CSCE 5013 Domain Specific Accelerators
Emerging NVM Technologies
David Andrews

Must...Have...More...Memory ☺

Machine Learning and Neuromorphic Architectures Need Memory

Processor in/near Memory (PIM) Architectures can scale processing with Memory Capacity.

CMOS compatible Non-Volatile Memories are Emerging.....

The Big Three

PCM Phase Change Memory

-Intel/Micron Optane/3D XPoint

MRAM Magnetic RAM (STT-RAM)

-Avalanche
-Everspin
-Samsung

RRAM Resistive RAM -> Memristors

-Rambus
-Fujitsu
-Panasonic

-Adesto
-Crossbar

RRAM

Resistive RAM: Can be ambiguously used as umbrella term referring to any device that vary resistance (MRAM, PCM).

Tighter definition: oxygen vacancy memory (OxRAM) and conductive bridging memory (CBRAM).

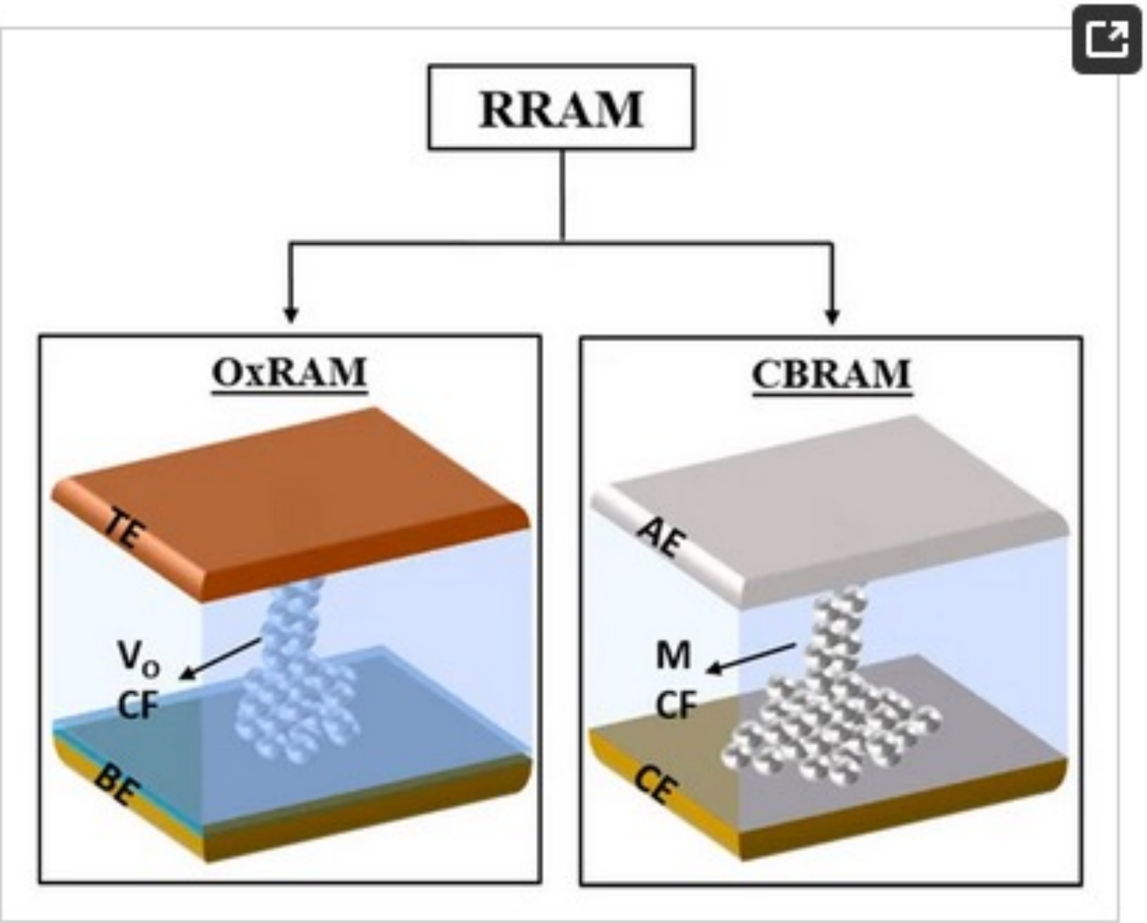
OxRAM: oxygen ions get removed by current in one direction and replaced when current is opposite direction.

-Removing oxygen ions -> decreases conductivity
-Replacing oxygen ions -> increases conductivity

CBRAM: Conductive Bridging RAM

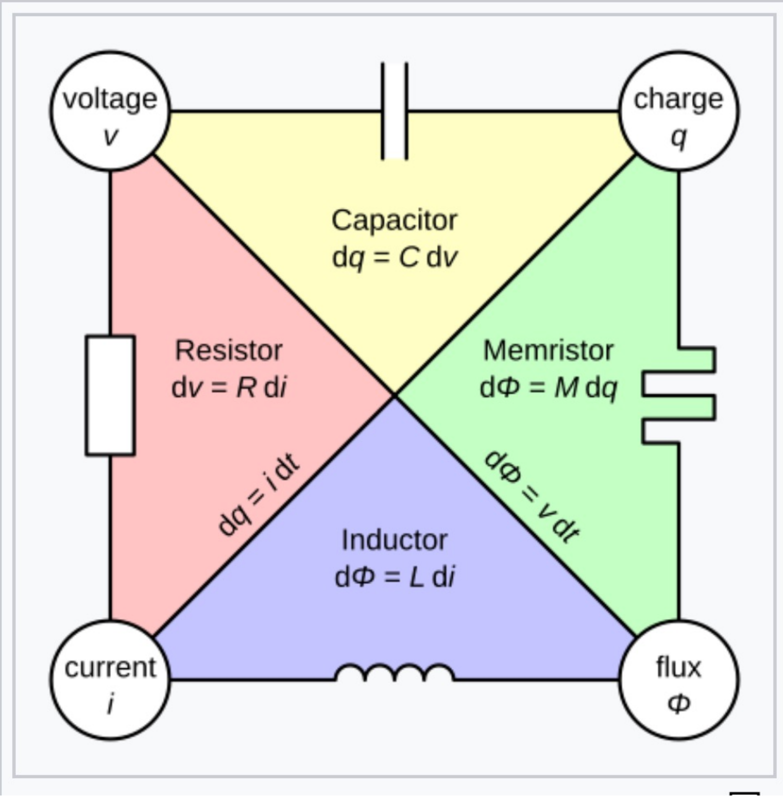
-TSMC 28 nm

RRAM



Memristors

- Theorized:
  - Leon Chua (1971)



$$i(t) = C\frac{dv}{dt} \quad \mathrm{d}q = C\,dv$$

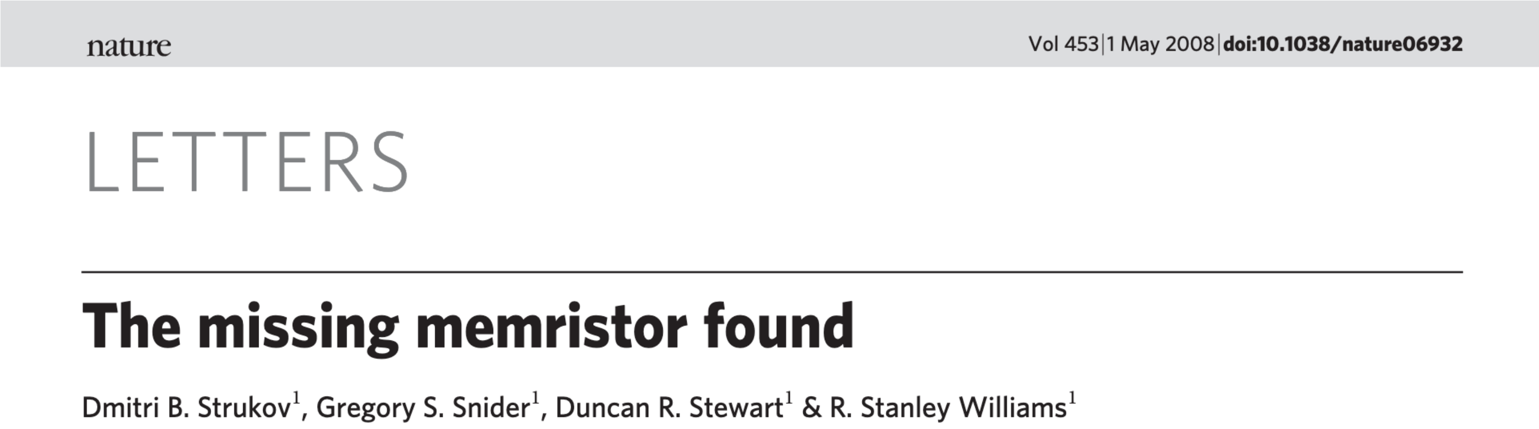
$$v(t) = L\frac{di}{dt} \quad \mathrm{d}\phi = L\,\mathrm{di}$$

$$v(t) = i(t)R \quad \mathrm{dv} = R\,\mathrm{di}$$

charge <-> current $\mathrm{dq} = \mathrm{idt}$

Flux <-> voltage $\mathrm{d}\phi = \mathrm{v\,di}$

Memristors



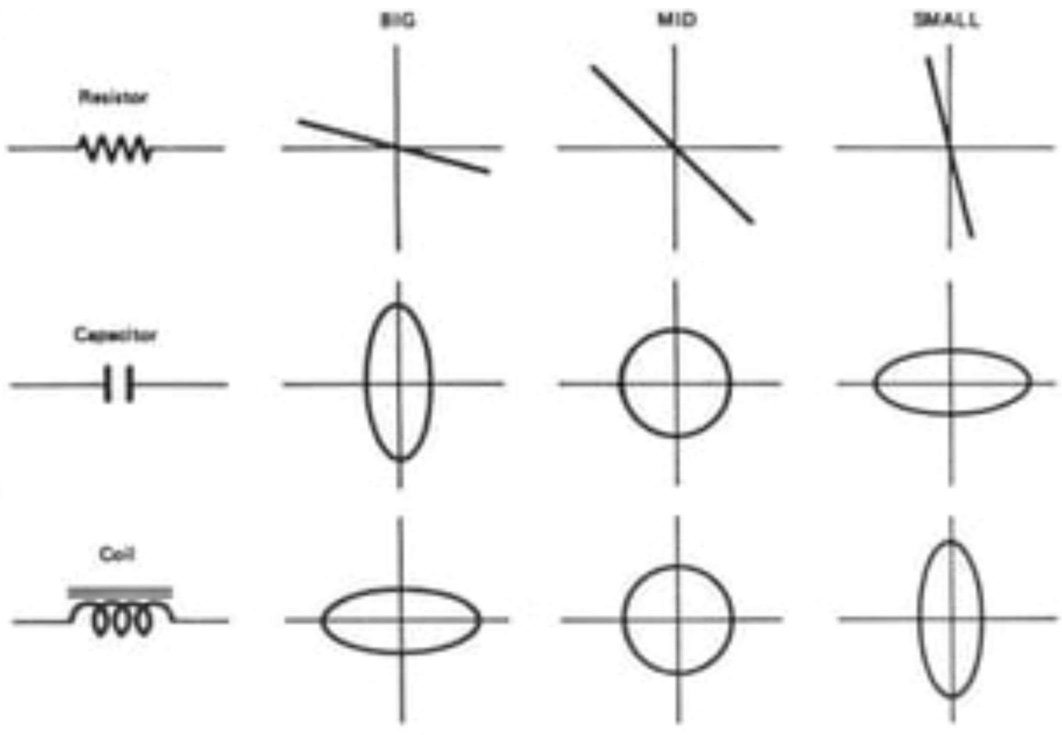
nature Vol 453|1 May 2008|**doi:10.1038/nature06932**

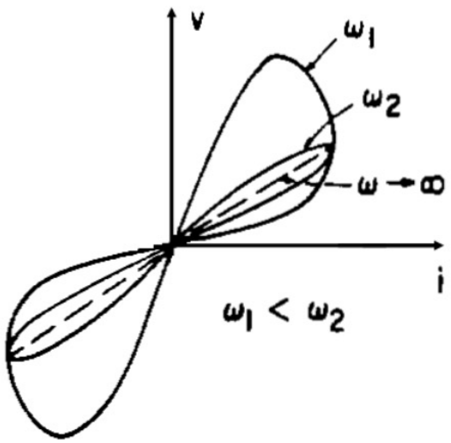
LETTERS

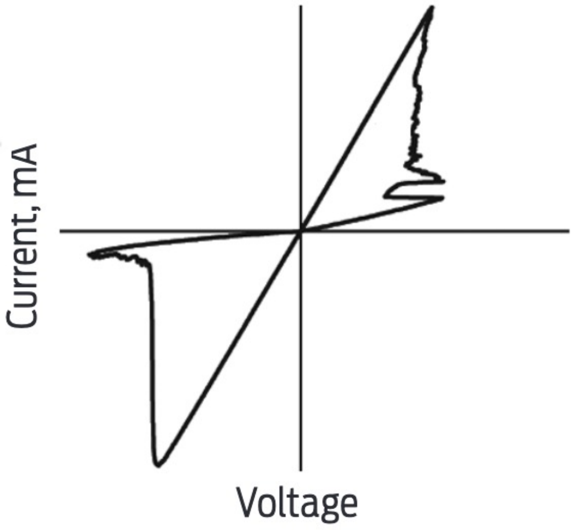
The missing memristor found

Dmitri B. Strukov[1], Gregory S. Snider[1], Duncan R. Stewart[1] & R. Stanley Williams[1]

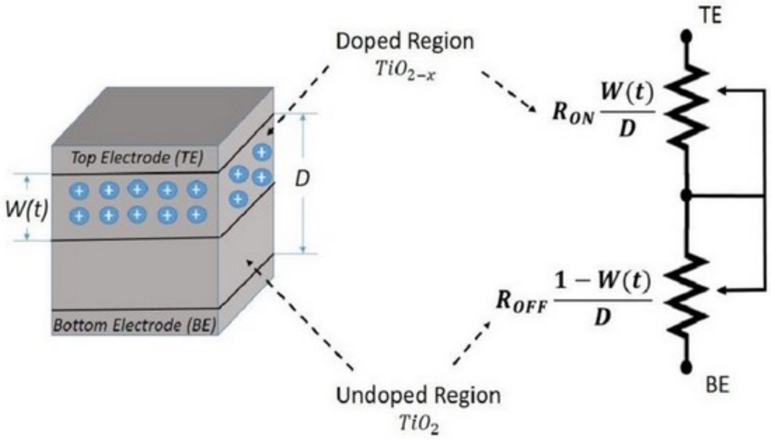
I-V Curves

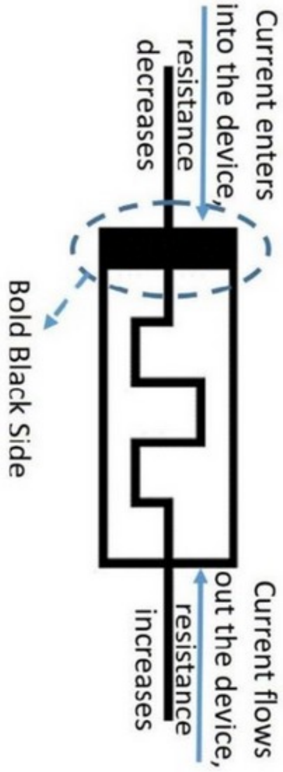

BIG
MID
SMALL
Resistor
Capacitor
Coil


v
ω_1
ω_2
$\omega \to \infty$
i
$\omega_1 < \omega_2$


Current, mA
Voltage

Memristor Operation




Current enters into the device, resistance decreases
Bold Black Side
Current flows out the device, resistance increases

$$R_{mem} = R_{on}(x) + R_{off}(1\text{-}x) \;\Big|\; x = \frac{w}{d}\ [0,1]$$

$$\frac{\text{Roff}}{\text{Ron}} >> 1000\text{x}$$

R_{on} = Conducting (low Resistance)

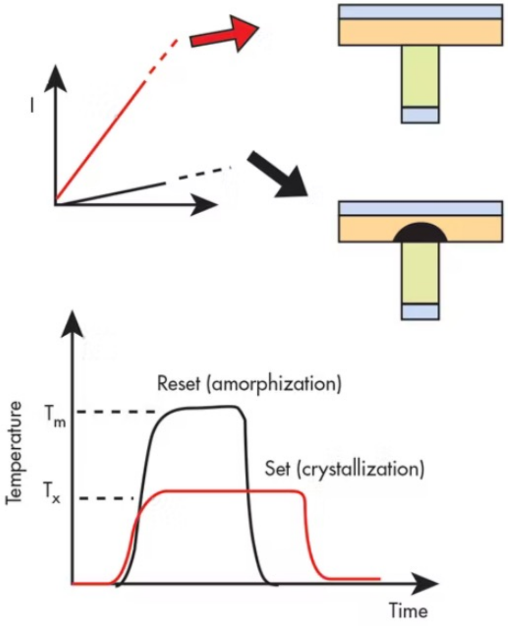
R_{off} = Resistance (high Resistance)

PCM

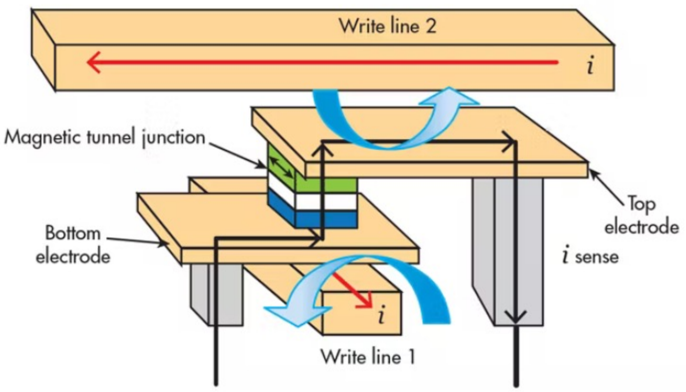
PCM works by changing the phase of a special kind of glass ($Ge_2Sb_2Te_5$) within the bit cell. Phase changed by heating

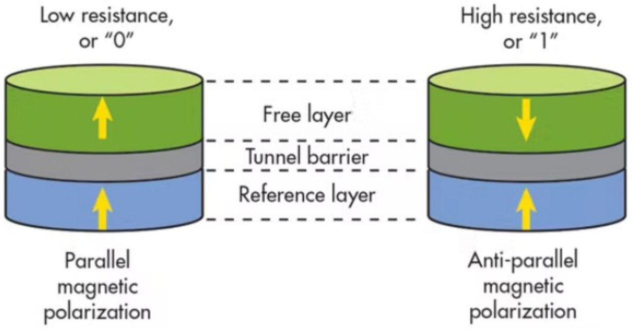
A higher heating current that is removed early causes glass to solidify into an amorphous nonconductive state (high heat -> Liquid -> high resistance)

Slower heating at lower temperature solidifies glass into a conductive crystalline structure (slow heat -> solid -> low resistance)



MRAM: Magnetic RAM (STT-RAM)





Perpendicular STT-RAM

Magnetic Tunnel Junction (MTJ) sandwiched magnetic plates:

-1 plate magnetic field set at factory

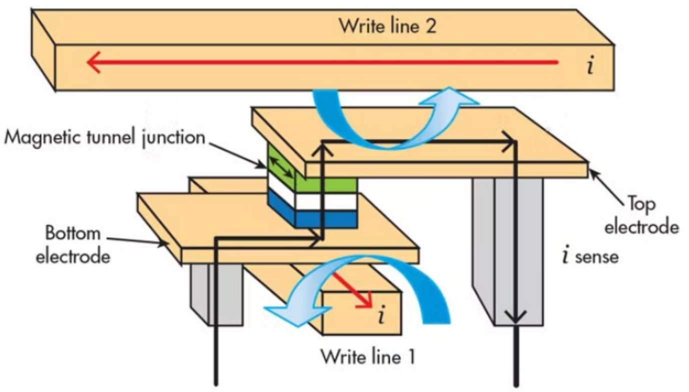
-2nd plate magnetic field variable by direction of current flow on write lines (Right hand rule)

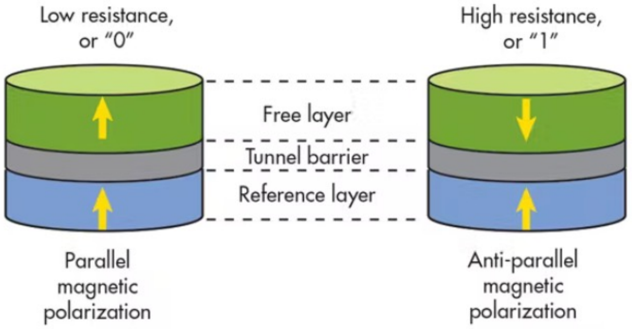
resistance of MTJ varied based on constructive/destructive combining of magnetic field

constructive -> low resistance – logic 0

destructive -> high resistance – logic 1

MRAM: Magnetic RAM (STT-RAM)





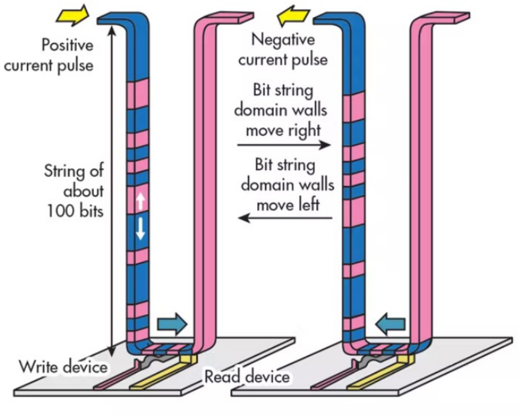
Perpendicular STT-RAM

Fabless Companies:

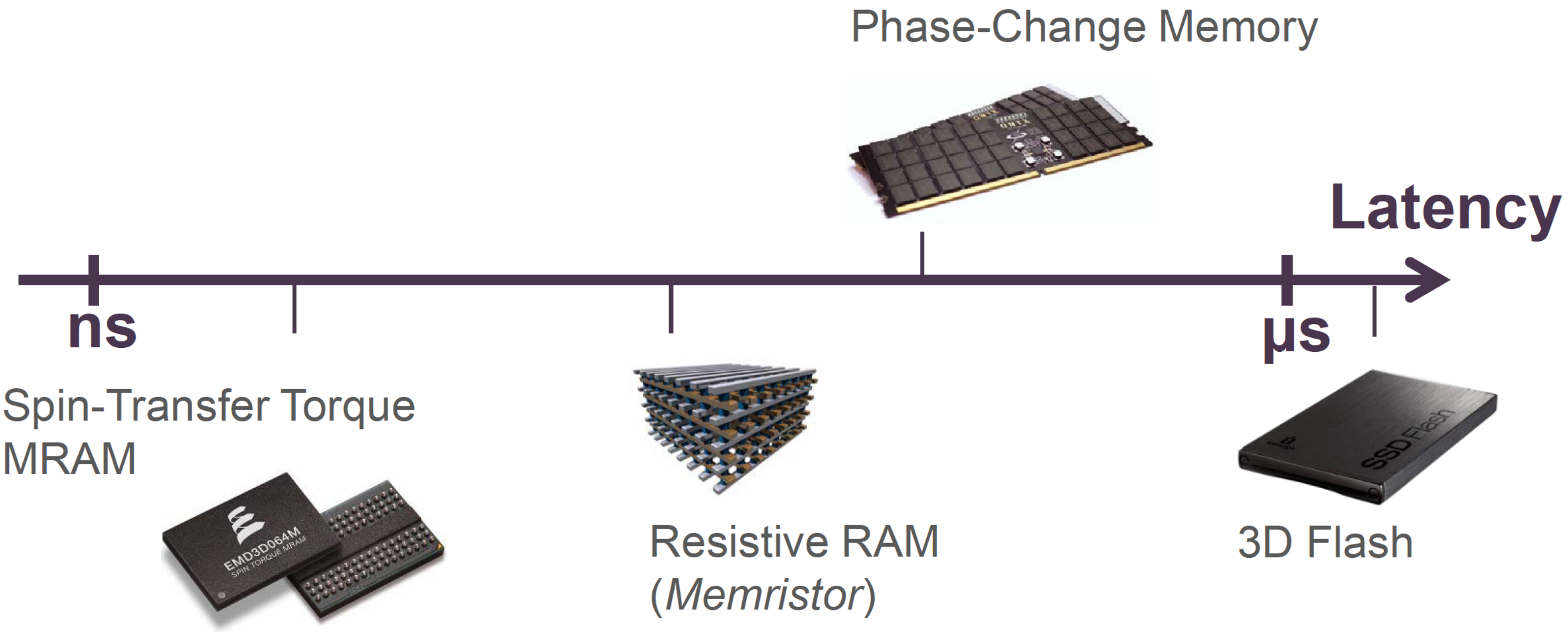
Everspin, Avalanche, Spin Transfer Technologies

Using Globalfoundries 14nm

IBM Racetrack →



Relative Speed Comparisons



Persistently stores data

Access latencies comparable to DRAM

Byte addressable (load/store) rather than block addressable (read/write)

More energy efficient and denser than DRAM

Haris Volos, et al. "Aerie: Flexible File-System Interfaces to Storage-Class Memory," *Proc. EuroSys 2014*.

Relative Comparisons

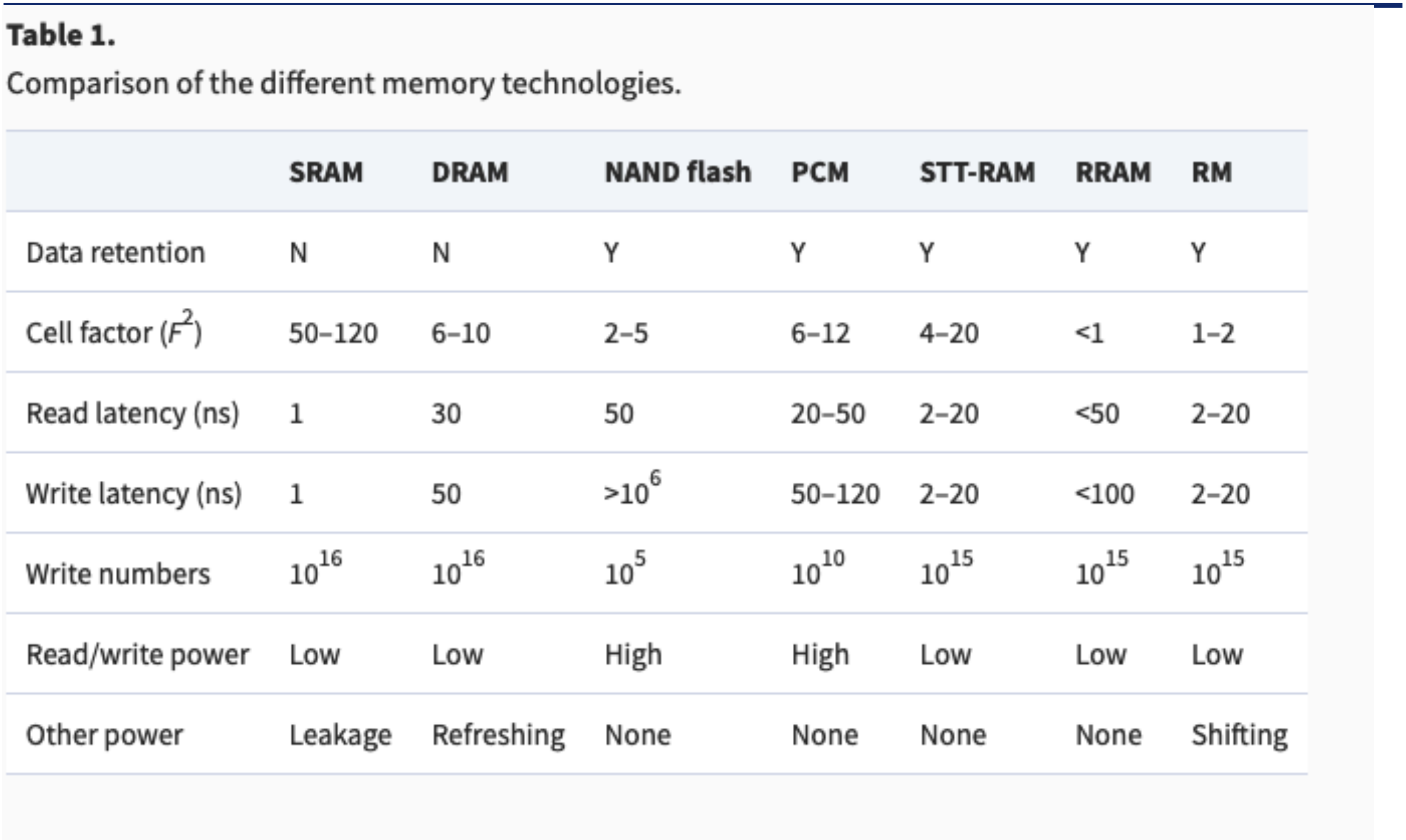


Table 1.

Comparison of the different memory technologies.

	SRAM	DRAM	NAND flash	PCM	STT-RAM	RRAM	RM
Data retention	N	N	Y	Y	Y	Y	Y
Cell factor (F^2)	50–120	6–10	2–5	6–12	4–20	<1	1–2
Read latency (ns)	1	30	50	20–50	2–20	<50	2–20
Write latency (ns)	1	50	$>10^6$	50–120	2–20	<100	2–20
Write numbers	10^{16}	10^{16}	10^5	10^{10}	10^{15}	10^{15}	10^{15}
Read/write power	Low	Low	High	High	Low	Low	Low
Other power	Leakage	Refreshing	None	None	None	None	Shifting

Gaungyu Sun et. al. "*Memory that never forgets: Emerging nonvolatile memory and the implication for Architecture design*"

- Assumption: same technology node Quantitative values are for single bit data.

Going beyond memory bottle neck

- Benchmarking Machine Learning, Artificial intelligence and neuromorphic computing memory elements
- Ultimate density can be achieved by going 3D

Parameters	**Volatile Memory**		**Non-Volatile Memory**		
	2D CMOS ASIC	**3D CMOS ASIC**	**3D CMOS RRAM**	**3D CMOS PCRAM**	**3D CMOS MRAM**
Read Time	Best	Best	Good	Poor	Comparable to RRAM
Write Time	Best (1ns)	Best (1ns)	Poor	Poor	Good (<10 ns)
Write Energy/bit	Poor(10 uJ)	Poor(10 uJ)	Good (200 pJ)	Good (500 pJ)	Best (50 pJ)
Area	OK	Good	Best(50 F^2)	Best (40-50 F^2)	Best (50-80 F^2)
Power	Good	Good	Good	Poor	Best
Throughput/W	Good	Good	Best	Poor	Best
Endurance (Cycles)	Best ($>10^{14}$)	Best ($>10^{14}$)	Bad (10^6)	Bad (10^8)	Good (10^12)
Retention	10Y 110 ^{0}C	10Y 110 ^{0}C	10Y 70 ^{0}C	10Y 70 ^{0}C	10Y 125 ^{0}C
Multi level bit			Available but not robust	**Available and Robust**	NO in STT, Hybrid S/MRAM integration?.
Maturity	Mature	Confident		**Mature**	
Synapse			Stochastic, binary, analog	low drift, analog	All spin device possible (spin wave as interconnect and MTJ as synapse

- Multi level bit memory increases the precision and accuracy. However, new circuit implementation algorithms can increase precision and accuracy to certain extent. Stochastic STT MRAM has shown improved precision!
- Multi level bit circuits are complex
- **Trade off: Multi Level Bit (RRAM/PCRAM) vs Endurance (MRAM)?**