Computing in/using memory

Asif Ali Khan

Fall Semester 2024

Department of Computer Systems Engineering

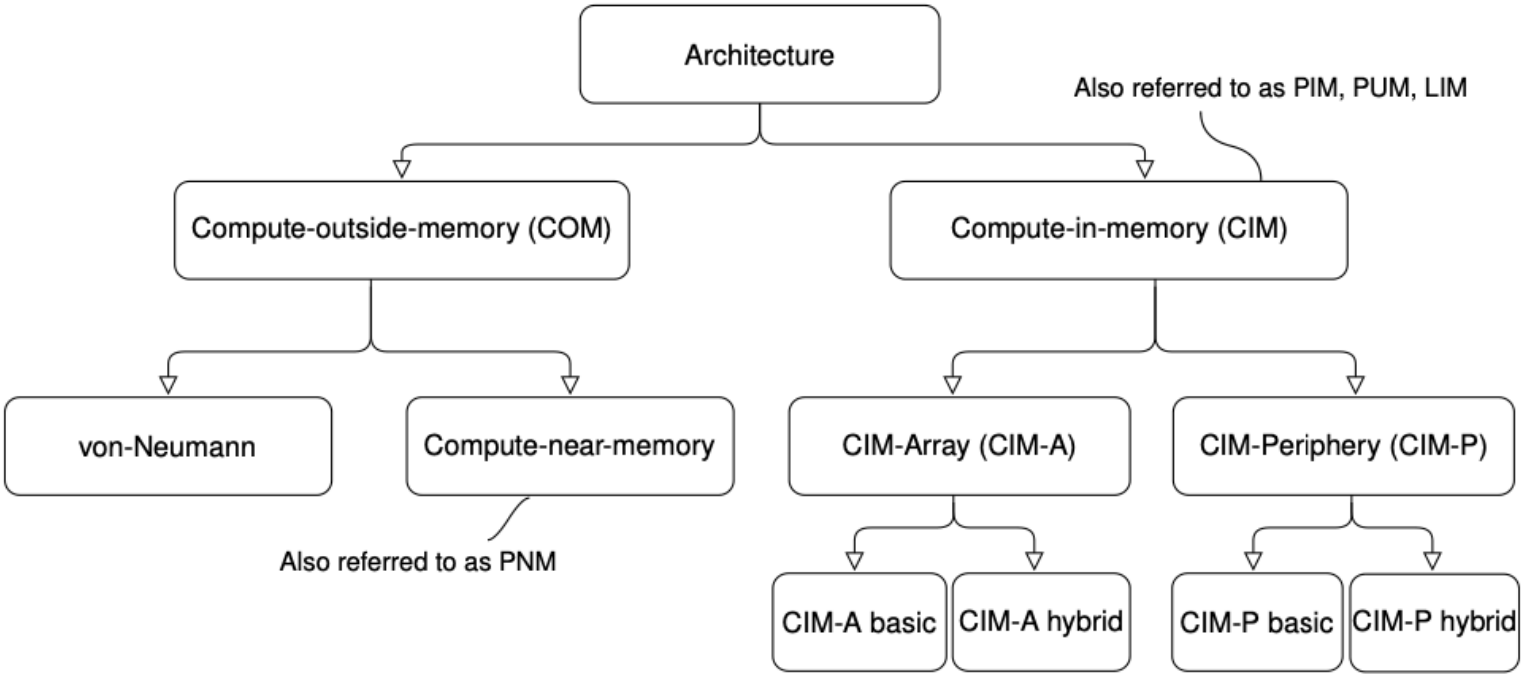
UET Peshawar, Pakistan

Nov 7, 2024

Recap: Near-memory computing architectures

- Compute-near-memory tries to mitigate the data movement over the external bus by integrating small compute units on/closer to the memory chips
- UPMEM, is a commercially available, general-purpose CNM system
- Samsung and SK Hynix have developed ML-specific CNM systems
- These systems come with their software stacks
- But their programmability is still challenging
- High-level compilation flows, e.g., Cinnamon, target lowering high-level representations to these emerging architectures

Terminology overview



Do read: Khan et al., "The Landscape of Compute-near-memory and Compute-in-memory: A Research and Commercial Overview", Arxiv 2024

Compute-in-memory (CIM)

- The CIM paradigm aims to completely eliminate the data movement in the system

Compute-in-memory (CIM)

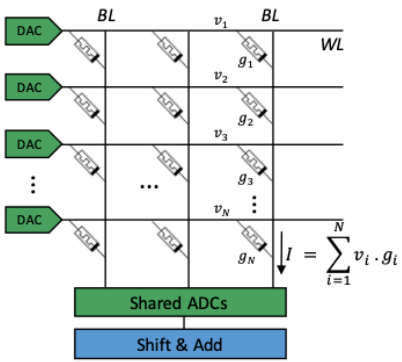
- The CIM paradigm aims to completely eliminate the data movement in the system
- The fundamental idea is to exploit the physical properties of the memory devices to perform computations

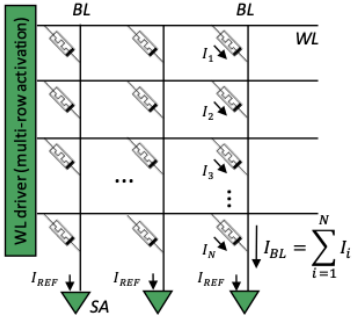
Compute-in-memory (CIM)

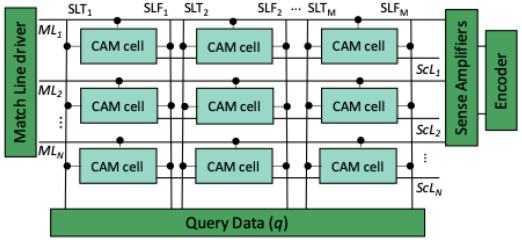
- The CIM paradigm aims to completely eliminate the data movement in the system
- The fundamental idea is to exploit the physical properties of the memory devices to perform computations
- Not every computation can be performed with every technology

Compute-in-memory (CIM)

- The CIM paradigm aims to completely eliminate the data movement in the system
- The fundamental idea is to exploit the physical properties of the memory devices to perform computations
- Not every computation can be performed with every technology

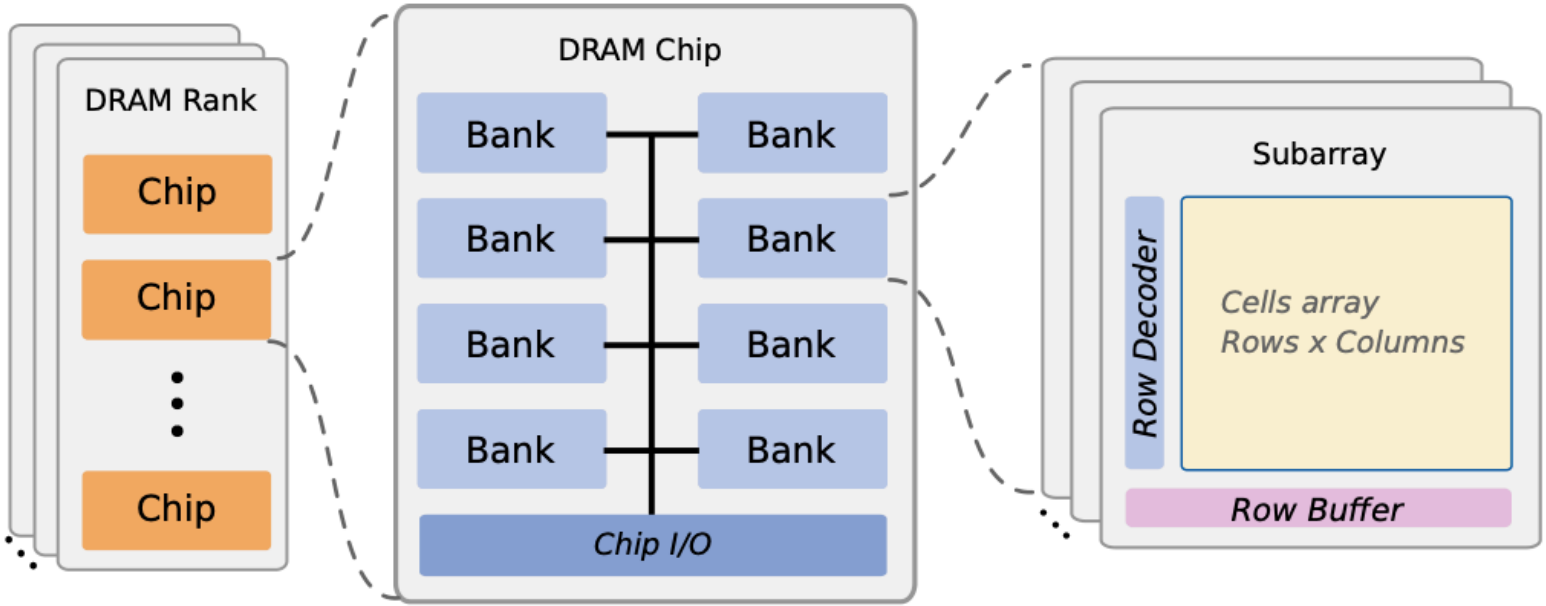






Khan et al., Arxiv 2024

The memory system organization



Khan et al., Arxiv 2024

DRAM and SRAM technologies

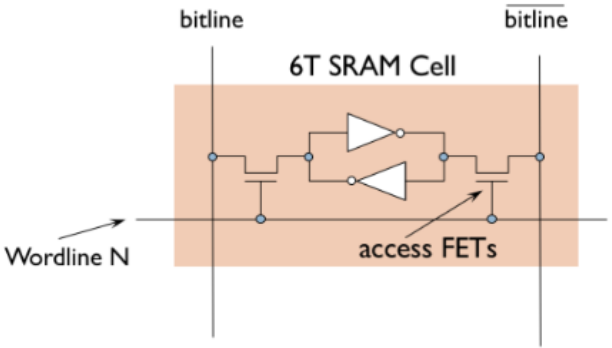
- The most mature and widely used memory technologies

DRAM and SRAM technologies

- The most mature and widely used memory technologies
- Have scaled nicely until recently

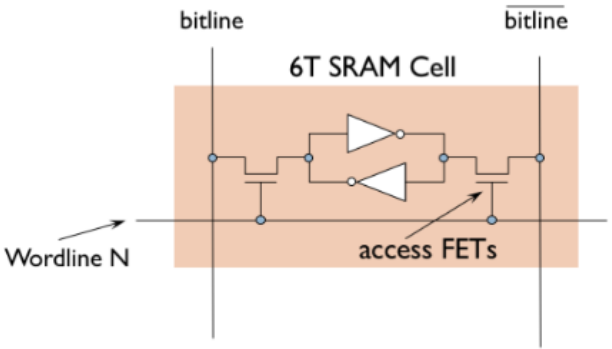
DRAM and SRAM technologies

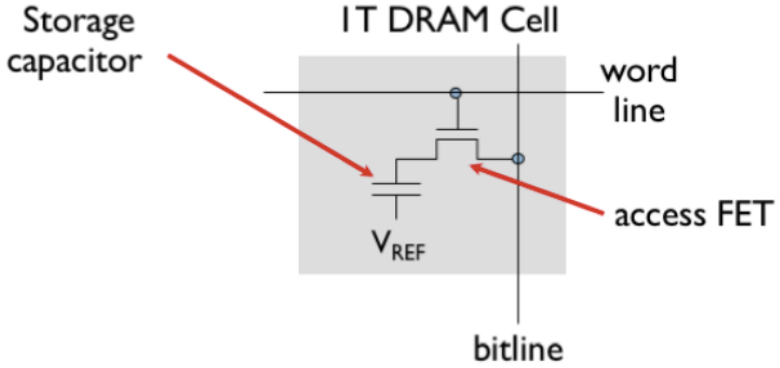
- The most mature and widely used memory technologies
- Have scaled nicely until recently



DRAM and SRAM technologies

- The most mature and widely used memory technologies
- Have scaled nicely until recently





SRAM and DRAM functionality

- SRAM read operation
  - Precharge all bitlines to Vdd and leave them floating

SRAM and DRAM functionality

- SRAM read operation
  - Precharge all bitlines to Vdd and leave them floating
  - Activate the wordline

SRAM and DRAM functionality

- SRAM read operation
  - Precharge all bitlines to Vdd and leave them floating
  - Activate the wordline
  - Each cell then pulls-down one of the bit-lines (depending on the cell value)

SRAM and DRAM functionality

- SRAM read operation
  - Precharge all bitlines to Vdd and leave them floating
  - Activate the wordline
  - Each cell then pulls-down one of the bit-lines (depending on the cell value)

- DRAM read operation
  - Precharge all bitlines to Vdd/2

SRAM and DRAM functionality

- SRAM read operation
  - Precharge all bitlines to Vdd and leave them floating
  - Activate the wordline
  - Each cell then pulls-down one of the bit-lines (depending on the cell value)

- DRAM read operation
  - Precharge all bitlines to Vdd/2
  - Activate wordline

SRAM and DRAM functionality

- SRAM read operation
  - Precharge all bitlines to Vdd and leave them floating
  - Activate the wordline
  - Each cell then pulls-down one of the bit-lines (depending on the cell value)

- DRAM read operation
  - Precharge all bitlines to Vdd/2
  - Activate wordline
  - Capacitor and bitline share charge
    - If capacitor is charged, bitlines voltage increases, i.e., Vdd/2 + δ
    - If capacitor is discharged, bitline's voltage becomes Vdd/2 - δ

SRAM and DRAM functionality

- SRAM read operation
  - Precharge all bitlines to Vdd and leave them floating
  - Activate the wordline
  - Each cell then pulls-down one of the bit-lines (depending on the cell value)

- DRAM read operation
  - Precharge all bitlines to Vdd/2
  - Activate wordline
  - Capacitor and bitline share charge
    - If capacitor is charged, bitlines voltage increases, i.e., Vdd/2 + δ
    - If capacitor is discharged, bitline's voltage becomes Vdd/2 - δ
  - Senseamps sense the difference to determine 1 or 0

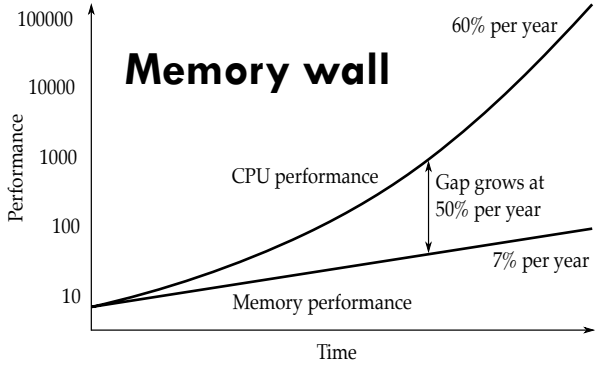
SRAM and DRAM functionality

- SRAM read operation
  - Precharge all bitlines to Vdd and leave them floating
  - Activate the wordline
  - Each cell then pulls-down one of the bit-lines (depending on the cell value)

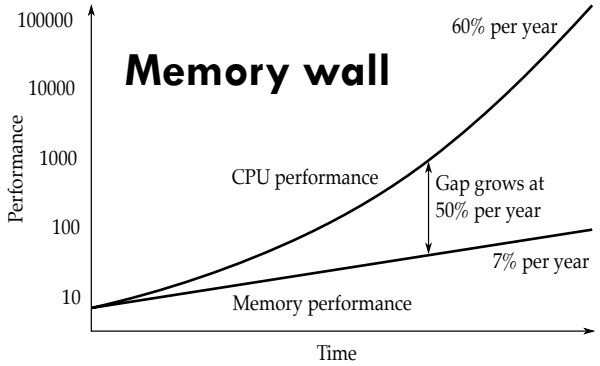
- DRAM read operation
  - Precharge all bitlines to Vdd/2
  - Activate wordline
  - Capacitor and bitline share charge
    - If capacitor is charged, bitlines voltage increases, i.e., Vdd/2 + δ
    - If capacitor is discharged, bitline's voltage becomes Vdd/2 - δ
  - Senseamps sense the difference to determine 1 or 0

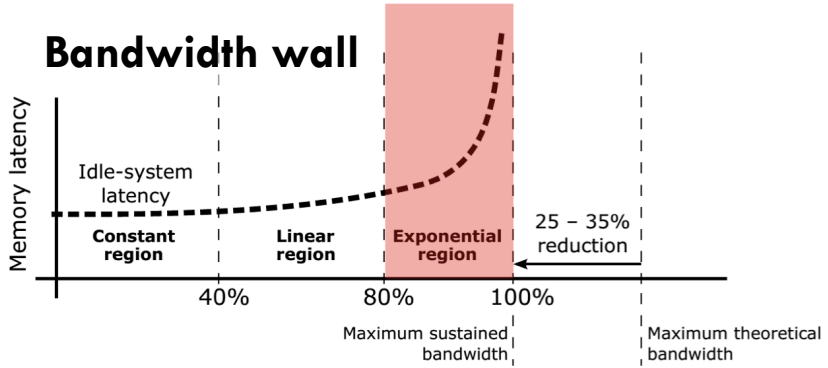
Note: Reads are destructive

Memory subsystem challenges

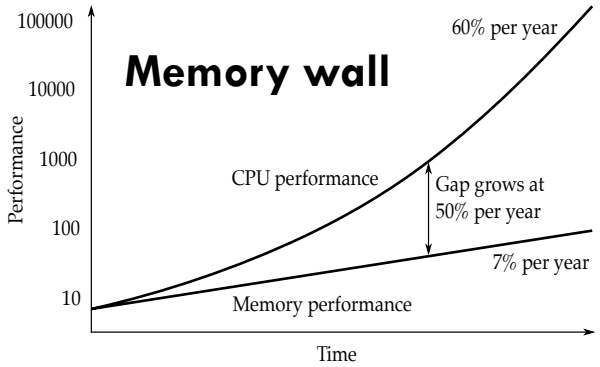


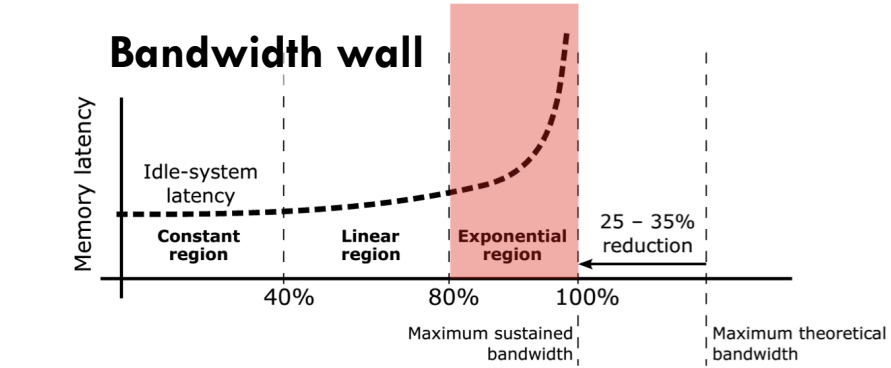
Memory subsystem challenges

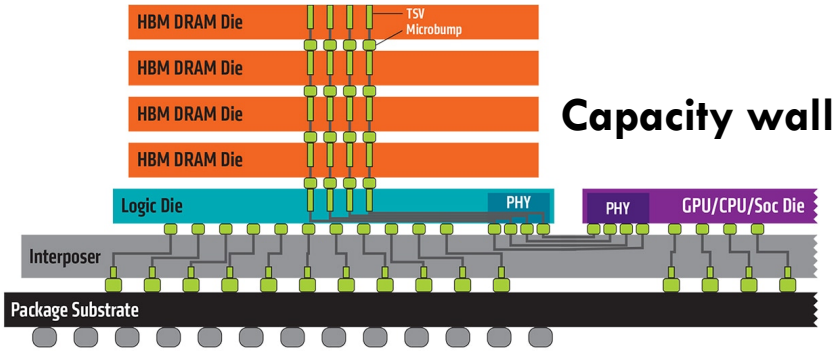




Memory subsystem challenges

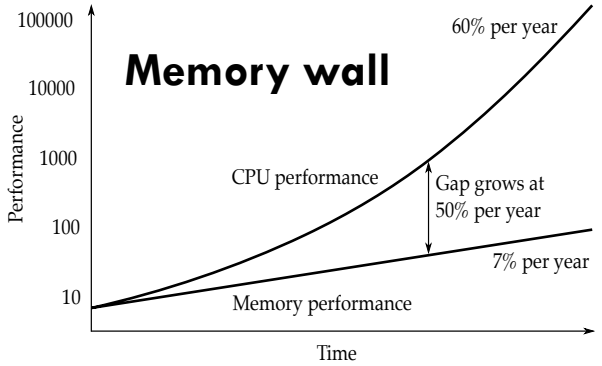


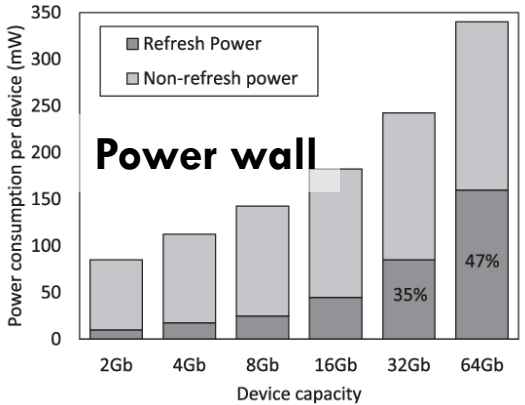




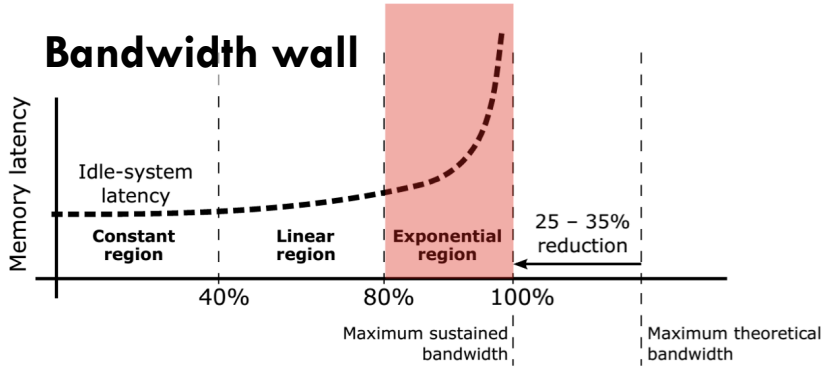
Capacity wall

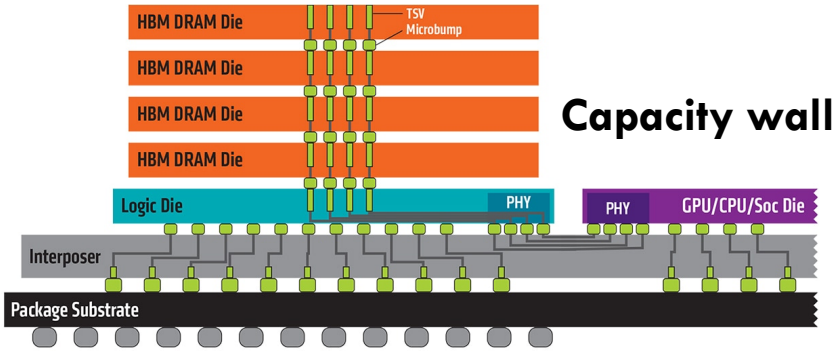
Memory subsystem challenges





Shin et al., IEEE Transaction on Computers, 2018

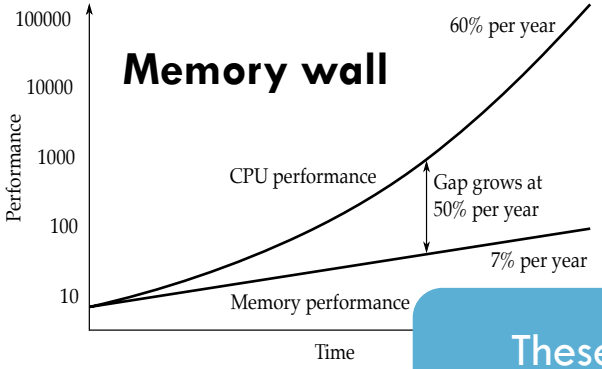


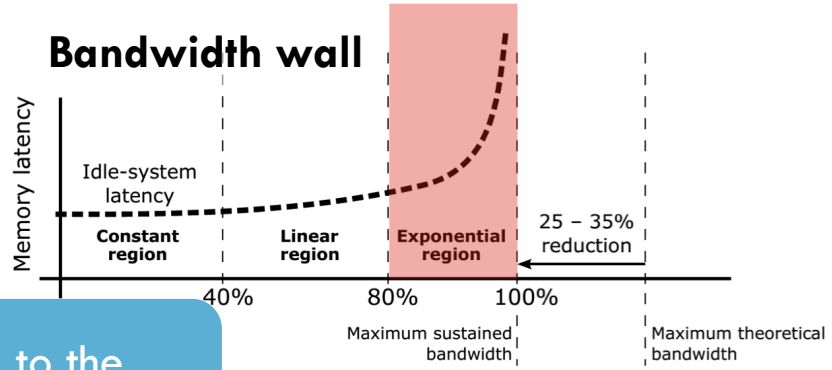


Capacity wall

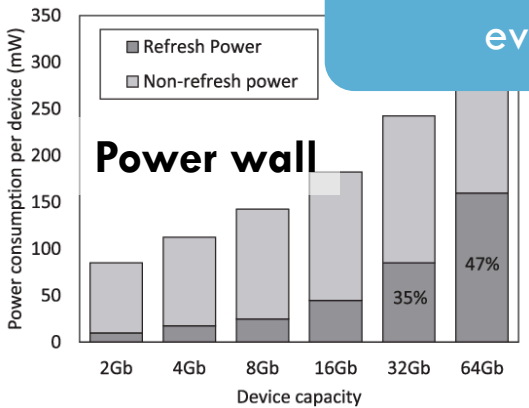
Sources: Report on the HPC application bottlenecks, ExaNoDe, 2017, AMD

Memory subsystem challenges

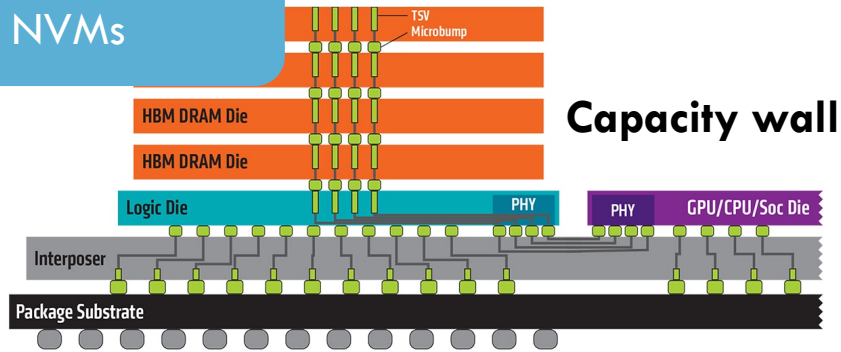




These ``walls`` led to the evolution of NVMs



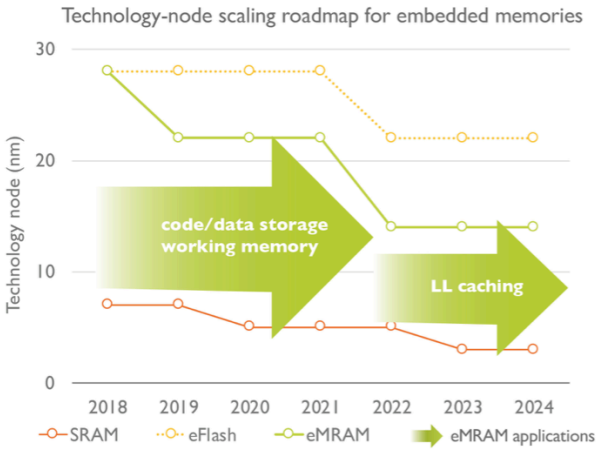
Shin et al., IEEE Transaction on Computers, 2018

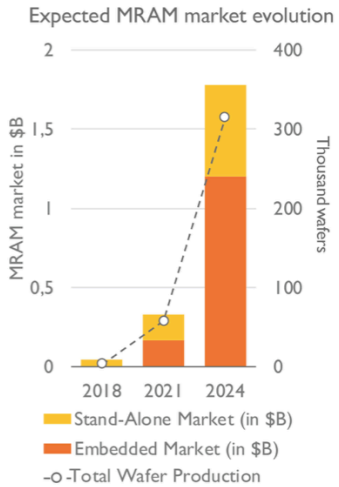


Sources: Report on the HPC application bottlenecks, ExaNoDe, 2017, AMD

The rise of nonvolatile memories (NVMs)

- Momentum is building around NVMs

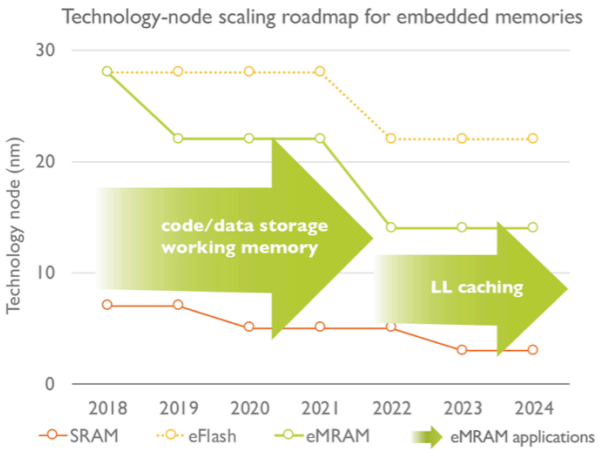


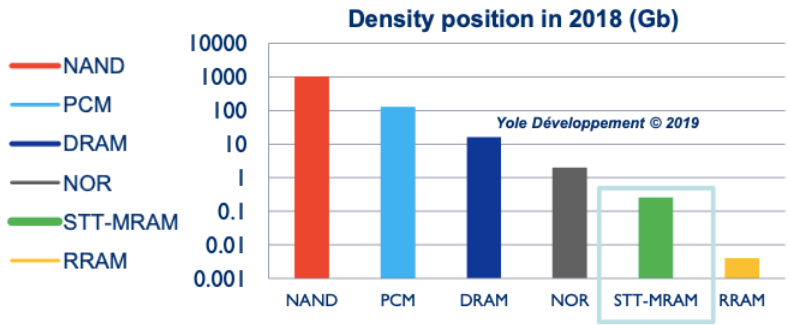


(Source: MRAM Technology and Business 2019 report, Yole Développement, 2019)

The rise of nonvolatile memories (NVMs)

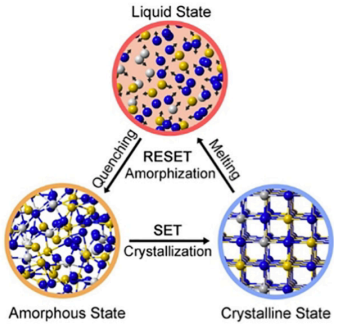
- Momentum is building around NVMs





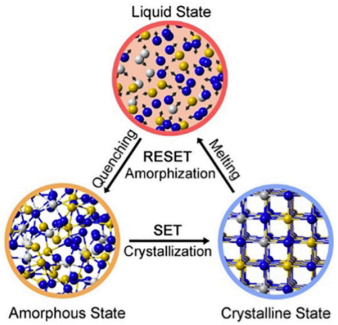
(Source: MRAM Technology and Business 2019 report, Yole Développement, 2019)

Emerging nonvolatile memories (NVM)

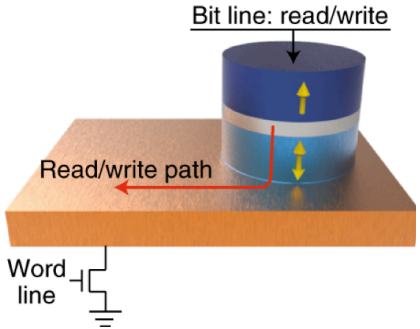


Zhang et al., 2020

Emerging nonvolatile memories (NVM)

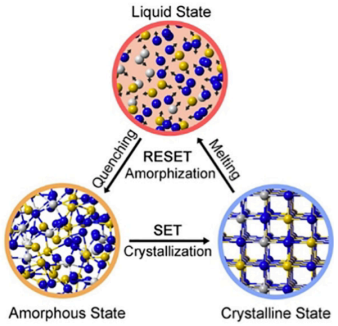


Zhang et al., 2020

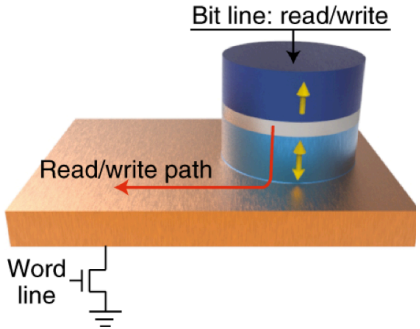


G. Yu, 2020

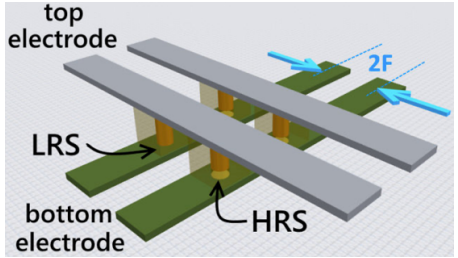
Emerging nonvolatile memories (NVM)



Zhang et al., 2020

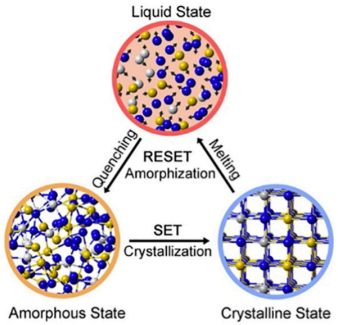


G. Yu, 2020

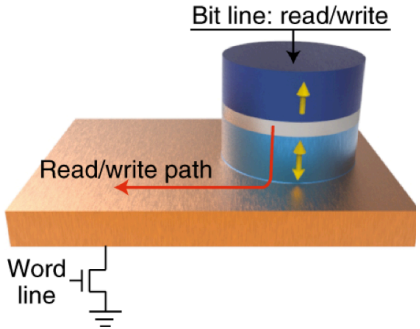


Lim et al, 2015

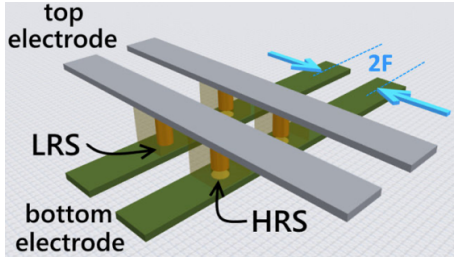
Emerging nonvolatile memories (NVM)



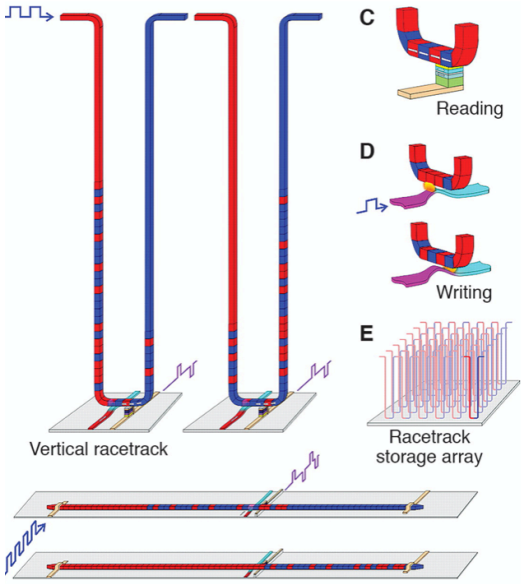
Zhang et al., 2020



G. Yu, 2020

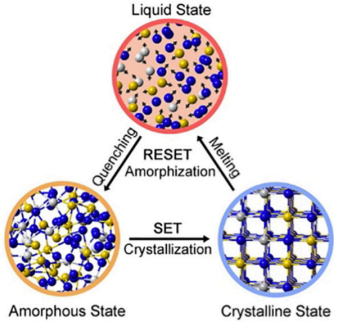


Lim et al, 2015

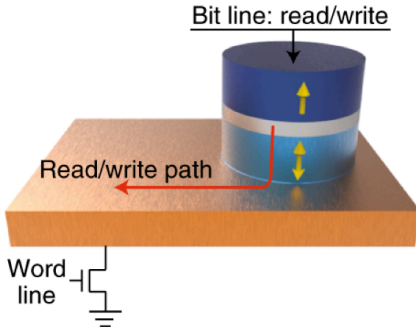


Parkin et al, 2008

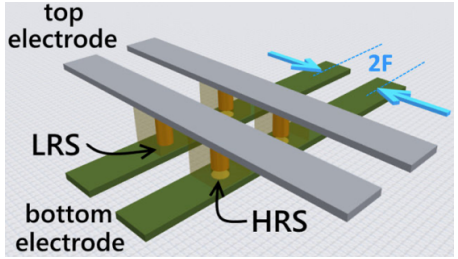
Emerging nonvolatile memories (NVM)



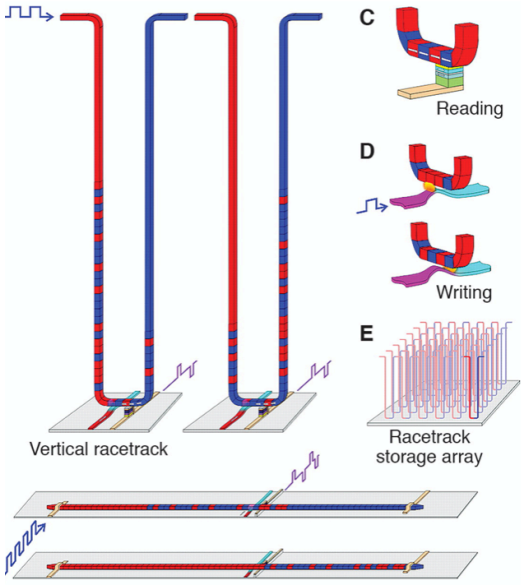
Zhang et al., 2020



G. Yu, 2020



Lim et al, 2015



Parkin et al, 2008

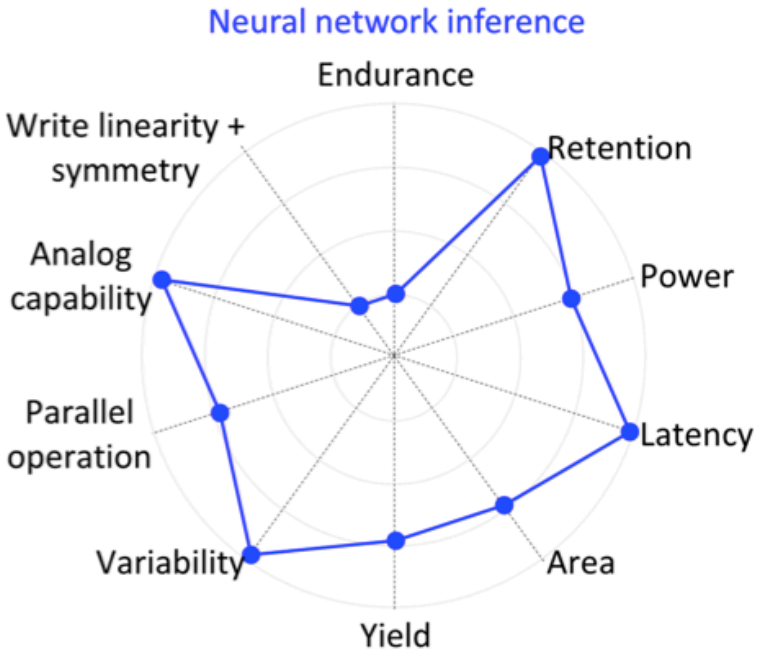
- Other options: FeFET, FeRAM etc
- Each technology has its strengths and challenges
- PCM and MRAM receive a lot of traction in industry

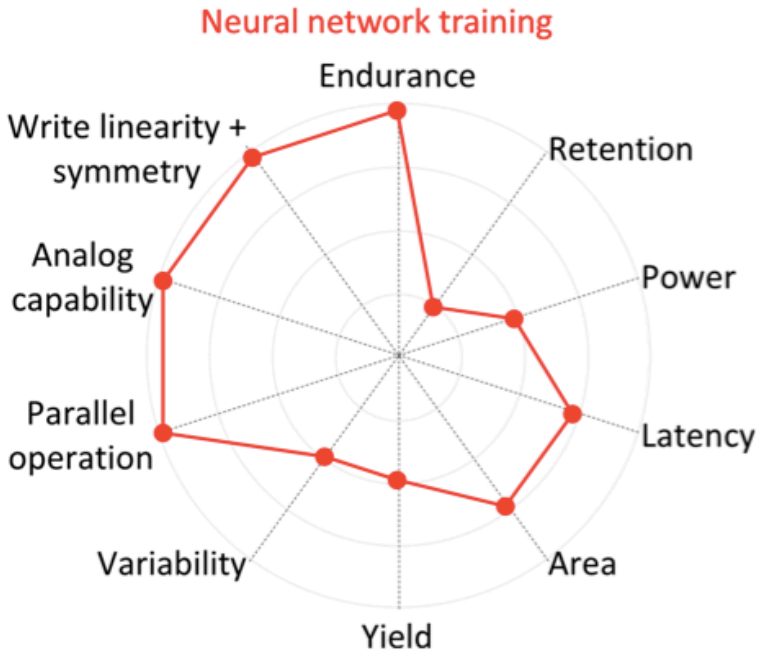
Memory technologies comparison

Device	SRAM	DRAM	RRAM	PCM	STT-MRAM	FeFET
Write time	$1-10ns$	$>20ns$	$>10ns$	$\sim 50ns$	$>10ns$	$\sim 10ns$
Read time	$1-10ns$	$>20ns$	$>10ns$	$>10ns$	$>10ns$	$\sim 10ns$
Drift	No	No	Weak	Yes	No	No
Write energy (per bit)	$1-10fJ$	$10-100fJ$	$0.1-1pJ$	$100pJ$	$\sim 100fJ$	$>1fJ$
Density	Low	Medium	High	High	Medium	High
Endurance	$>10^{16}$	$>10^{16}$	$>10^5-10^8$	$>10^5-10^8$	$>10^{15}$	$>10^{15}$
Retention	Low	Very Low	Medium	long	Medium	long

Khan et al., Arxiv 2024

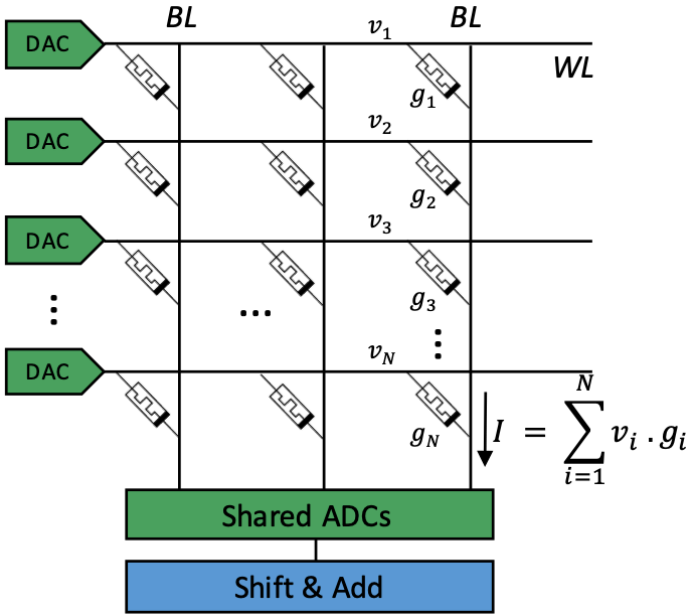
Parameter's relevance to applications





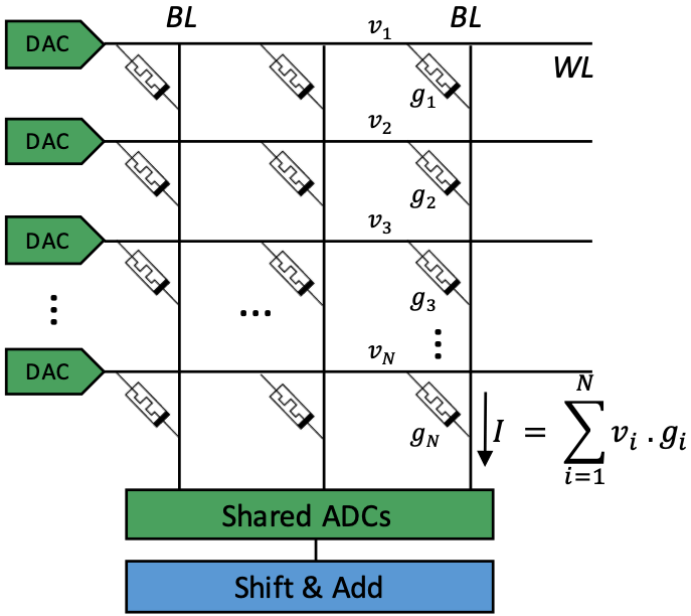
Khan et al., Arxiv 2024

CIM crossbar



Khan et al., Arxiv 2024

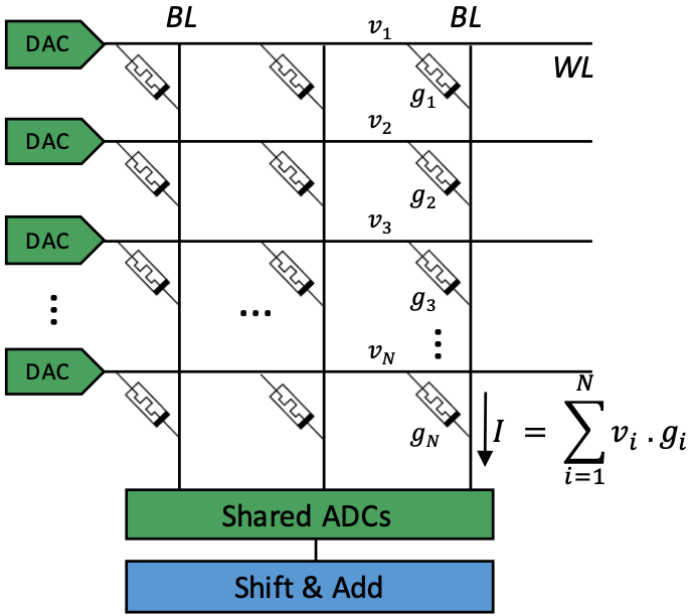
CIM crossbar



- Program one operand into memristors devices (conductance)

Khan et al., Arxiv 2024

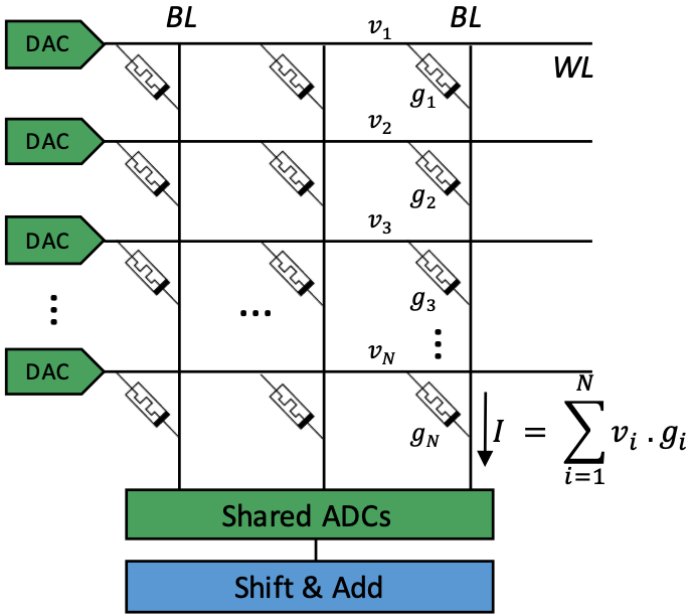
CIM crossbar



- Program one operand into memristors devices (conductance)
- Enable all wordlines simultaneously and apply another operand as input

Khan et al., Arxiv 2024

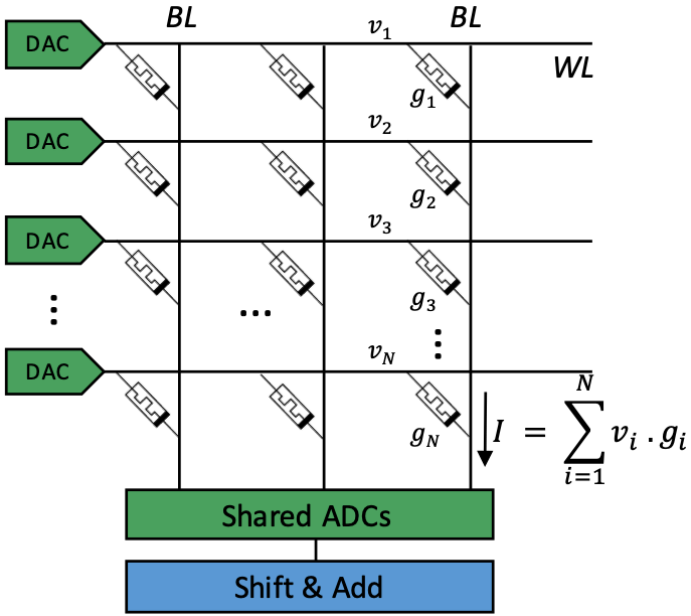
CIM crossbar



- Program one operand into memristors devices (conductance)
- Enable all wordlines simultaneously and apply another operand as input
- The accumulated current at the bitlines using kirchoff's law produces the outcome of dot product

Khan et al., Arxiv 2024

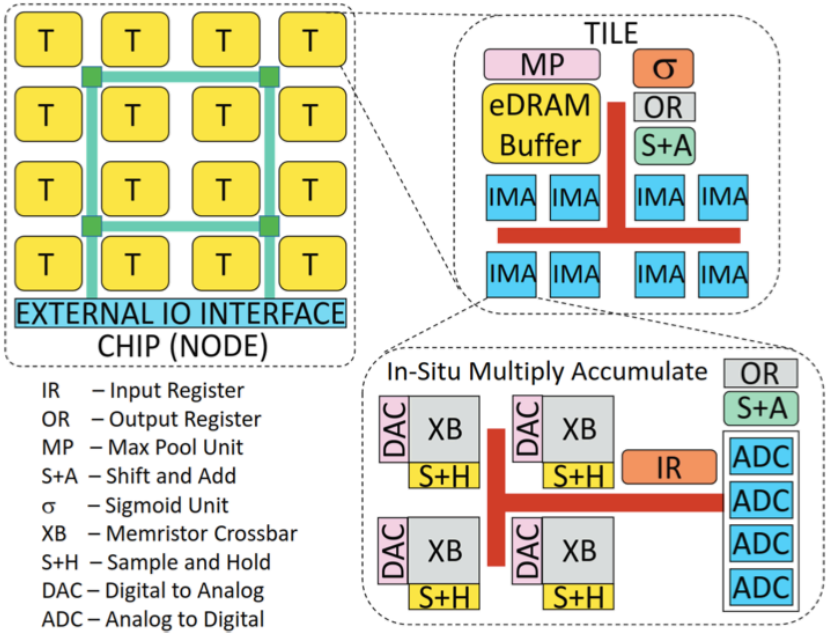
CIM crossbar



- Program one operand into memristors devices (conductance)
- Enable all wordlines simultaneously and apply another operand as input
- The accumulated current at the bitlines using kirchoff's law produces the outcome of dot product
- Is an approximation and not the accurate result

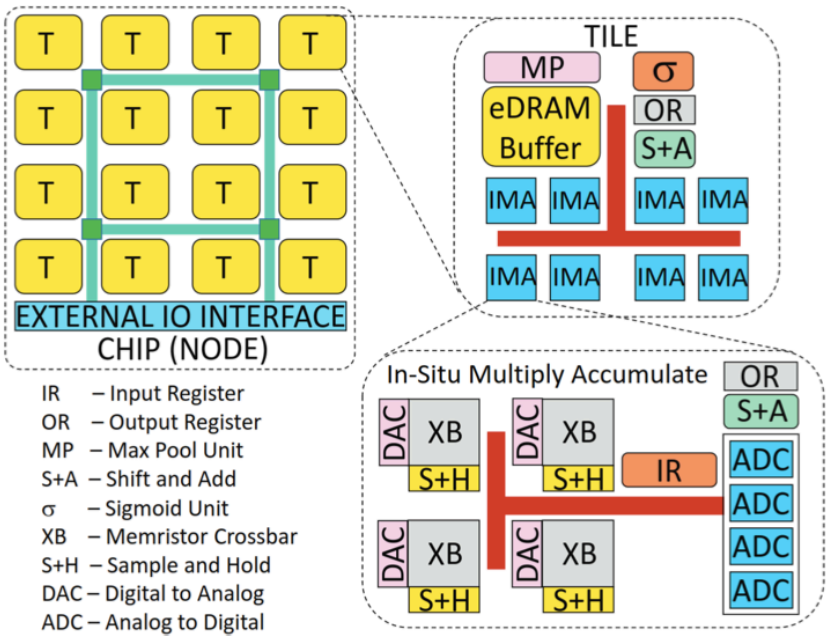
Khan et al., Arxiv 2024

ISAAC accelerator



A. Shafiee et al., "Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," ACM SIGARCH Computer Architecture News, vol. 44, no. 3, pp. 14–26, 2016.

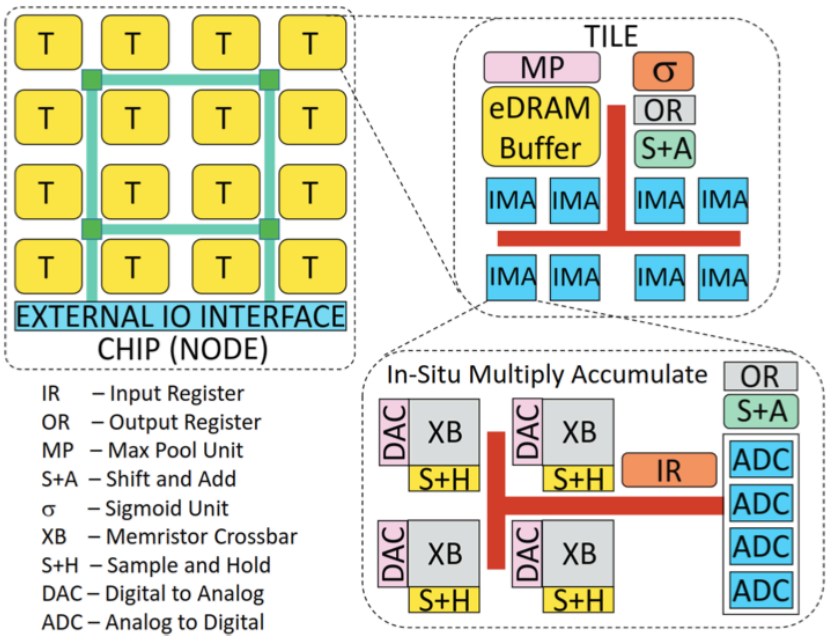
ISAAC accelerator



- One of the pioneering work from HP

A.Shafieeetal.,"Isaac:Aconvolutionalneuralnetworkacceleratorwithin-situanalogarithmeticincrossbars,"ACM SIGARCH Computer Architecture News, vol. 44, no. 3, pp. 14–26, 2016.

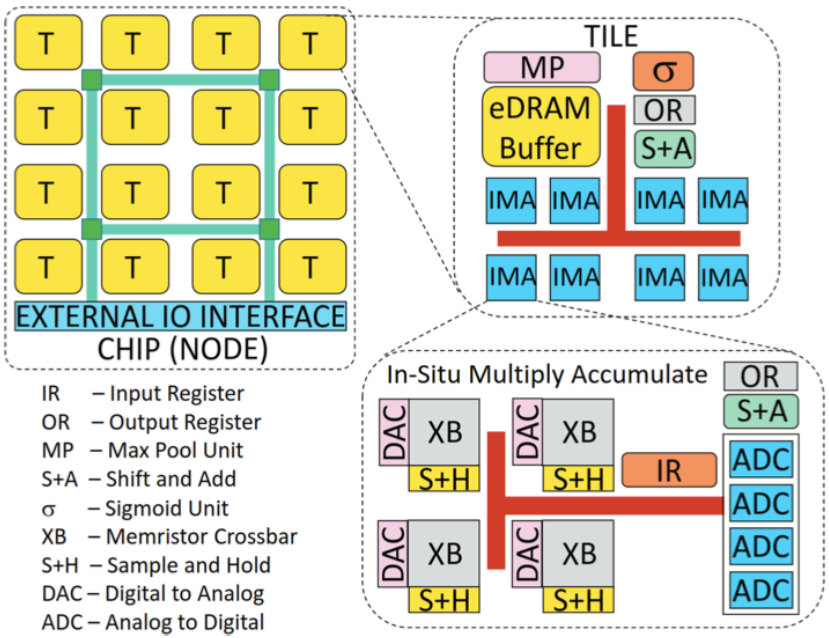
ISAAC accelerator



- One of the pioneering work from HP
- The memristive devices can store multiple bits per cell

A.Shafieeetal.,"Isaac:Aconvolutionalneuralnetworkacceleratorwithin-situanalogarithmeticincrossbars,"ACM SIGARCH Computer Architecture News, vol. 44, no. 3, pp. 14–26, 2016.

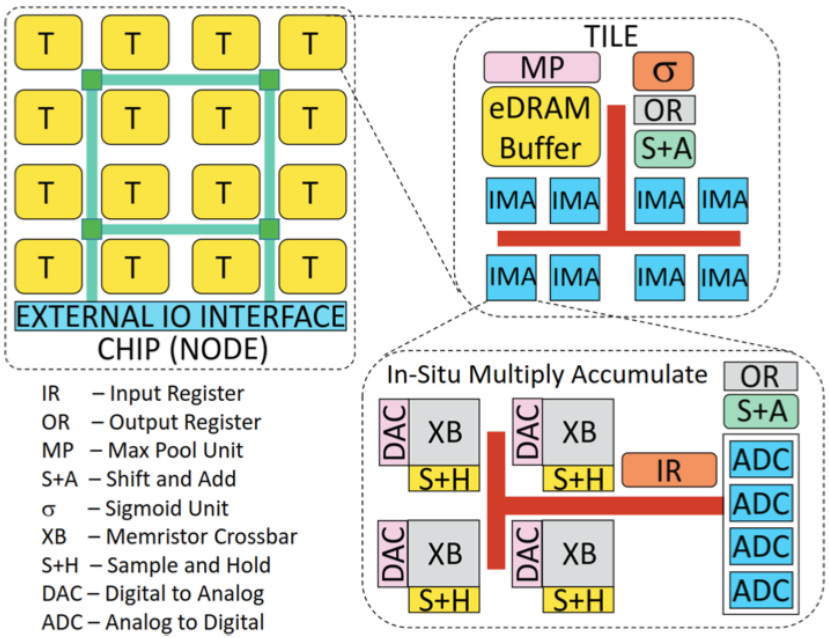
ISAAC accelerator



- One of the pioneering work from HP
- The memristive devices can store multiple bits per cell
- For higher precision input, it needs be bit-sliced and stored in multiple columns

A.Shafieeetal.,"Isaac:Aconvolutionalneuralnetworkacceleratorwithin-situanalogarithmeticincrossbars,"ACM SIGARCH Computer Architecture News, vol. 44, no. 3, pp. 14–26, 2016.

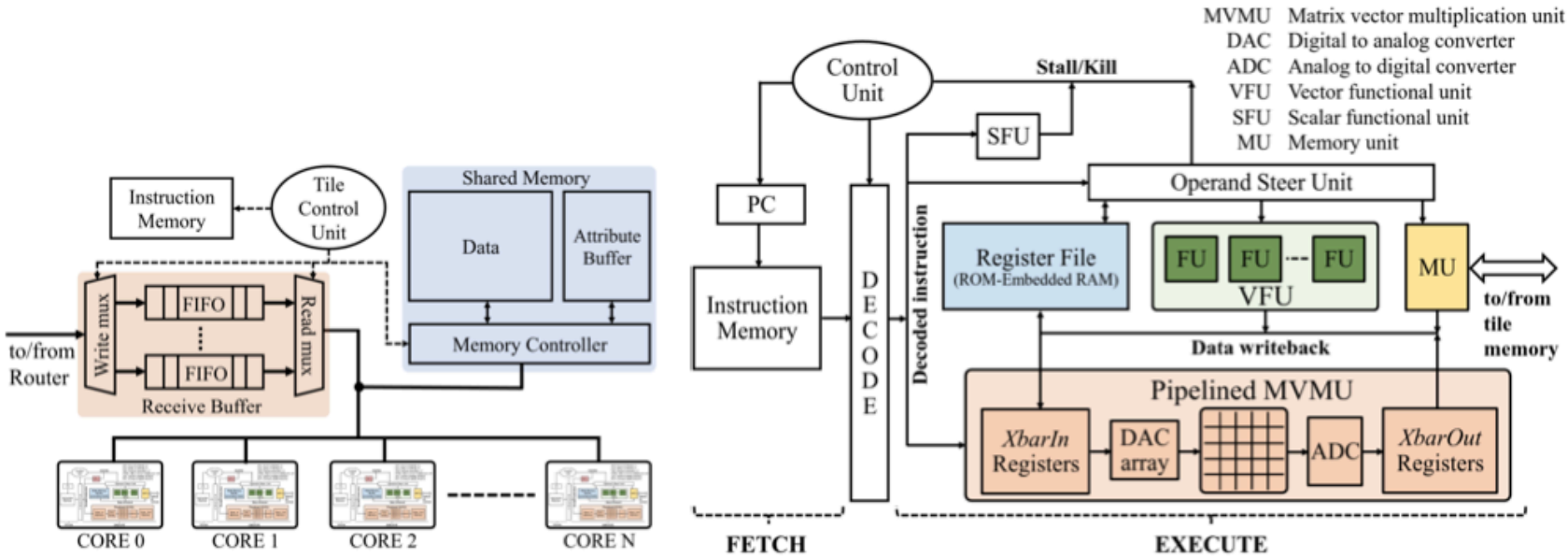
ISAAC accelerator



- One of the pioneering work from HP
- The memristive devices can store multiple bits per cell
- For higher precision input, it needs be bit-sliced and stored in multiple columns
- The shift-and-add (S+A) circuitry then takes care of the accumulation

A.Shafieeetal.,"Isaac:Aconvolutionalneuralnetworkacceleratorwithin-situanalogarithmeticincrossbars,"ACM SIGARCH Computer Architecture News, vol. 44, no. 3, pp. 14–26, 2016.

The PUMA architecture



(a) PUMA's tile architecture

(b) PUMA's core architecture

Ankit et al., ASPLOS 2019

Thank you!

asif.ali@uetpeshawar.edu.pk