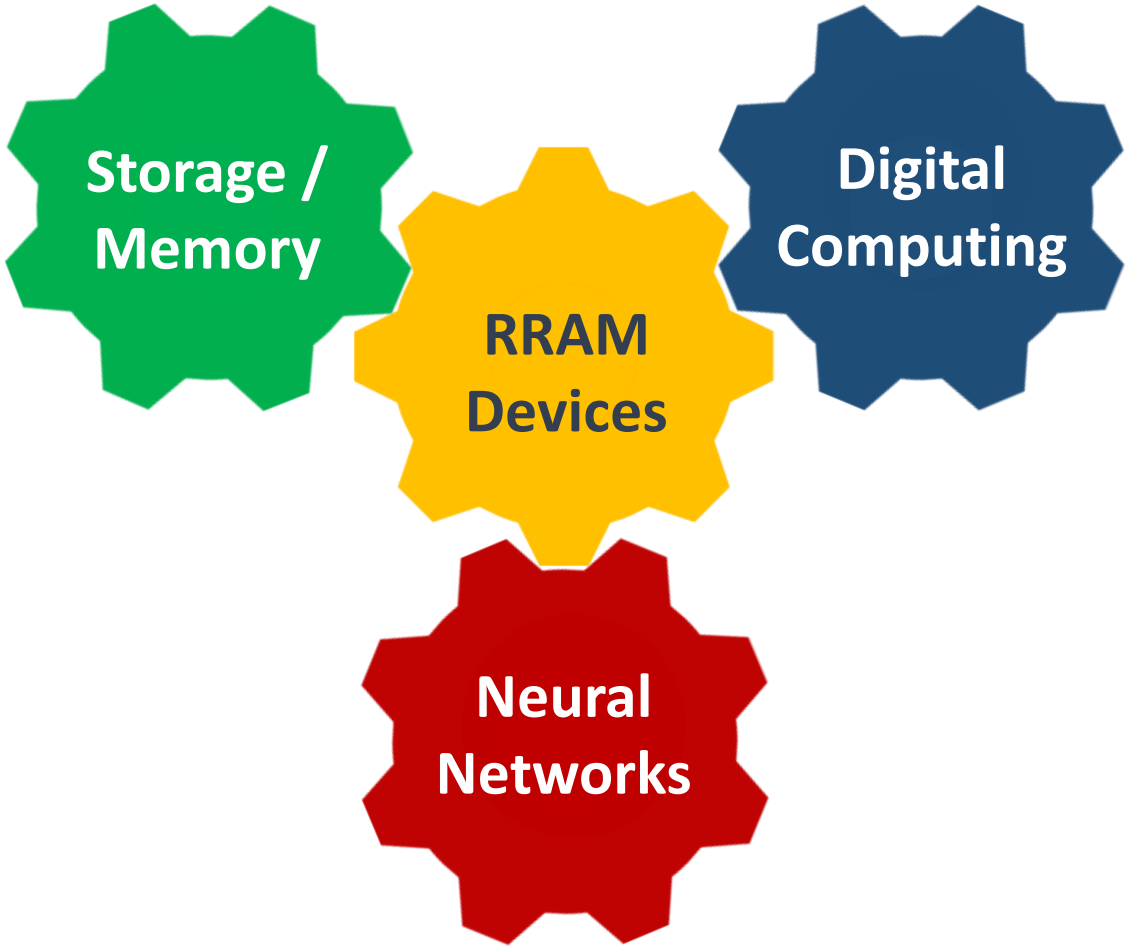

Storage / Memory
RRAM Devices
Digital Computing
Neural Networks

Reconfigurable RRAM Computing



New Computing Devices	✓
No Memory Bottleneck	✓
Classical Process	✓
Cognitive Process	✓
Low Power Consumption	✓
Scalable	✓

Reconfigurable RRAM Computing



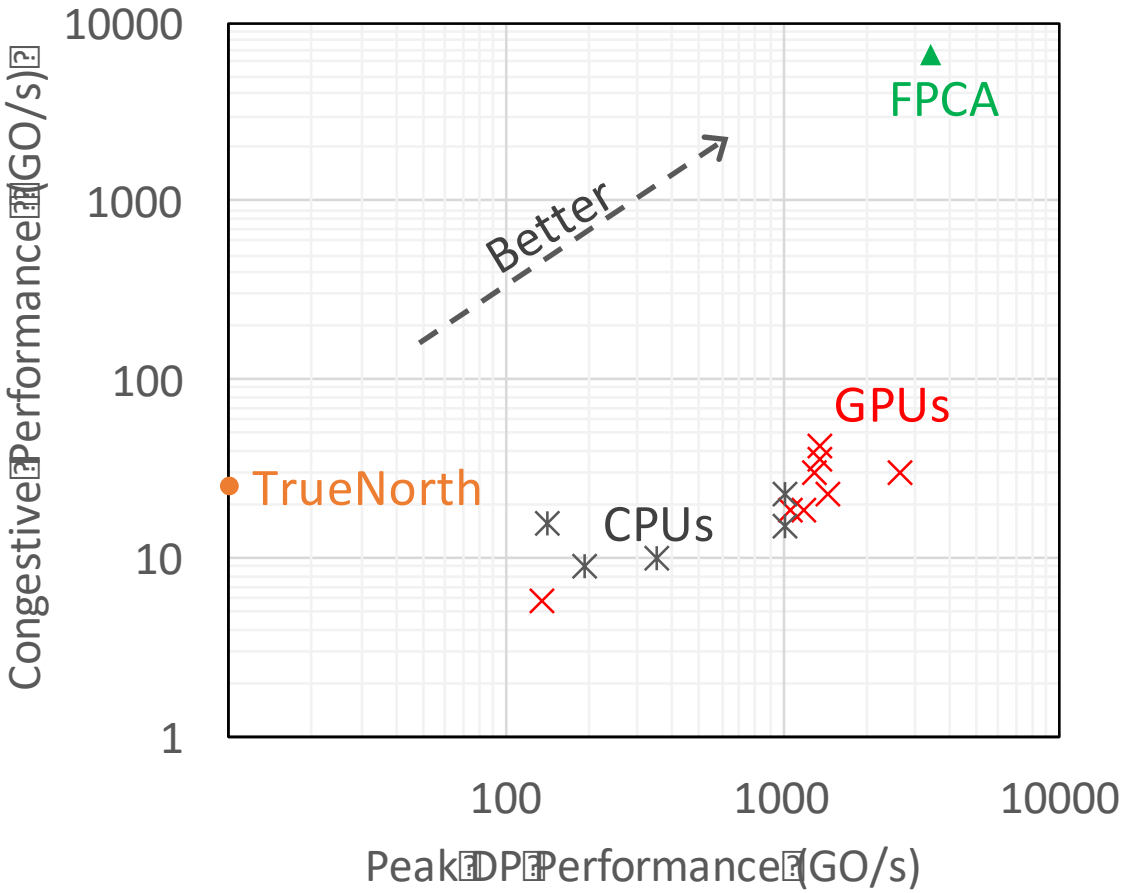


Photo by Mohammed Zidan - 2013