Non-Volatile Memories & I/O Systems

Manu Awasthi


@mnwsth

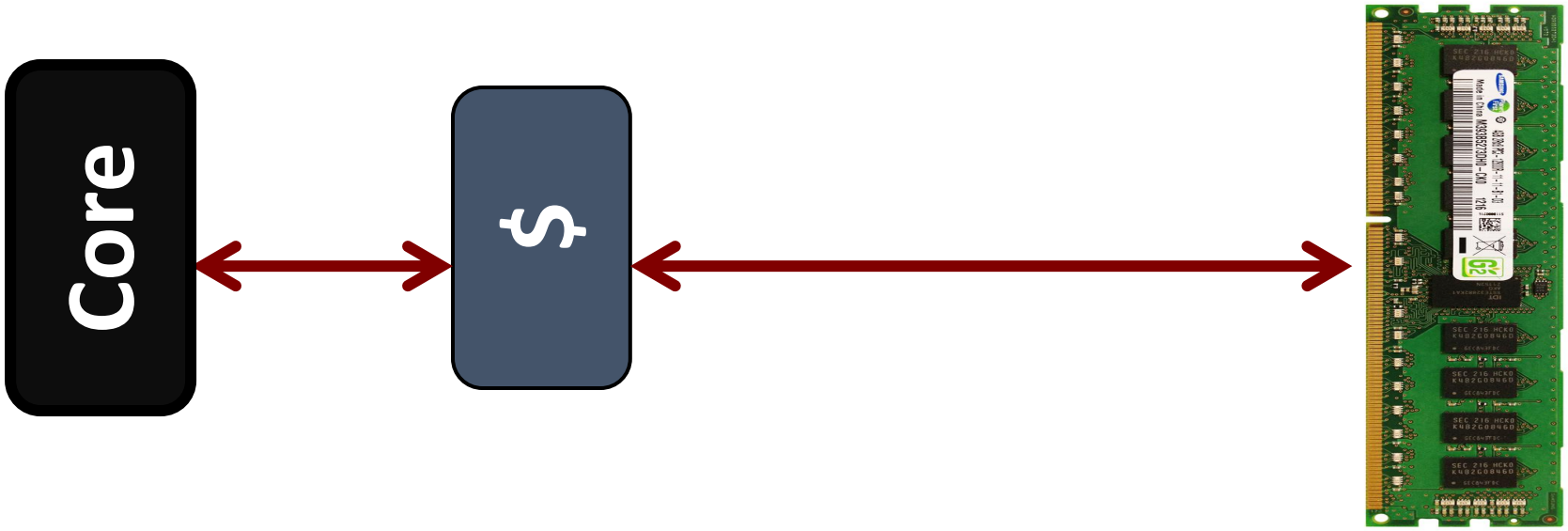
Computer Architecture Winter School 2020

Thought for the day

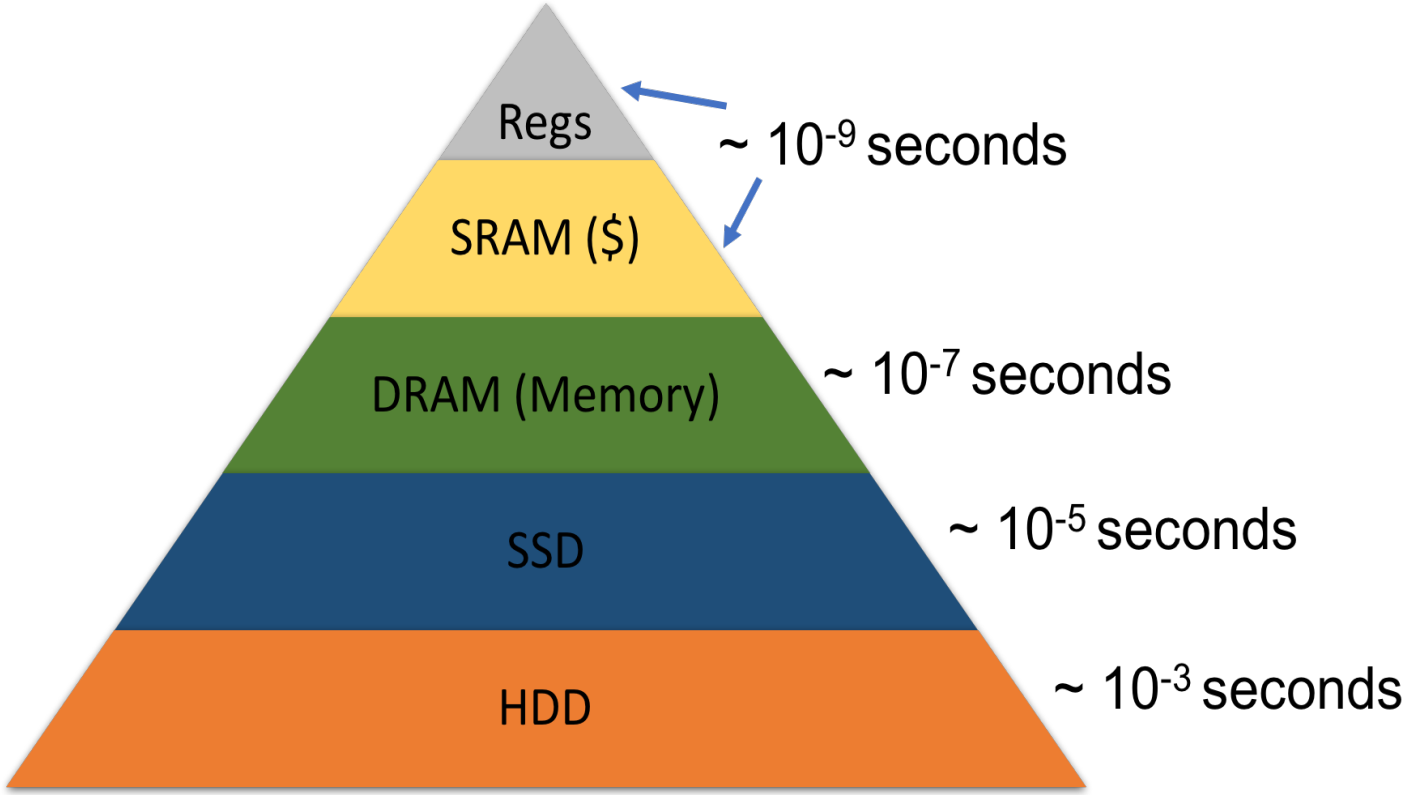
"What counts, I found, is not what I cover, but what you uncover"

(Apologies to Walter Lewin)

The Memory Hierarchy

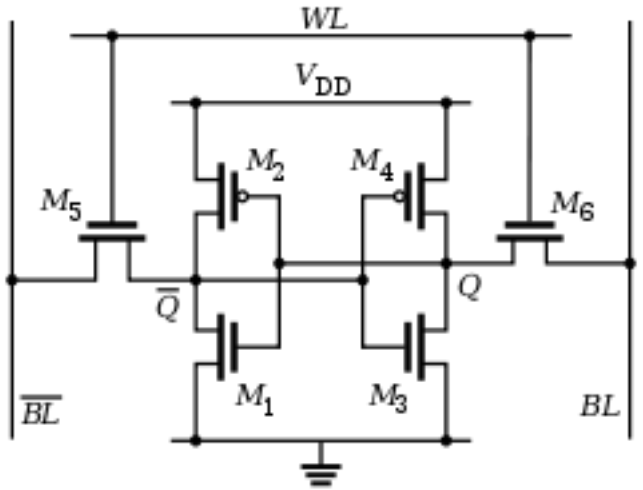


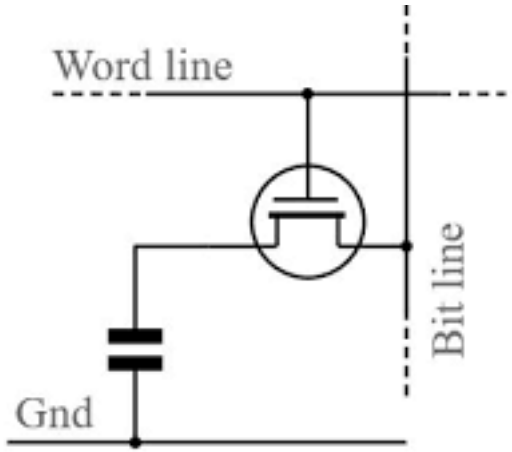
Latency Matters



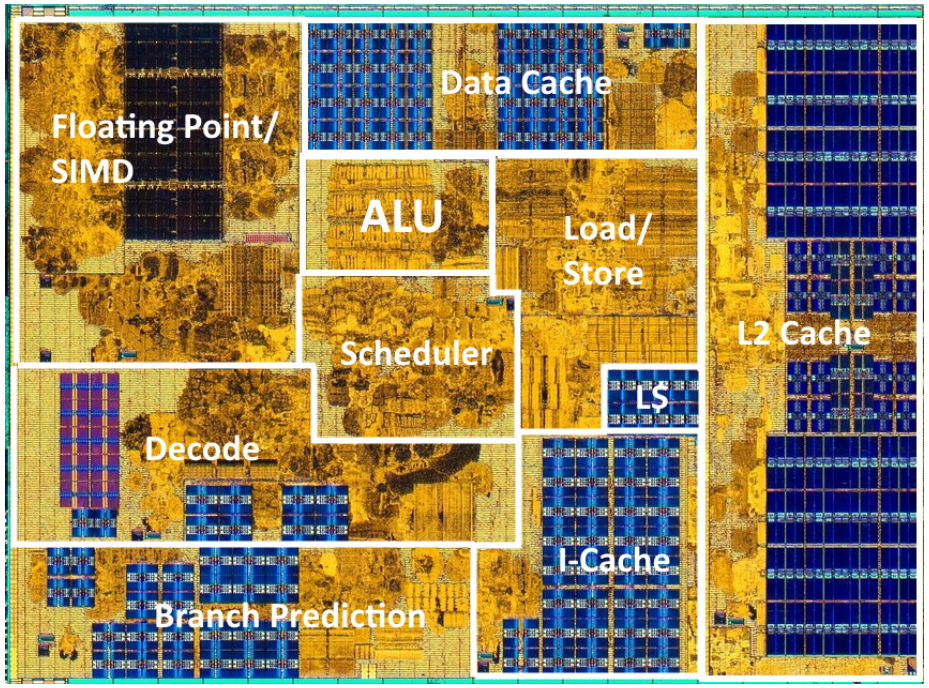
The Memory mainstays

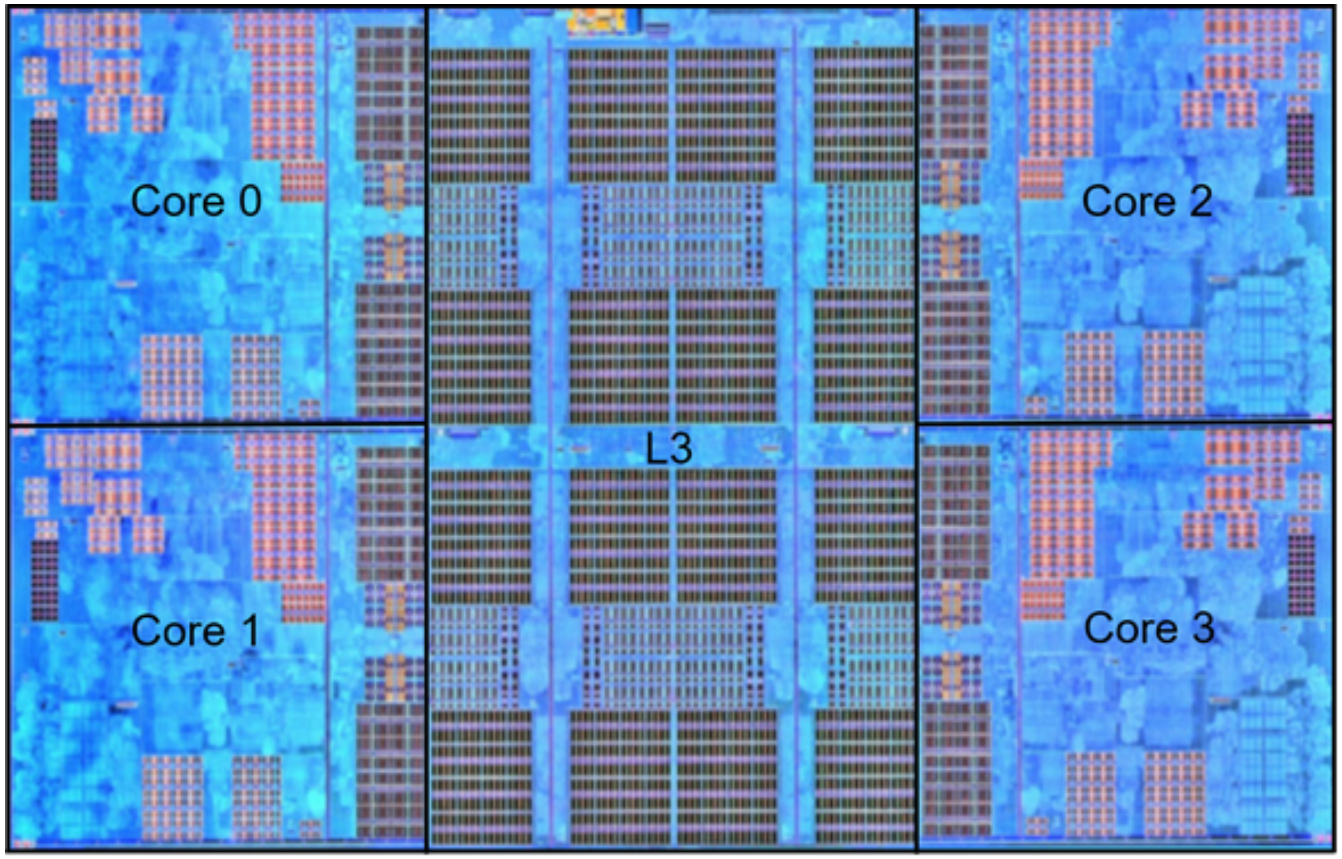
- SRAM and DRAM
- SRAM for on-chip caches; DRAM for off-chip memory
- Few 10 – 100s MB SRAM, multiple GBs of DRAM





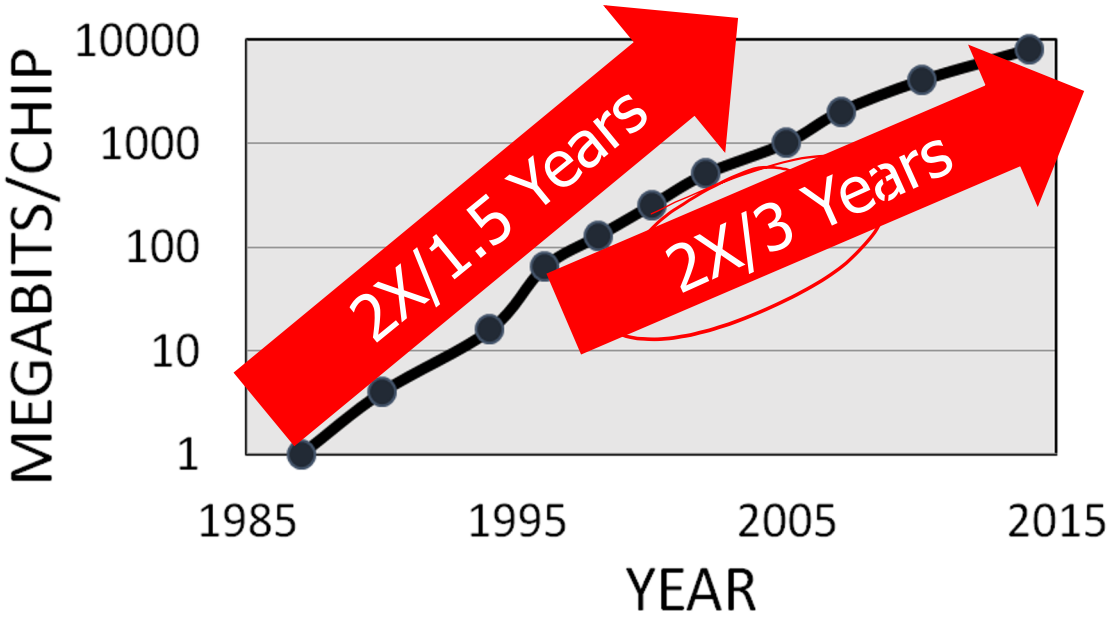
Issues with existing SRAM caches





https://en.wikichip.org/wiki/amd/microarchitectures/zen

DRAM Scaling Challenge

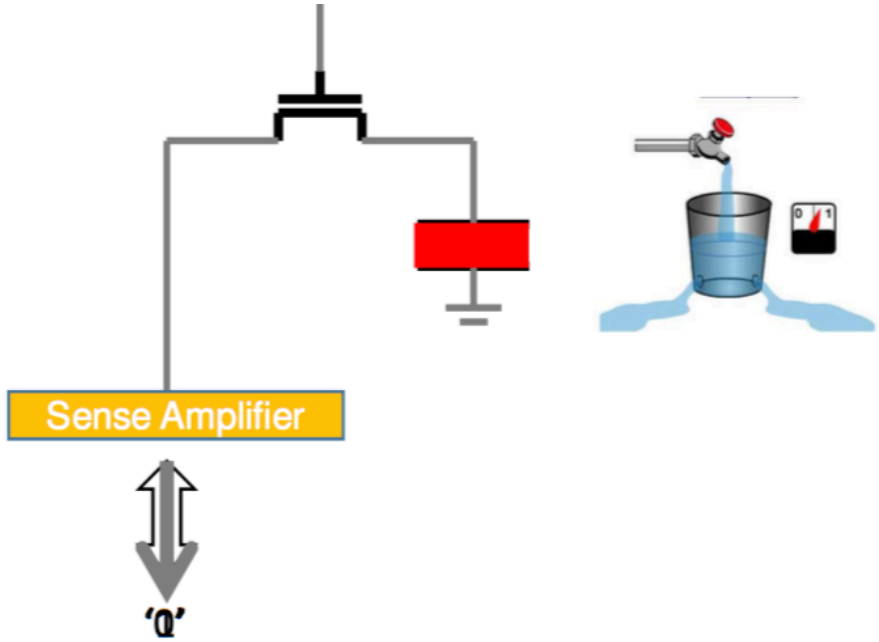


DRAM Density Scaling slowing down

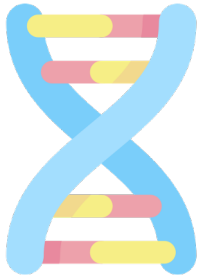
Solving the DRAM Scaling Challenge, Samira Khan, ARM Research Summit 2018

Contemporary DRAM

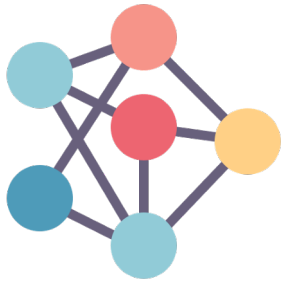
- Typical, volatile DRAM: information is stored as charge on a capacitor
- Optimized for low energy consumption, still need to be refreshed
- Possibly at the end of scaling (depends on who you ask)



The Workloads



Genomics



Neural Nets

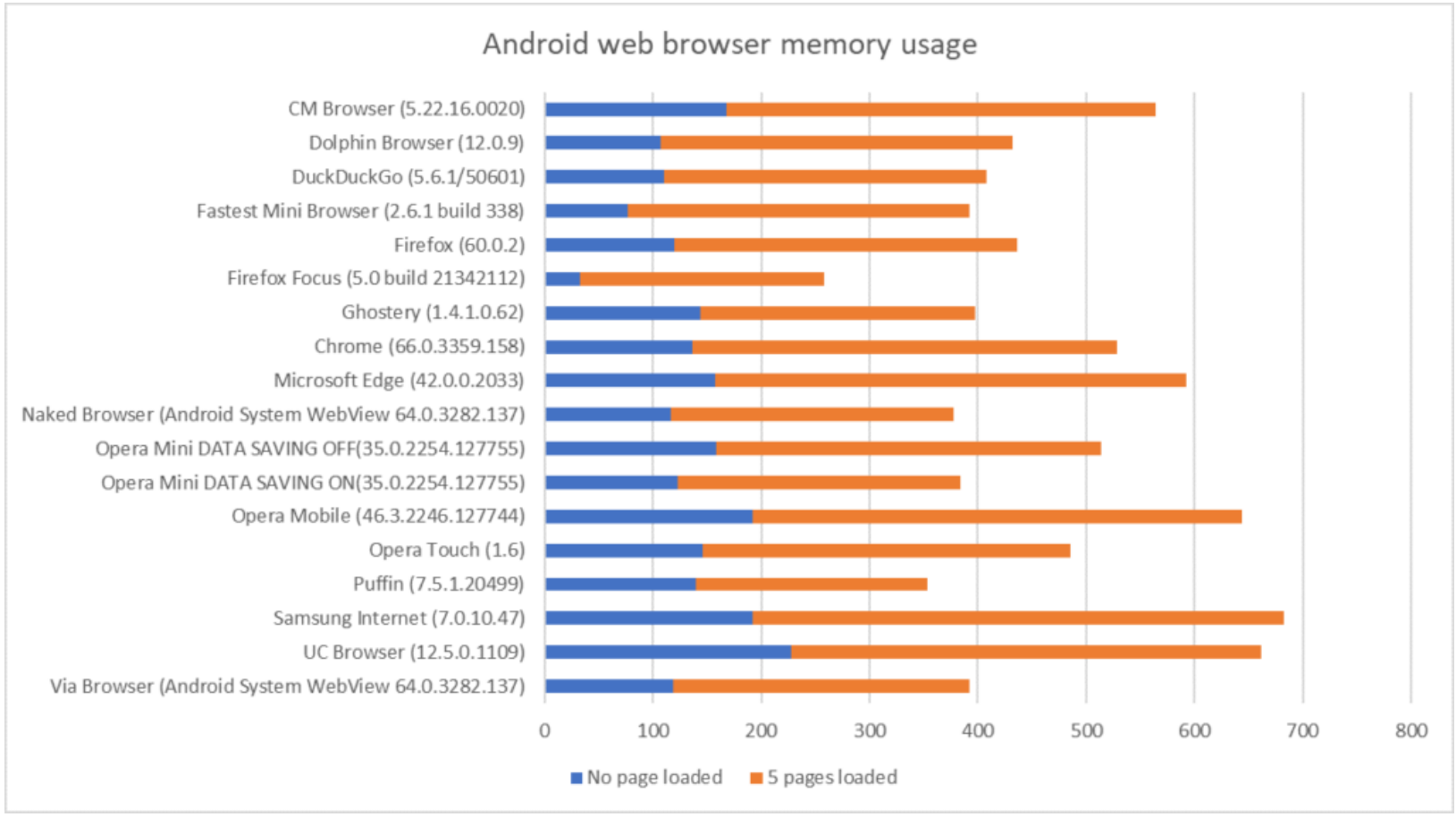
amazon DynamoDB

In-Memory Frameworks

Virtual Reality

Workloads require higher memory capacity

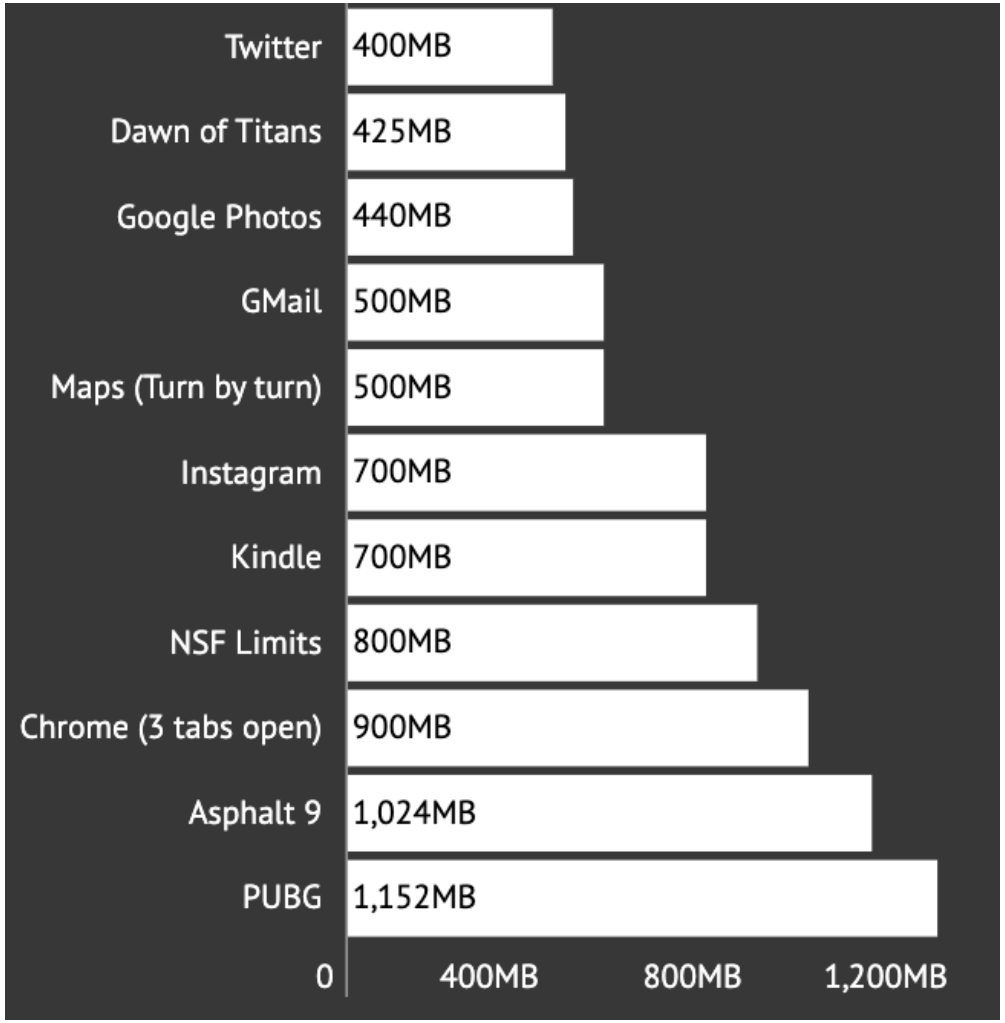
Memory Capacity : Handhelds



Applications are becoming feature rich, with increasing memory capacity requirements

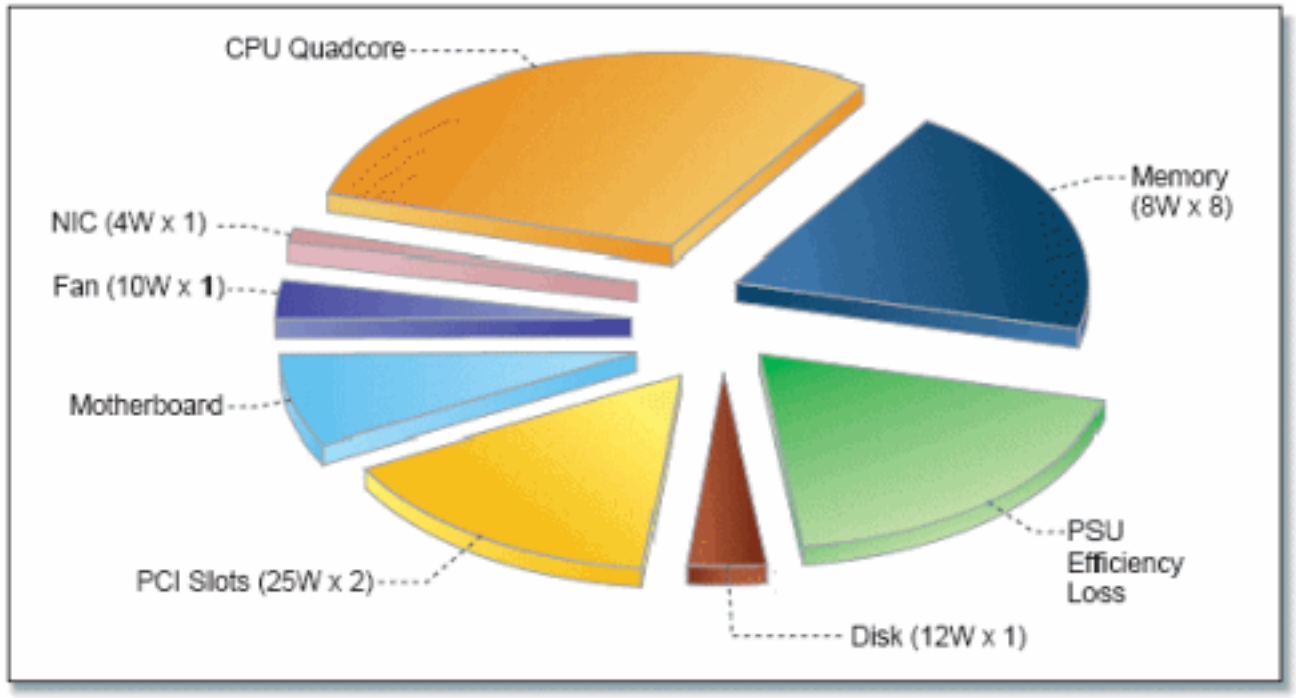
https://eitik.com/17-android-browser-tested-for-memory-usage-in-2018/

Memory Capacity : Handhelds



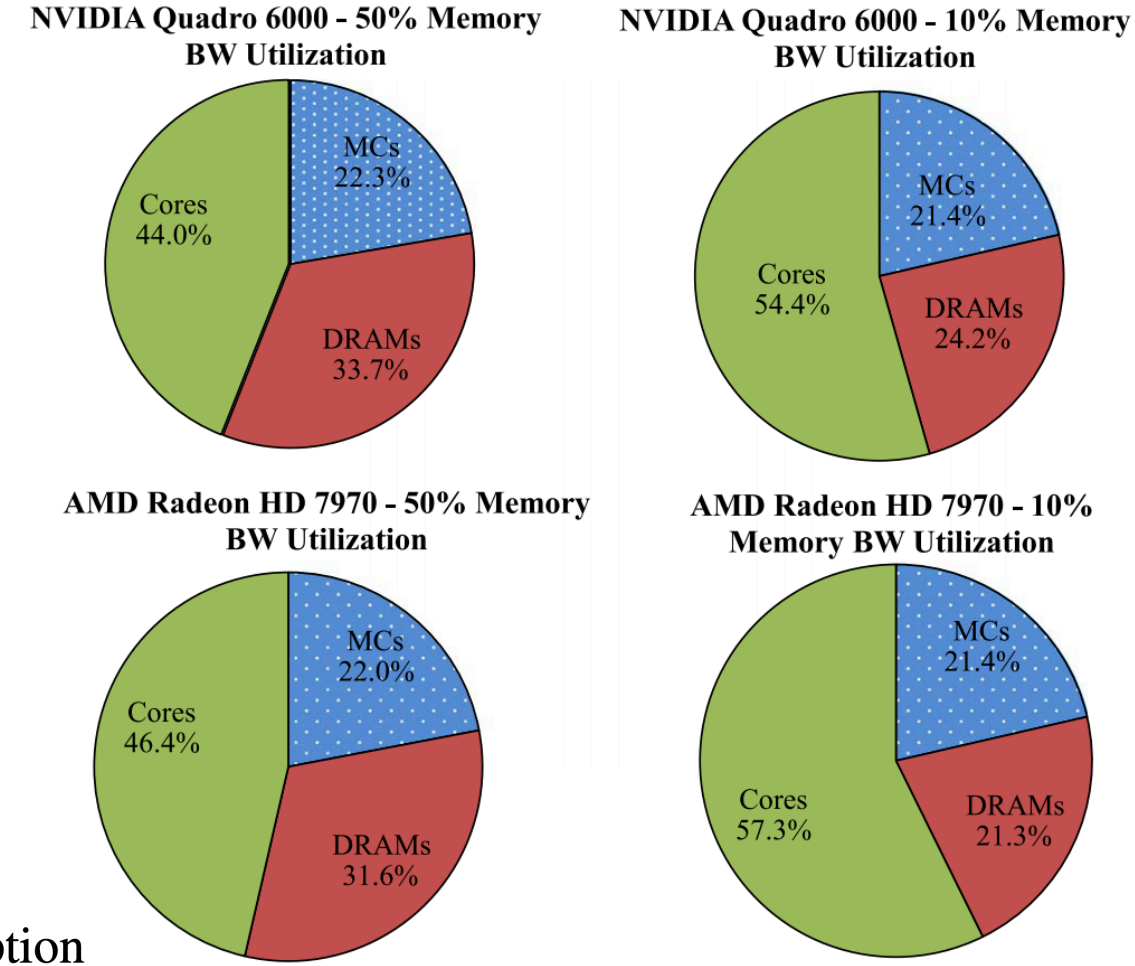
https://www.androidauthority.com/how-much-ram-do-you-need-in-smartphone-2019-944920/

Energy Consumption : Servers



Servers

Must be the CPUs, no?

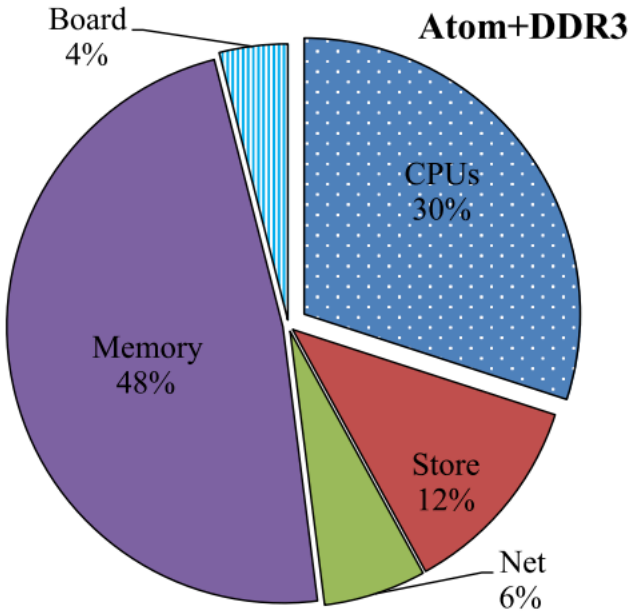


Data Center Energy Consumption Modeling: A Survey

Miyuru Dayarathna, Yonggang Wen, *Senior Member, IEEE*, and Rui Fan

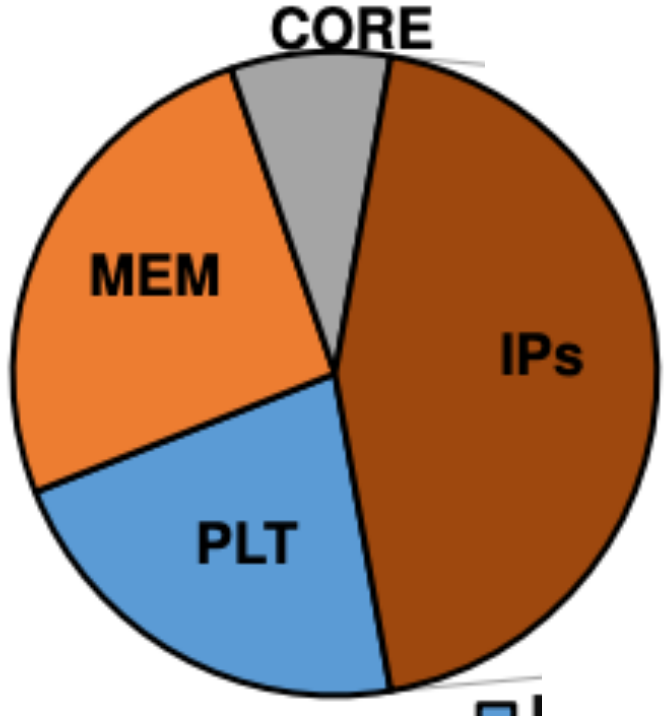
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7279063

Not you too, Handhelds!



Data Center Energy Consumption Modeling: A Survey

Miyuru Dayarathna, Yonggang Wen, *Senior Member, IEEE*, and Rui Fan

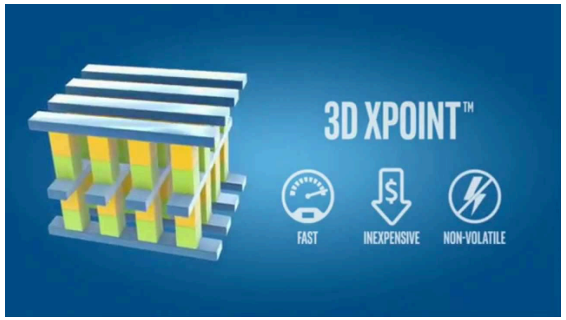


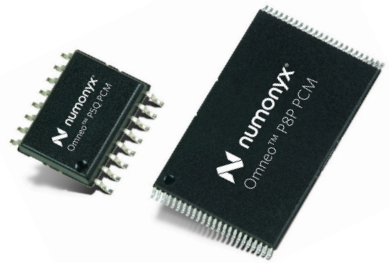
MP Game

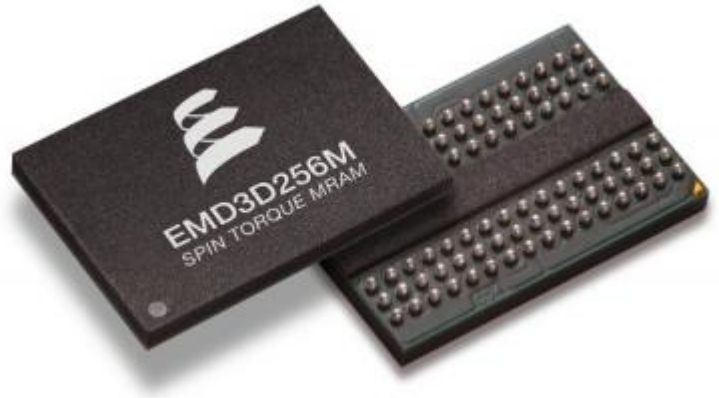
Domain Knowledge Based Energy Management in Handhelds, Nachiappan et al. HPCA 2015

So, what do we know?

Non-Volatile Memory Technologies

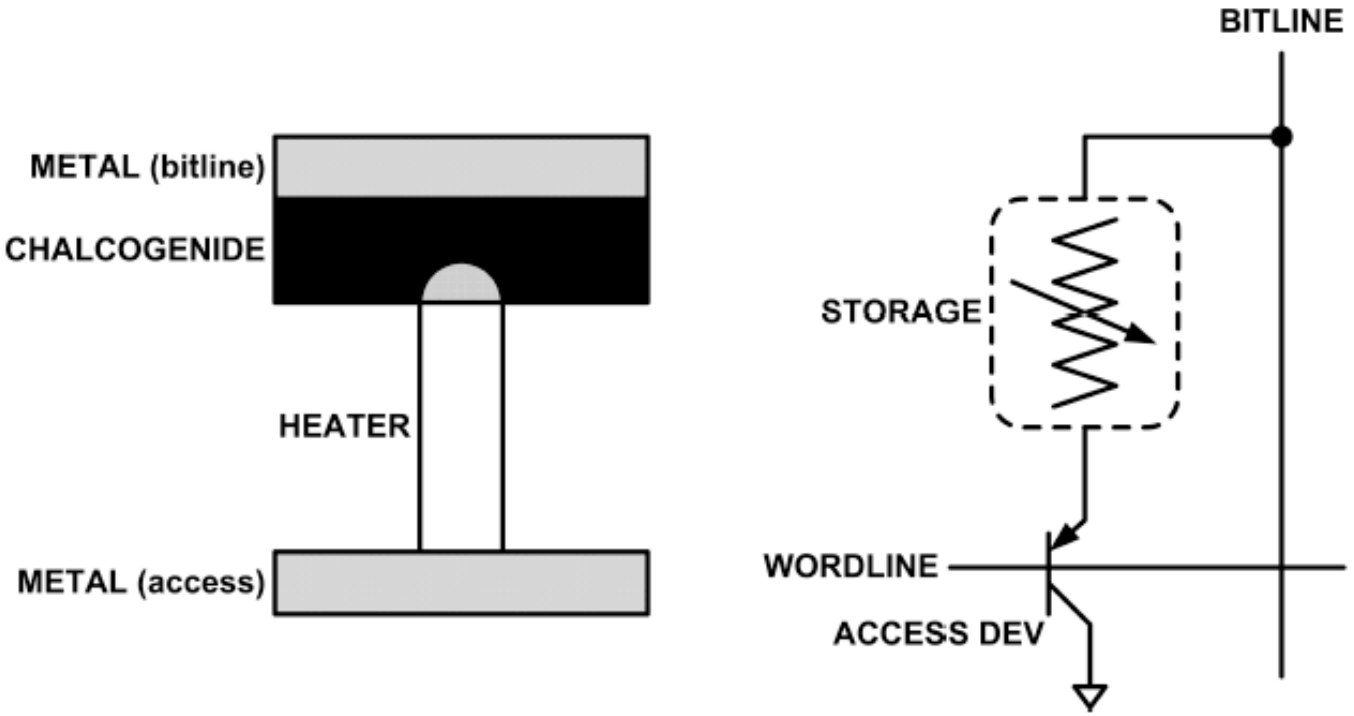






- Been around since 1960s, renewed interest with the projected decline of DRAM
- Many candidates : Phase Change Memory (PCM), Spin-Torque Transfer Memory (STT-RAM), 3D-Xpoint, Resistive RAM (ReRAM) etc.
- Vary in underlying mechanism for storing information

Phase Change Memory Primer

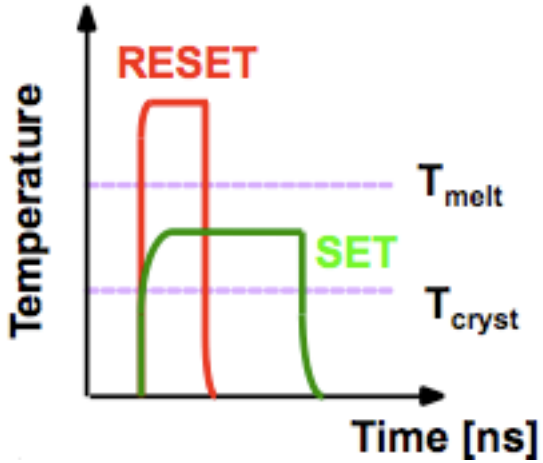


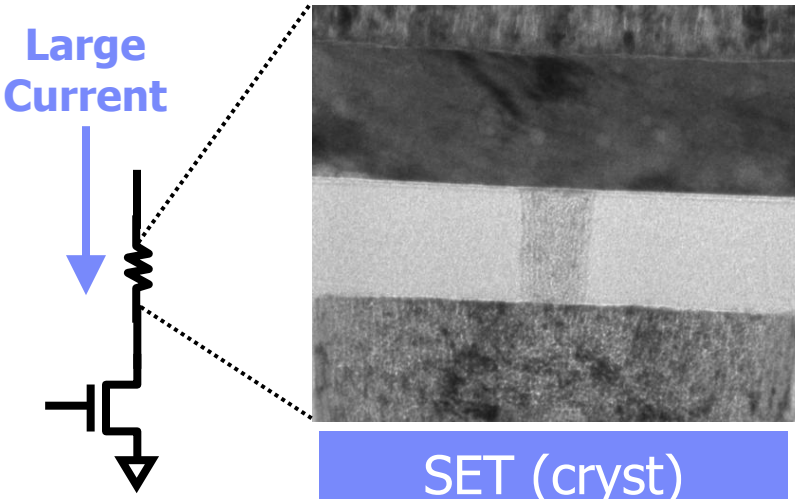
PCM is resistive memory: High resistance (0), Low resistance (1)
PCM cell can be switched between states reliably and quickly

Slide Courtesy: Moinuddin Qureshi, IBM

PCM Working Example

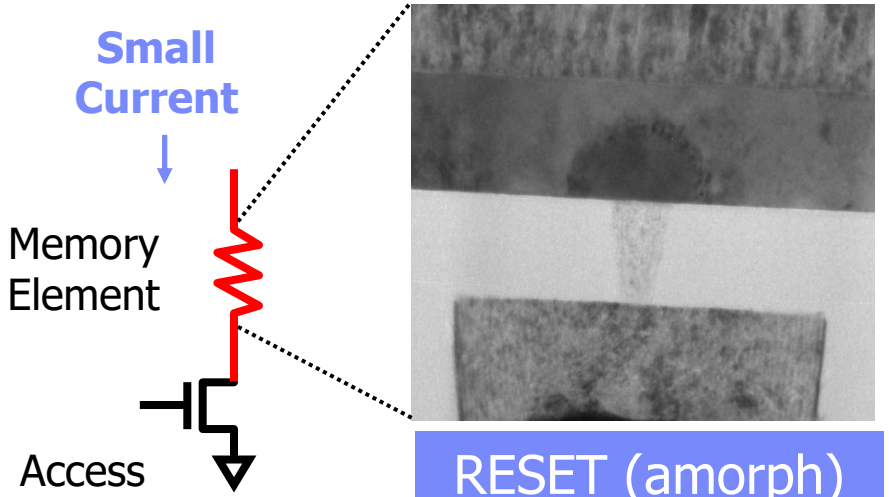
- Write: change phase via current injection
  - **SET**: sustained current to heat cell above T_{cryst}
  - **RESET**: cell heated above T_{melt} and quenched
- Read: detect phase via material resistance





SET (cryst)
Low resistance

10^3-10^4 Ω



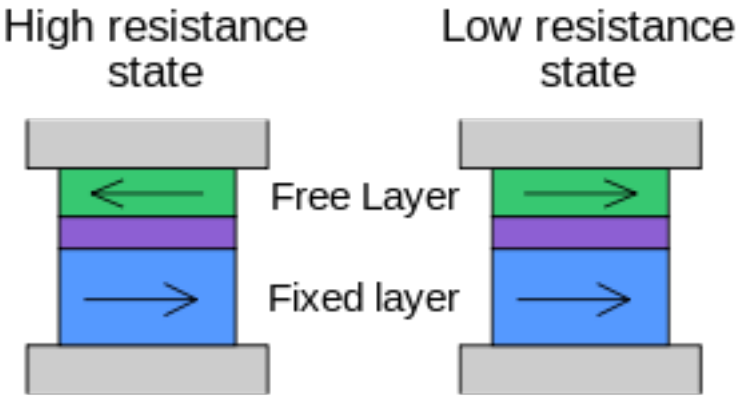
RESET (amorph)
High resistance

10^6-10^7 Ω

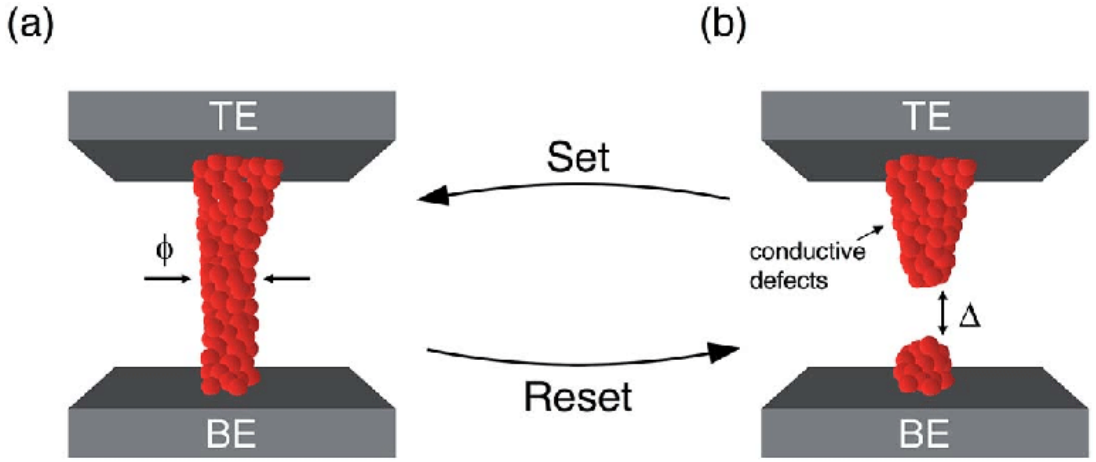
Slide Courtesy: Moinuddin Qureshi, IBM

Other Non-Volatile Memories

- Spin Torque Transfer RAM (STT-RAM)
  - Inject current to change magnet polarity
  - Resistance determined by polarity



- Memristors/RRAM/ReRAM
  - Inject current to change atomic structure
  - Resistance determined by atom distance



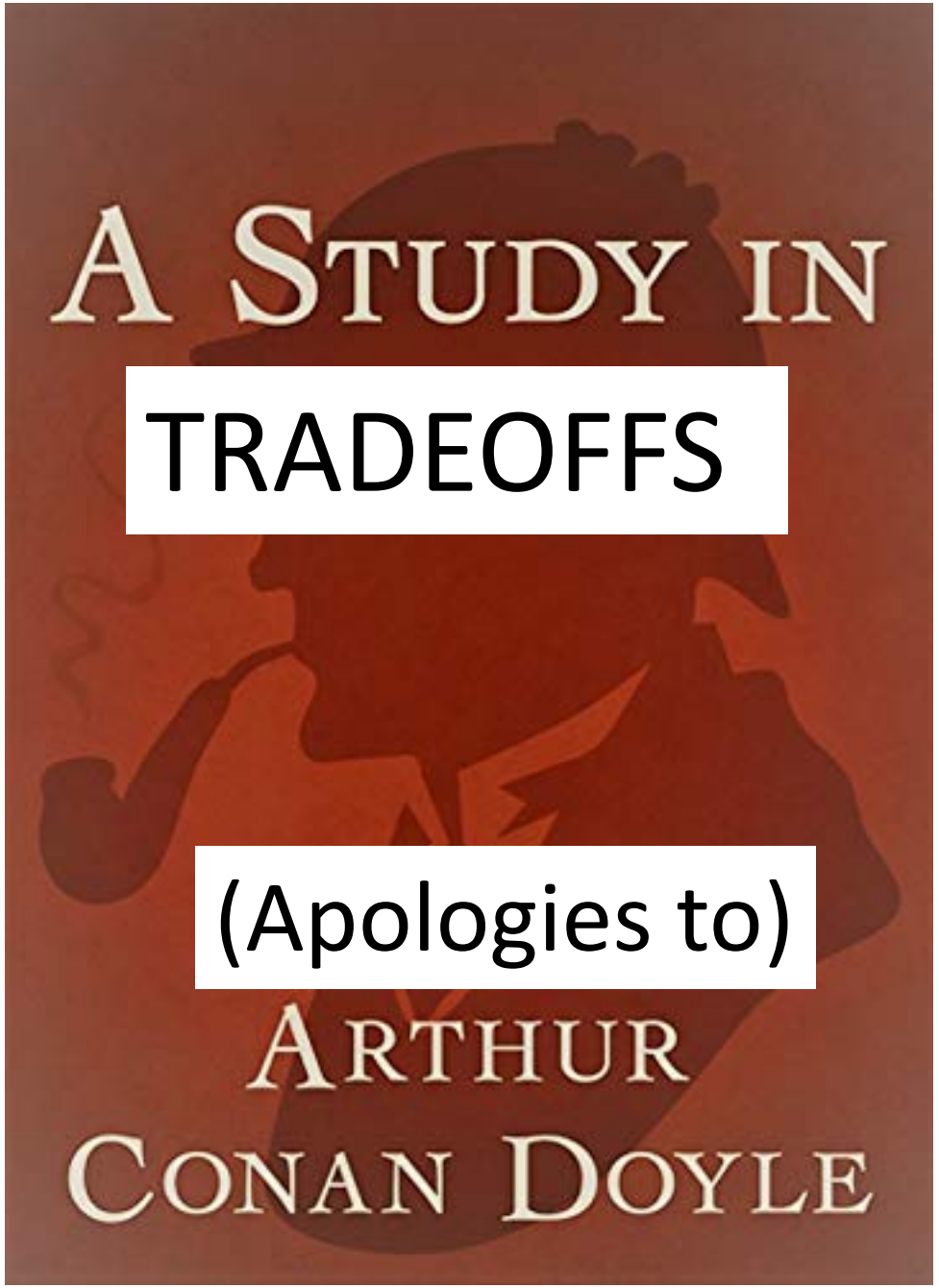
The Good, the Bad and the Ugly

+ Many candidates: PCM, STT-MRAM, others
+ Higher areal density : Higher compared to DRAM; much higher than SRAM
+ No Refresh / Standby energy
- Higher access latencies
- Asymmetric read / write energies
- Limited lifetimes

	Cell size	Access Granularity	Read Latency	Write Latency	Erase Latency	Endurance	Standby Power
HDD	N/A	512 B	5 ms	5 ms	N/A	$> 10^{15}$	1W
SLC Flash	$4–6F^2$	4 KB	25 μs	500 μs	2 ms	$10^4 - 10^5$	0
DRAM	$6–10F^2$	64 B	50 ns	50 ns	N/A	$> 10^{15}$	Refresh power
PCM	$4–12F2$	64 B	50 ns	500 ns	N/A	$10^8 - 10^9$	0
STT-RAM	$6–50F^2$	64 B	10 ns	50 ns	N/A	$> 10^{15}$	0
ReRAM	$4–10F^2$	64 B	10 ns	50 ns	N/A	10^{11}	0

A Survey of Software Techniques for Using Non-Volatile Memories for Storage and Main Memory Systems, Mittal et al., IEEE TPDS 2016

NVMs



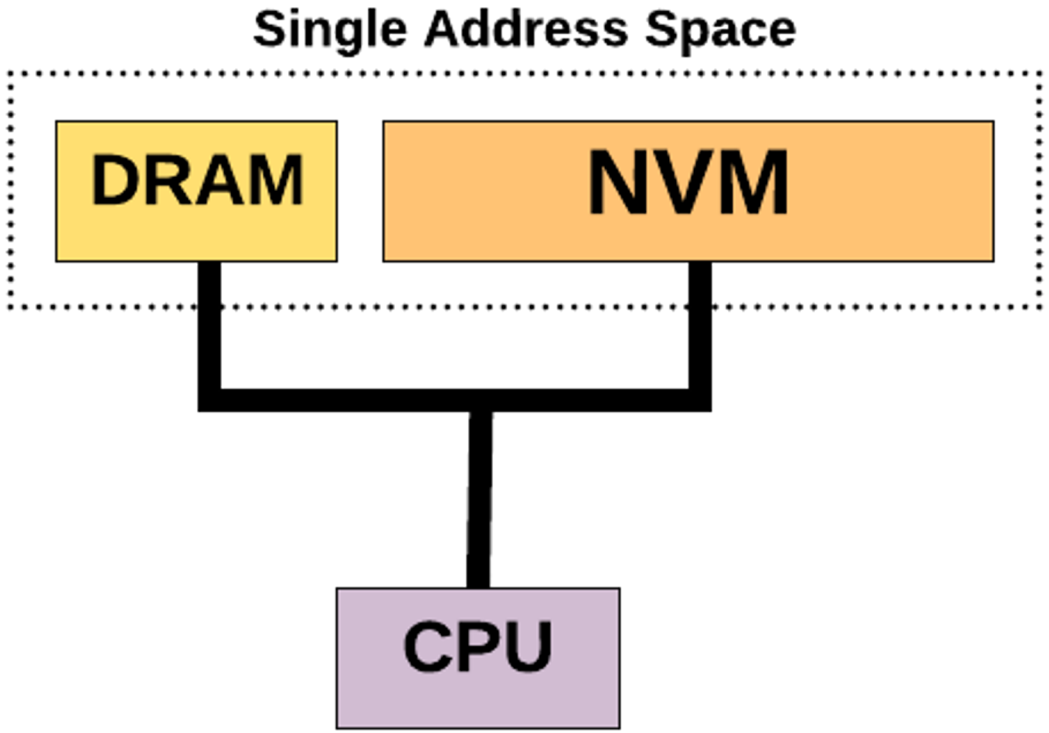
The Question

How and where do we use these emerging technologies to design memory systems?

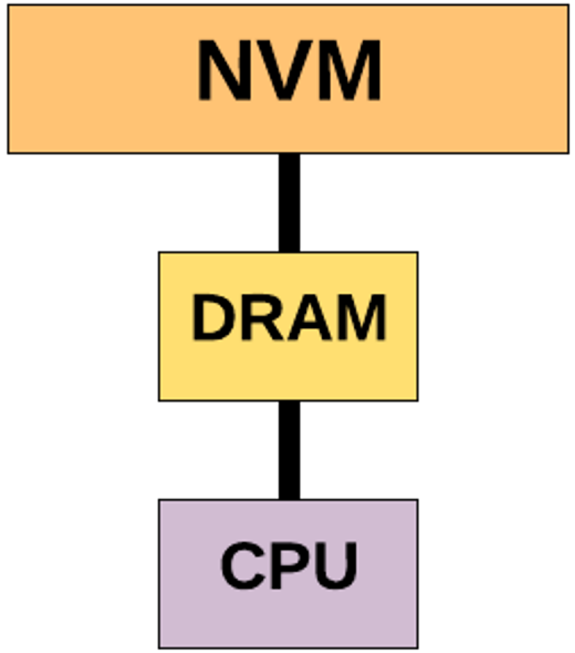
Use the Pluses, Overcome the minuses

Hybrid Main Memory

Use DRAM and NVM synergistically

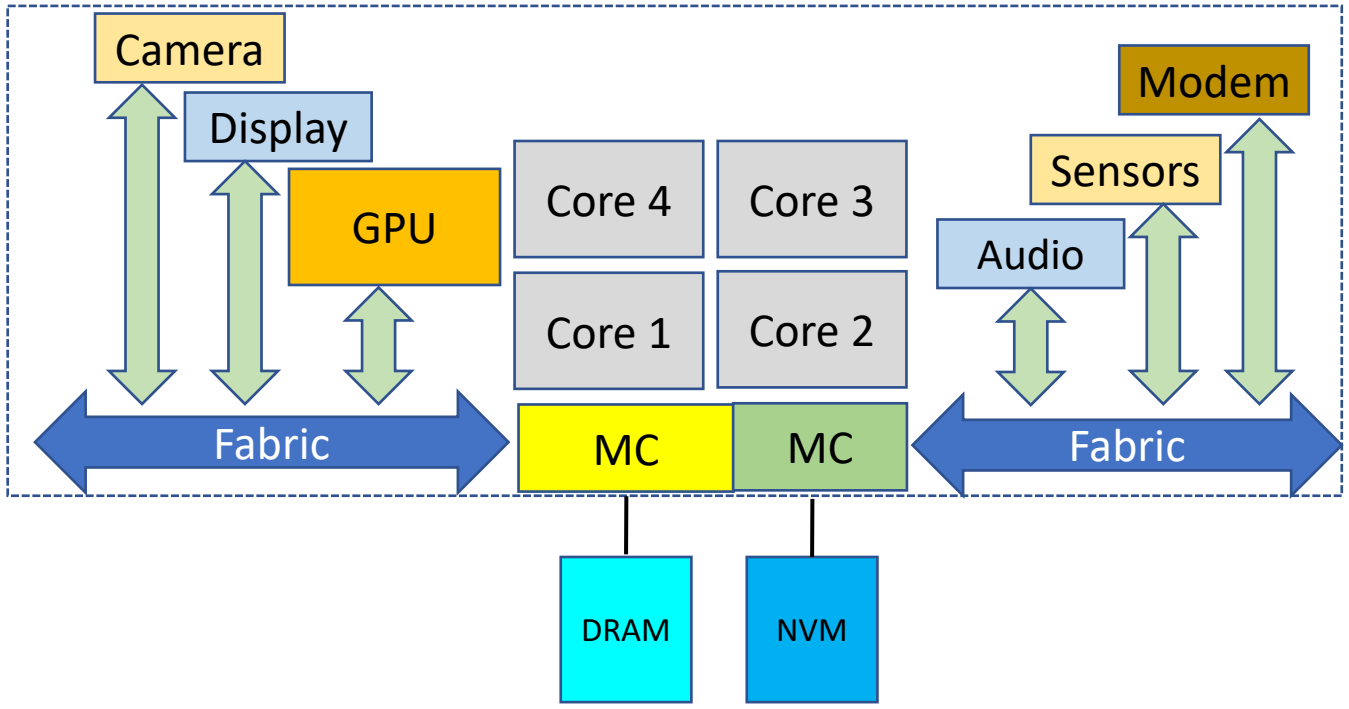


Single Address Space Variant



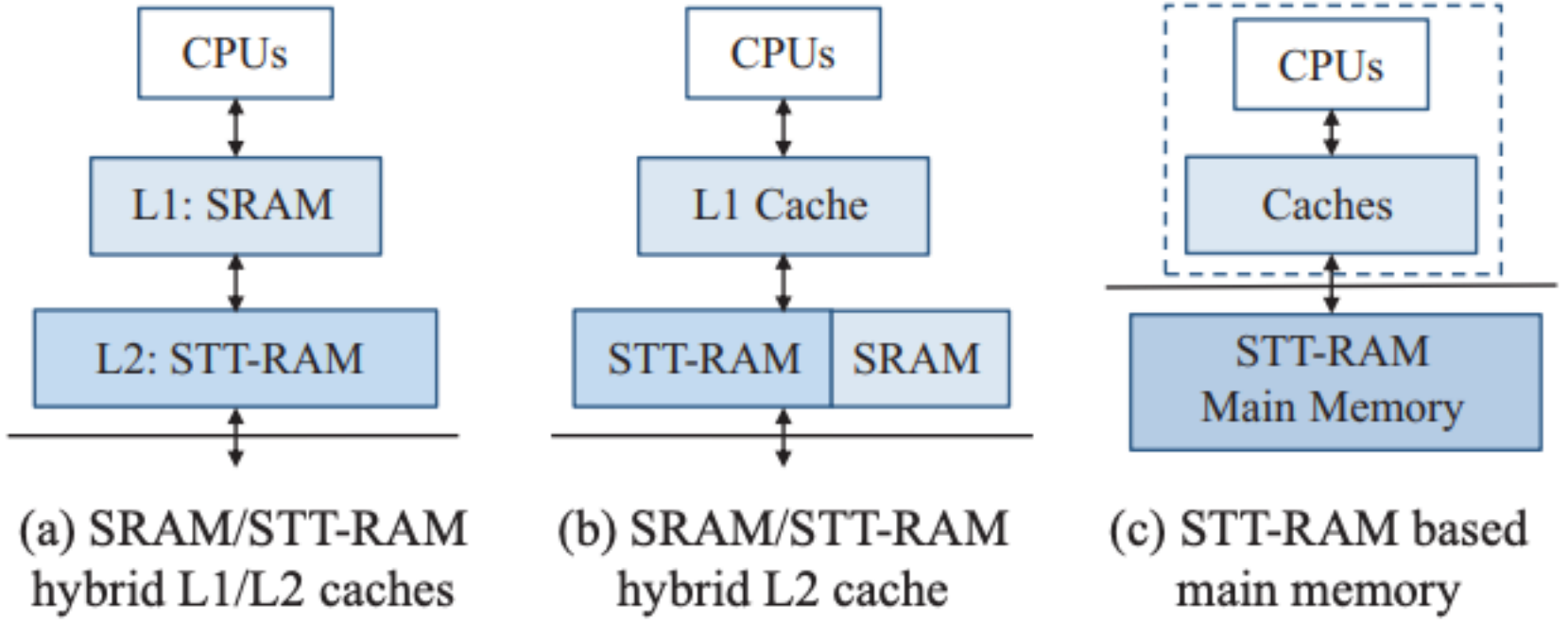
DRAM as a Cache Variant

Hybrid Main Memory in Handhelds



Exploring Non-Volatile Main Memory Architectures for Handheld Devices, **DATE 2018**

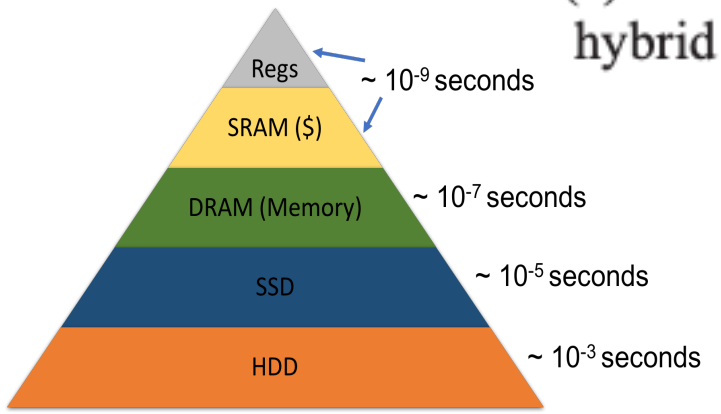
NVM in Caches?



(a) SRAM/STT-RAM hybrid L1/L2 caches

(b) SRAM/STT-RAM hybrid L2 cache

(c) STT-RAM based main memory



Architecture Design with STT-RAM: Opportunities and Challenges

Ping Chi†, Shuangchen Li†, Yuanqing Cheng†, Yu Lu‡, Seung H. Kang‡, Yuan Xie†
†Department of Electrical and Computer Engineering, University of California, Santa Barbara, USA
‡Qualcomm Incorporated, San Diego, USA
†{pingchi, shuangchenli, yuanqing, yuanxie}@ece.ucsb.edu, ‡yu.lu@qualcomm.com

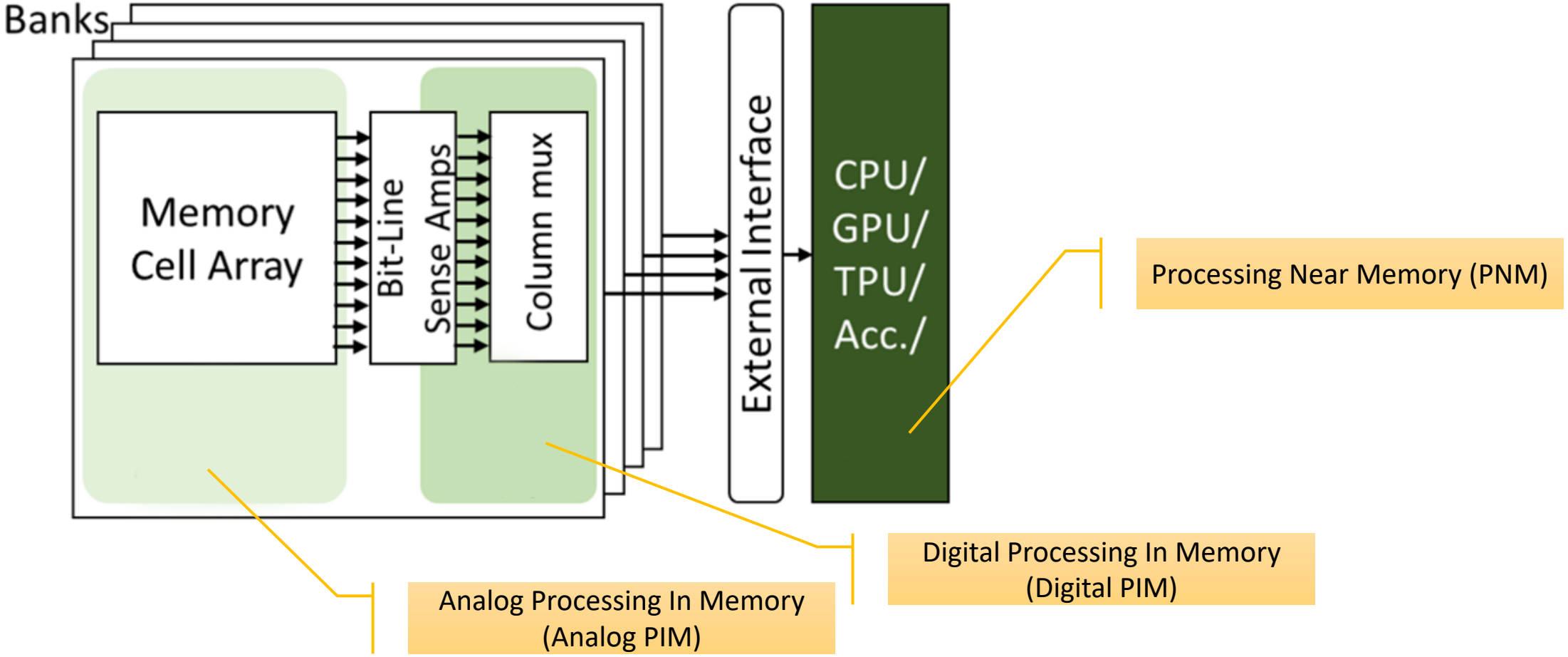
Key Takeaways

- Memory hierarchy has growing requirements in terms of latency, capacity, bandwidth, energy
- Contemporary technologies cannot keep up with demands
- NVMs might be a good alternative if:
  - We can use the pluses they offer and
  - Overcome the inherent minuses of the technologies

Research Problems (50K feet view)

- What are your metrics of interest?
- Which technology to use?
- Which level to use it at?
- What are the requirements of your workloads?

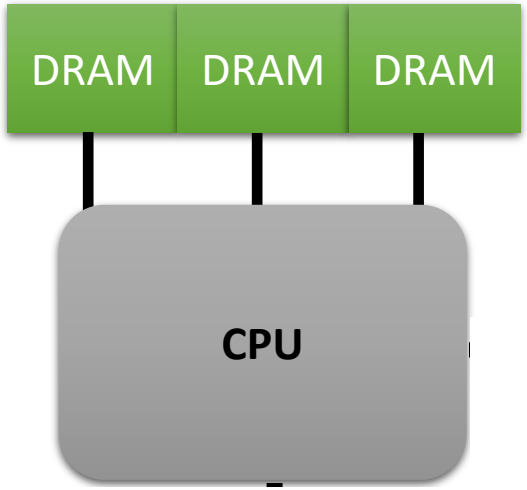
Processing Near Memory vs In Memory



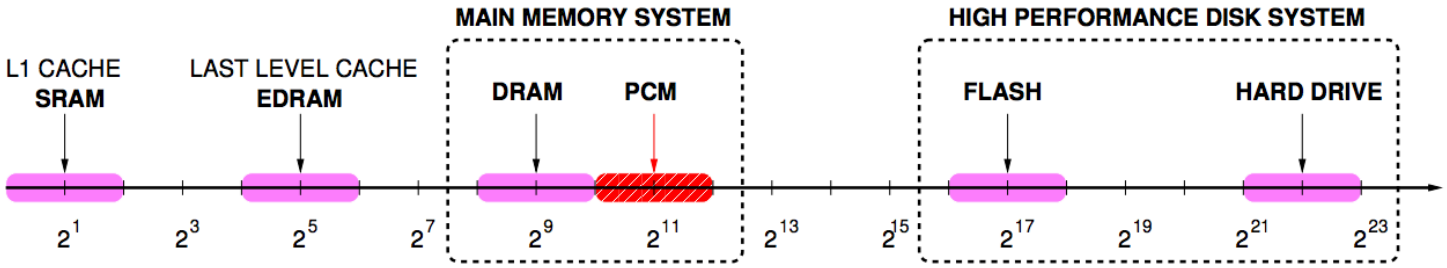
He, Mingxuan, et al. "Newton: A DRAM-maker's Accelerator-in-Memory (AiM) Architecture for Machine Learning."

A hard segue

Computer Systems

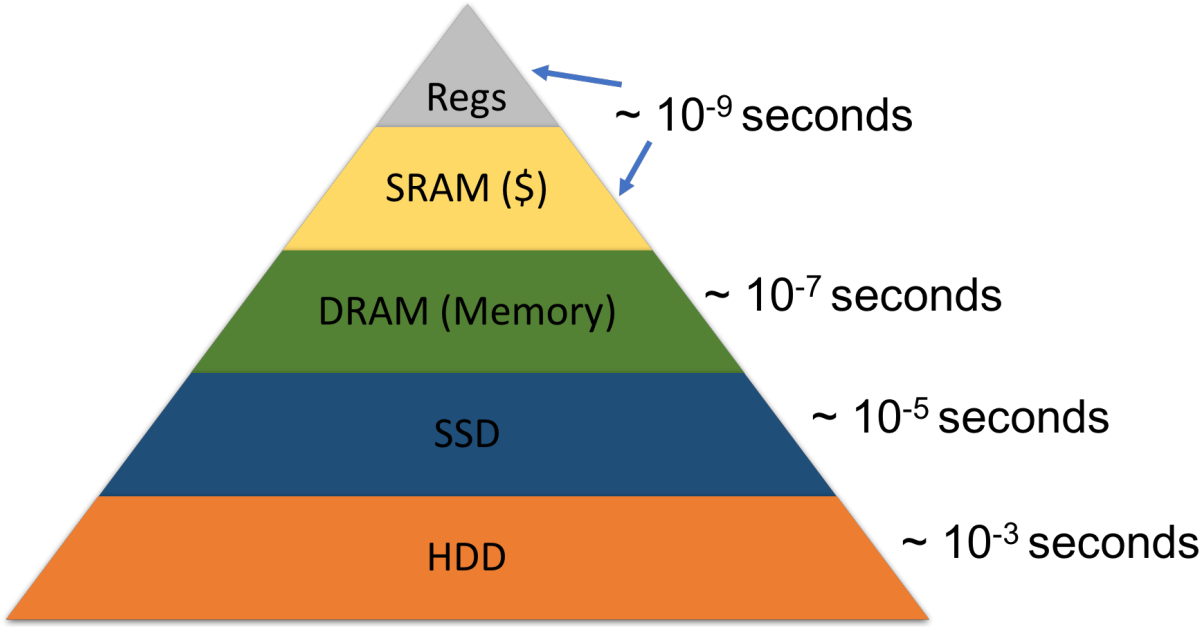


Why study storage?



Typical Access Latency (in terms of processor cycles for a 4 GHz processor)

Scalable High Performance Main Memory System Using Phase-Change Memory Technology, Qureshi et al , ISCA 2009



Data Generation Trends

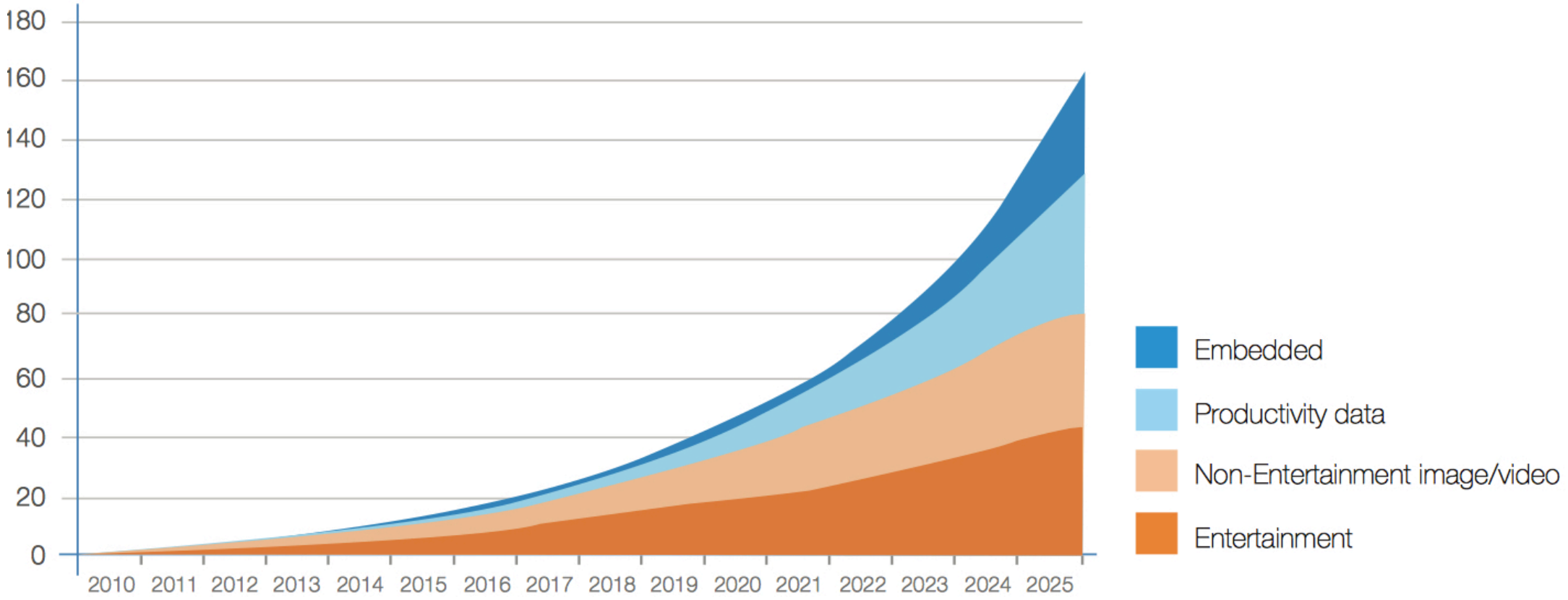
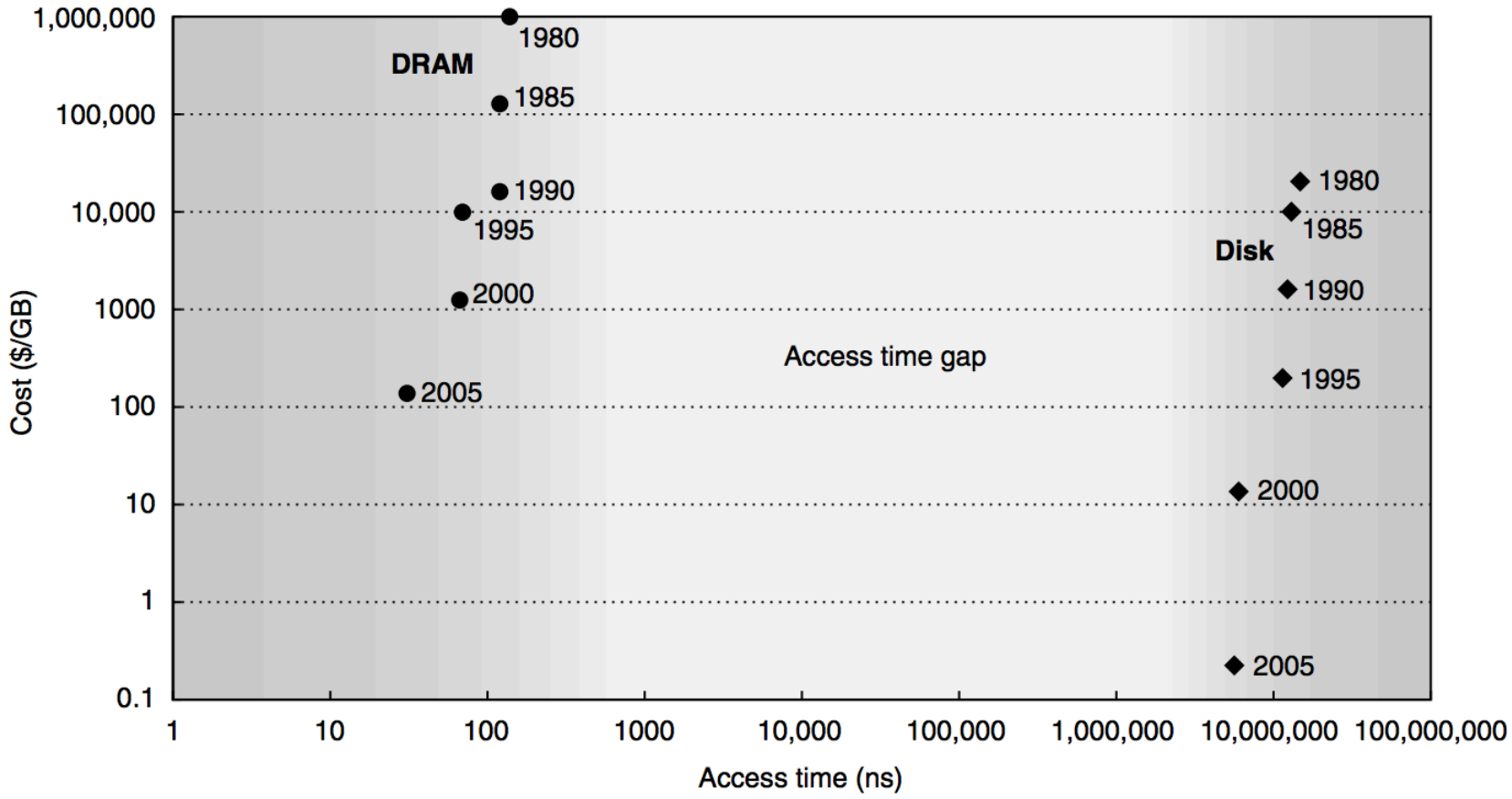


Image courtesy - https://www.storagenewsletter.com/2017/04/05/total-ww-data-to-reach-163-zettabytes-by-2025-idc/

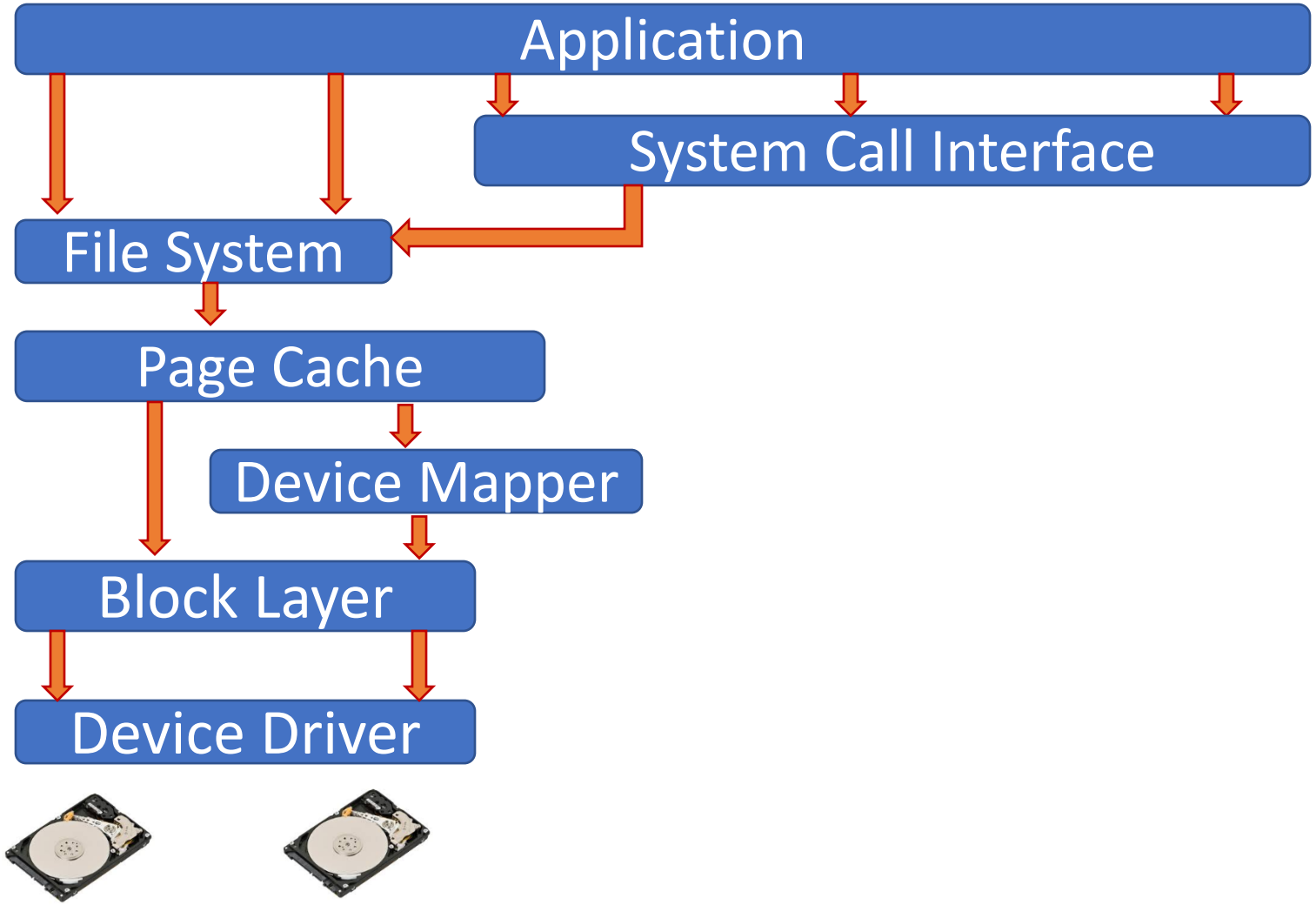
Cost versus access time for DRAM and HDD



Storage Systems: First Things First

- Storage devices are accessed differently than memory devices
- Access granularity
  - Block – minimum 512B, typically 4 – 8 KB
- Access is usually through a deeper, software stack
- Much higher access latencies
  - Milliseconds vs ns
- Interfaces are slower than everything that we have seen so far
  - SATA, SAS, PCIe
- Metrics for comparison
  - Latency, Bandwidth, IOPS

The Simplified I/O Software Stack



I/O Workloads Considerations

- Random and Sequential

What Does “Random” Mean?

KEYS ARE ALL OVER

A QUICK BROWN FOX JUMPED OVER A LAZY DOG

IMAGINE THAT THE KEYBOARD IS A DISK DRIVE

Slide courtesy https://www.snia.org/sites/default/files/ESF/Performance_Benchmarking_3_Block_Components_FinalPPT.pptx

What Does “Sequential” Mean?

EVERY KEY IS NEXT TO PREVIOUS

1 2 3 4 5 6

IMAGINE THAT THE KEYBOARD IS A DISK DRIVE

Slide courtesy https://www.snia.org/sites/default/files/ESF/Performance_Benchmarking_3_Block_Components_FinalPPT.pptx

“Sequential Read” Example

Slide courtesy https://www.snia.org/sites/default/files/ESF/Performance_Benchmarking_3_Block_Components_FinalPPT.pptx

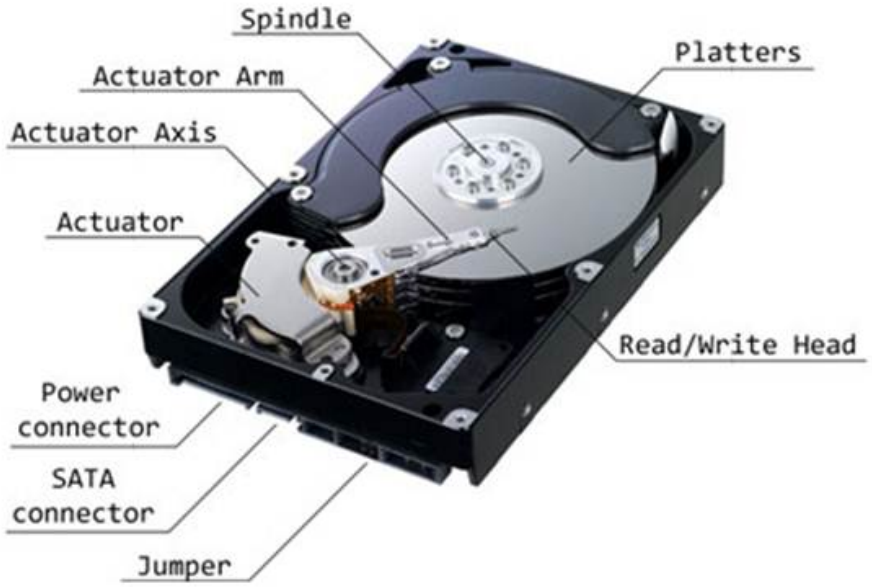
Jim Gray saw it coming!

Tape is Dead
Disk is Tape
Flash is Disk
RAM Locality is King

Jim Gray
Microsoft
December 2006

Magnetic Disks/ Hard Disk Drives/ HDDs

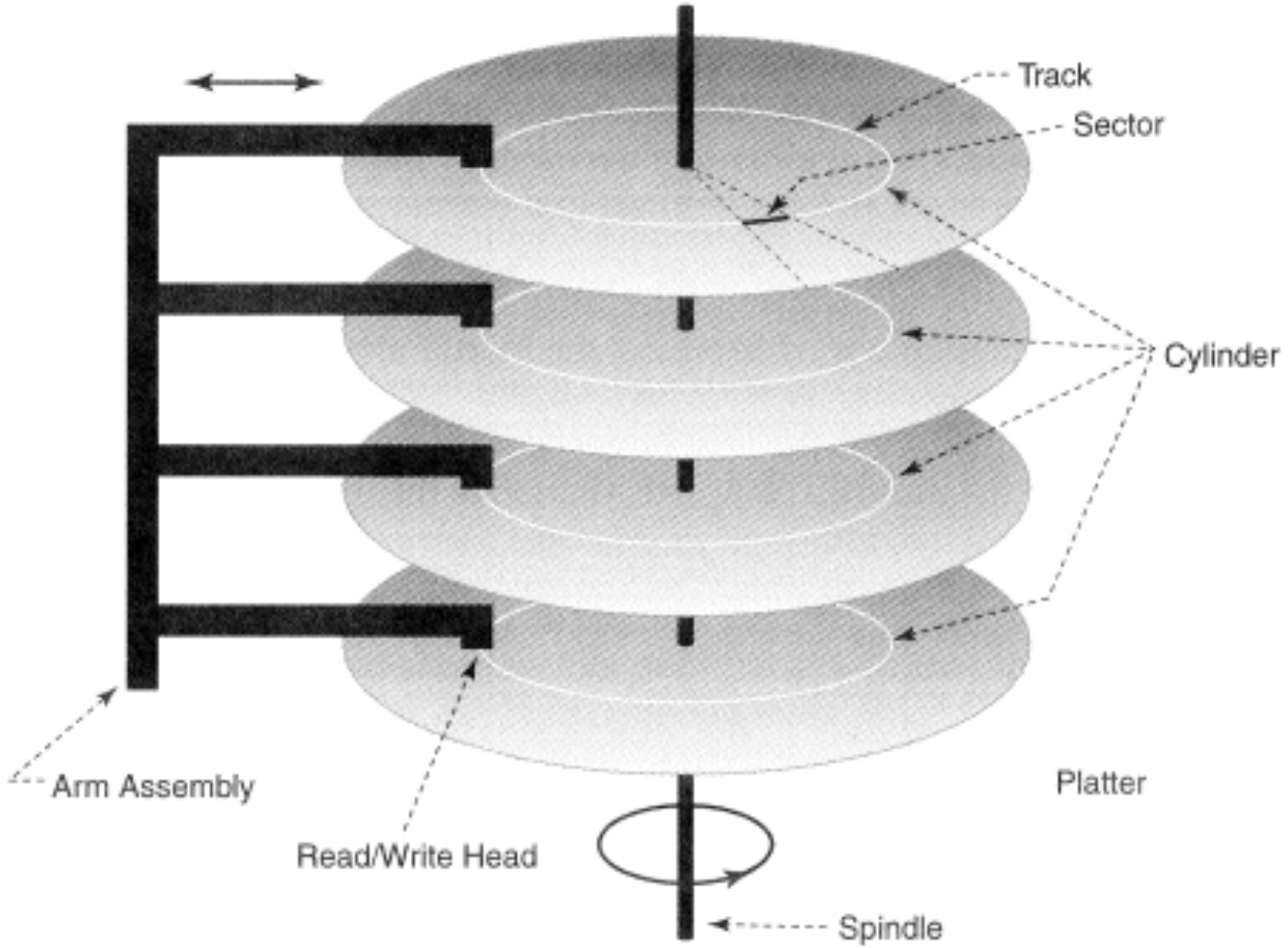
- A magnetic disk with *platters* (magnetic recording material on both sides
- Each platter has concentric *tracks* (5 - 30K); each track divided into *sectors* (512 B)
- A movable arm holds the heads for each disk surface and moves them in tandem



http://nptel.ac.in/courses/115103038/module4/lec28/images/image001.jpg

Slide courtesy: Rajeev Balasubramonian

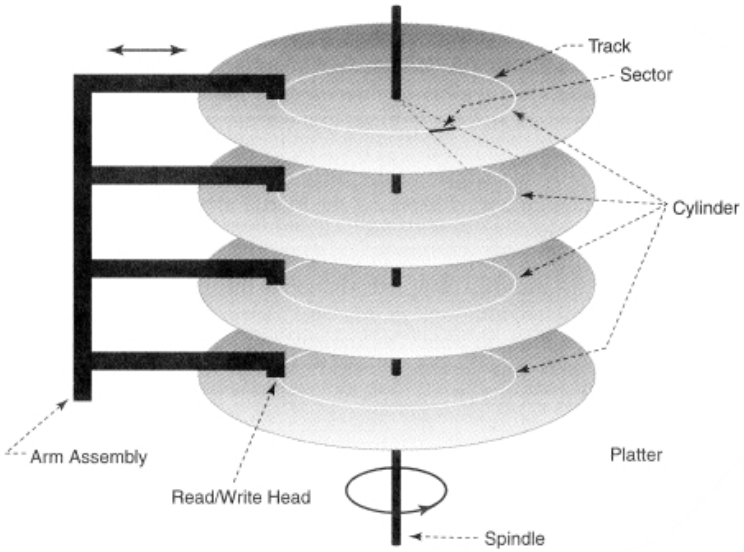
Hard Disk Drive



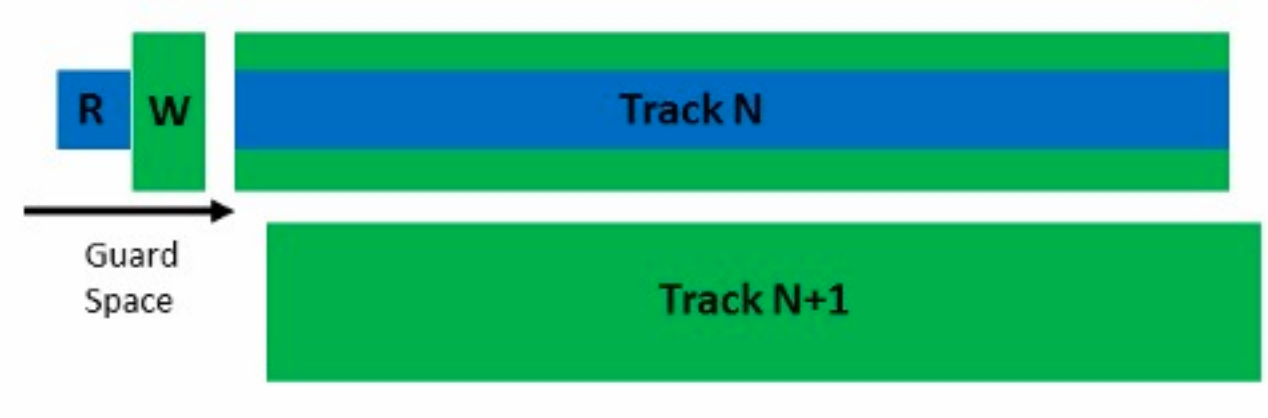
http://jcsites.juniata.edu/faculty/rhodes/dbms/diskorg.htm

Disk Latency

- Seek latency: Time to move the arm to the correct track; takes 5 to 12 ms on average; can take less if there is spatial locality
- Rotational latency : time taken to rotate the correct sector under the head; typically ~ 2 ms (15,000 RPM)
- Transfer time: time taken to transfer a block of bits out of the disk and is typically 100s MB/second
- Other overheads, depending on disk design



Shingled Magnetic Recording HDDs



Typical Drive

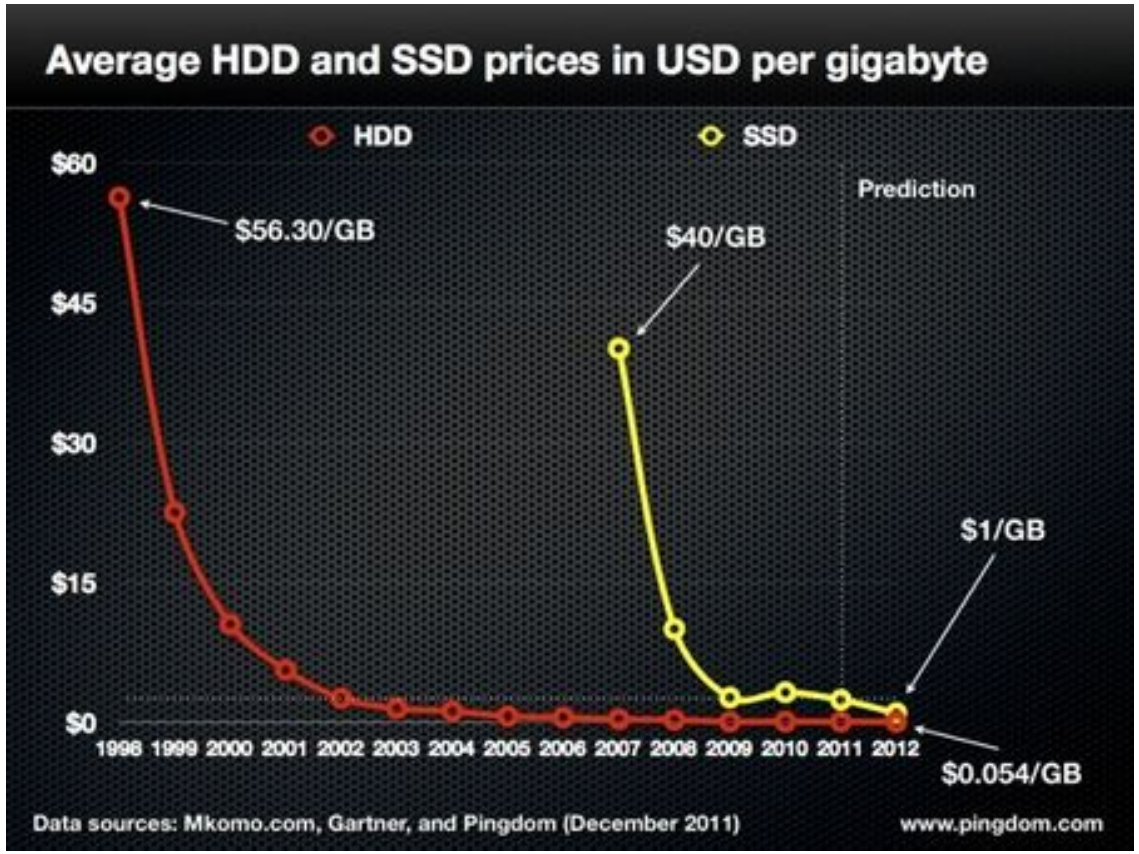
SMR Hard Disk Drive

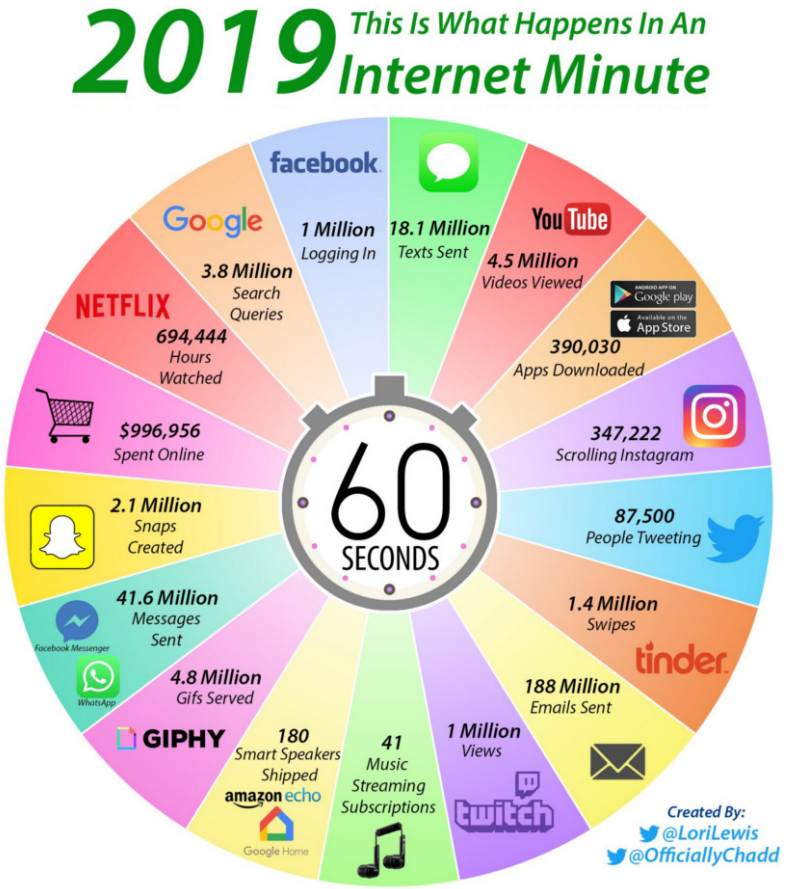
- Reduce the guard space between the tracks to increase density
- Overlapping tracks look like roof shingles, hence the same

Figure courtesy - http://www.enterprisestorageforum.com/storage-hardware/shingled-drives-for-re-roofing-your-storage-1.html

SSDs – Why now?

- NAND Flash is around since 80s – why the increase in interest?

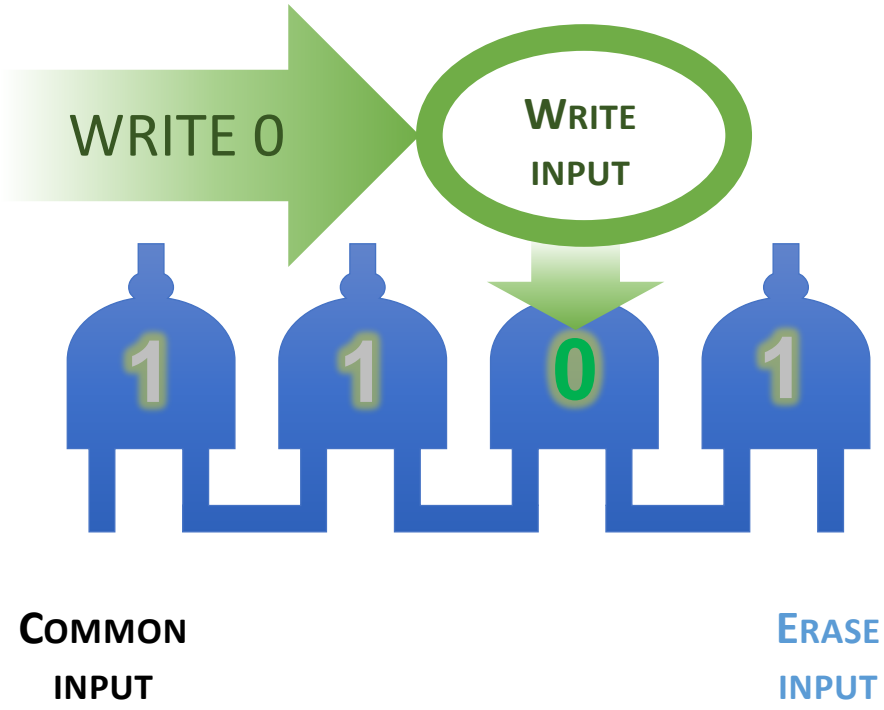




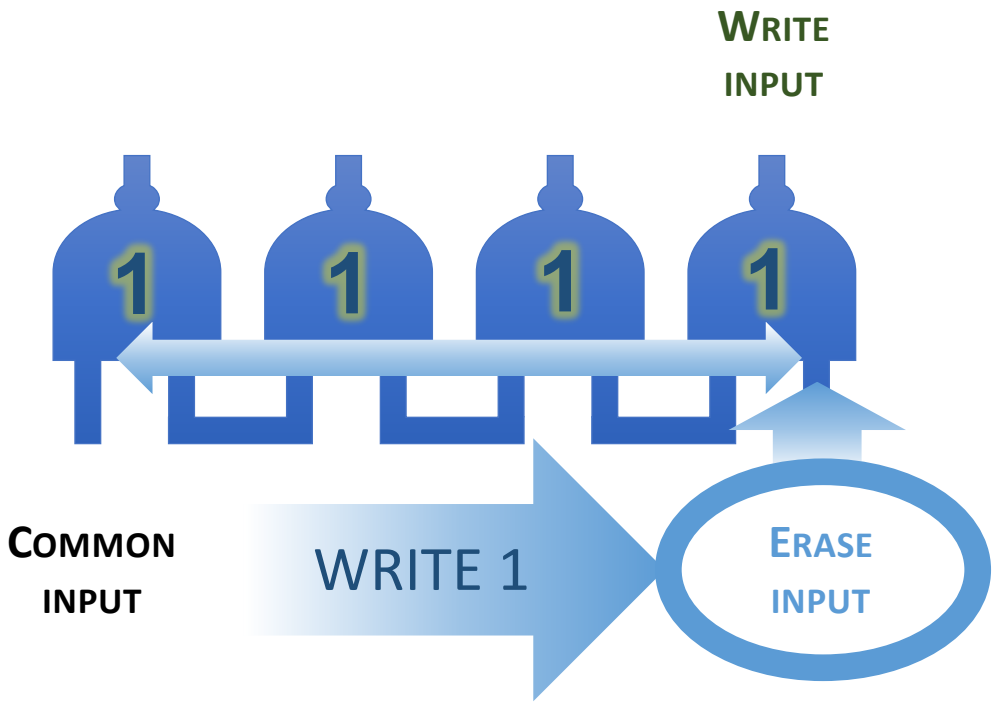
Slide courtesy : Michael Freedman, COS 518: *Advanced Computer Systems* Lecture 8, Harvard University

Flash And NAND Gates

EVERY NAND CAN BE SET TO 0 INDIVIDUALLY



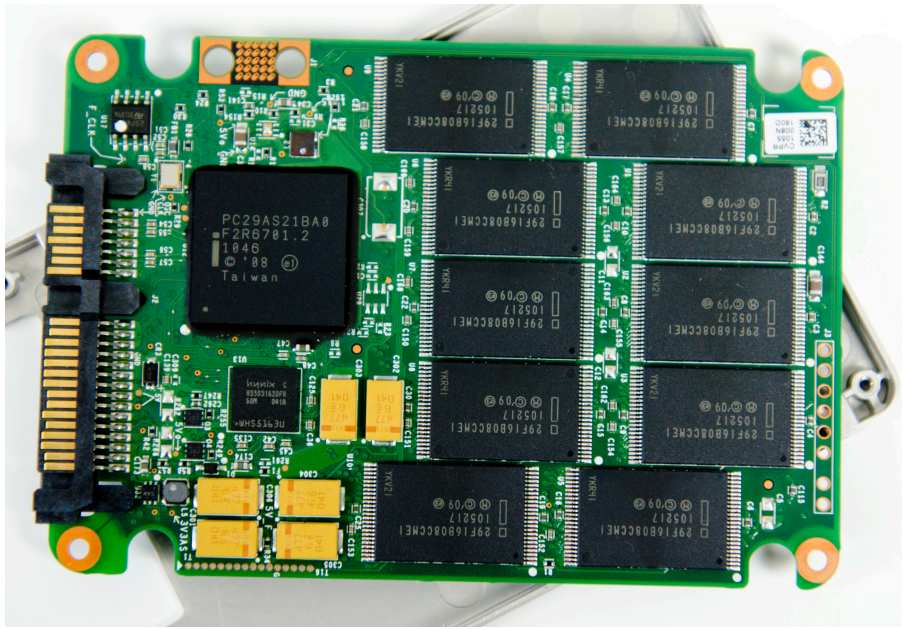
TO SET BACK TO 1, AN ENTIRE GROUP NEEDS TO BE RESET



Slide courtesy https://www.snia.org/sites/default/files/ESF/Performance_Benchmarking_3_Block_Components_FinalPPT.pptx

NAND Flash: Architecture

- Pages: 4-16 KB, assembled into -
- Blocks: 128KB, 256KB, 4MB, 8 MB



Block 1000 (data)

PPN	data
0	x
1	y
2	z
3	

Block 2000 (free)

PPN	data
0	
1	
2	
3	

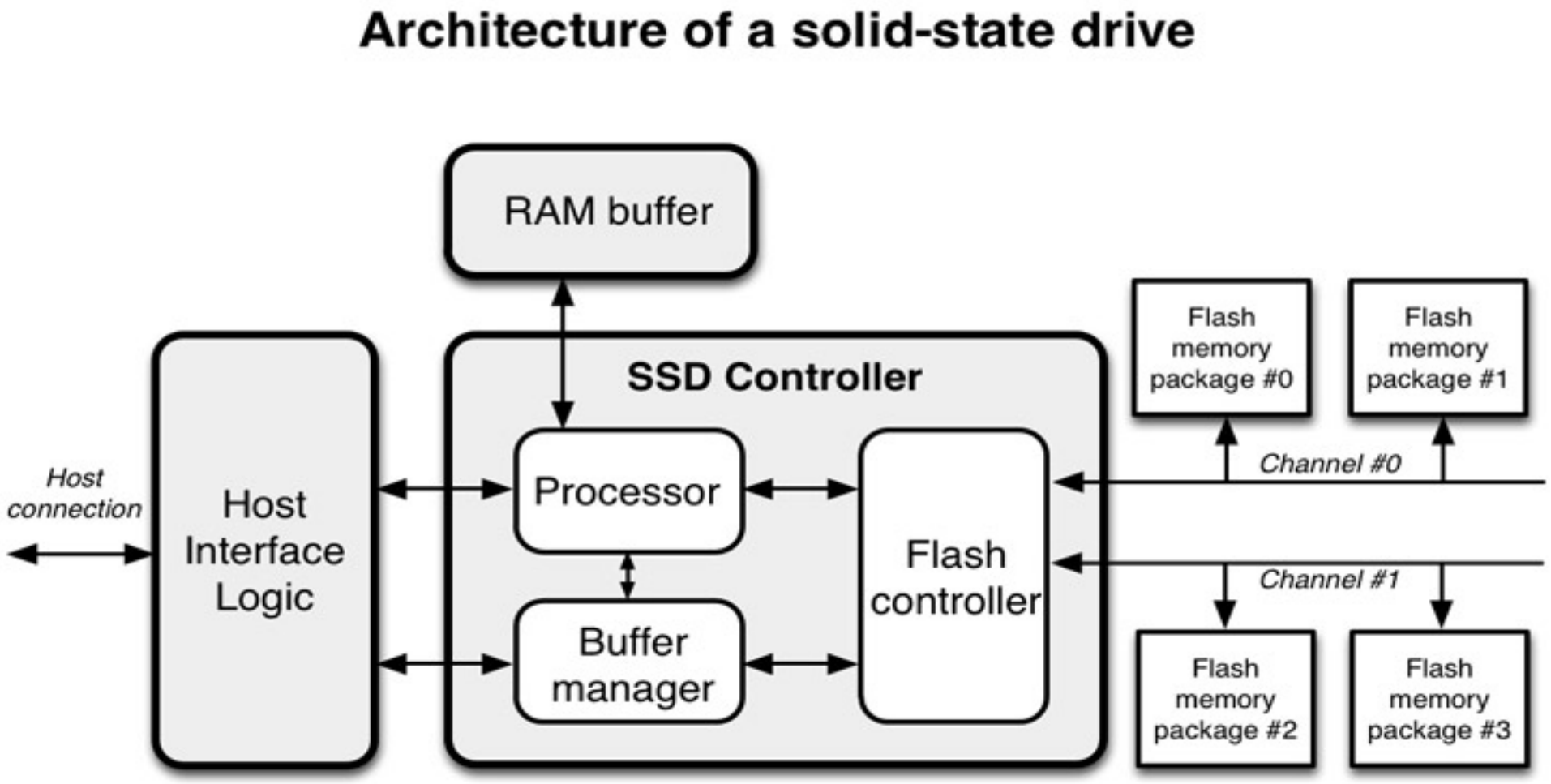
Slide courtesy : Michael Freedman, COS 518: *Advanced Computer Systems* Lecture 8, Harvard University

Flash Operations

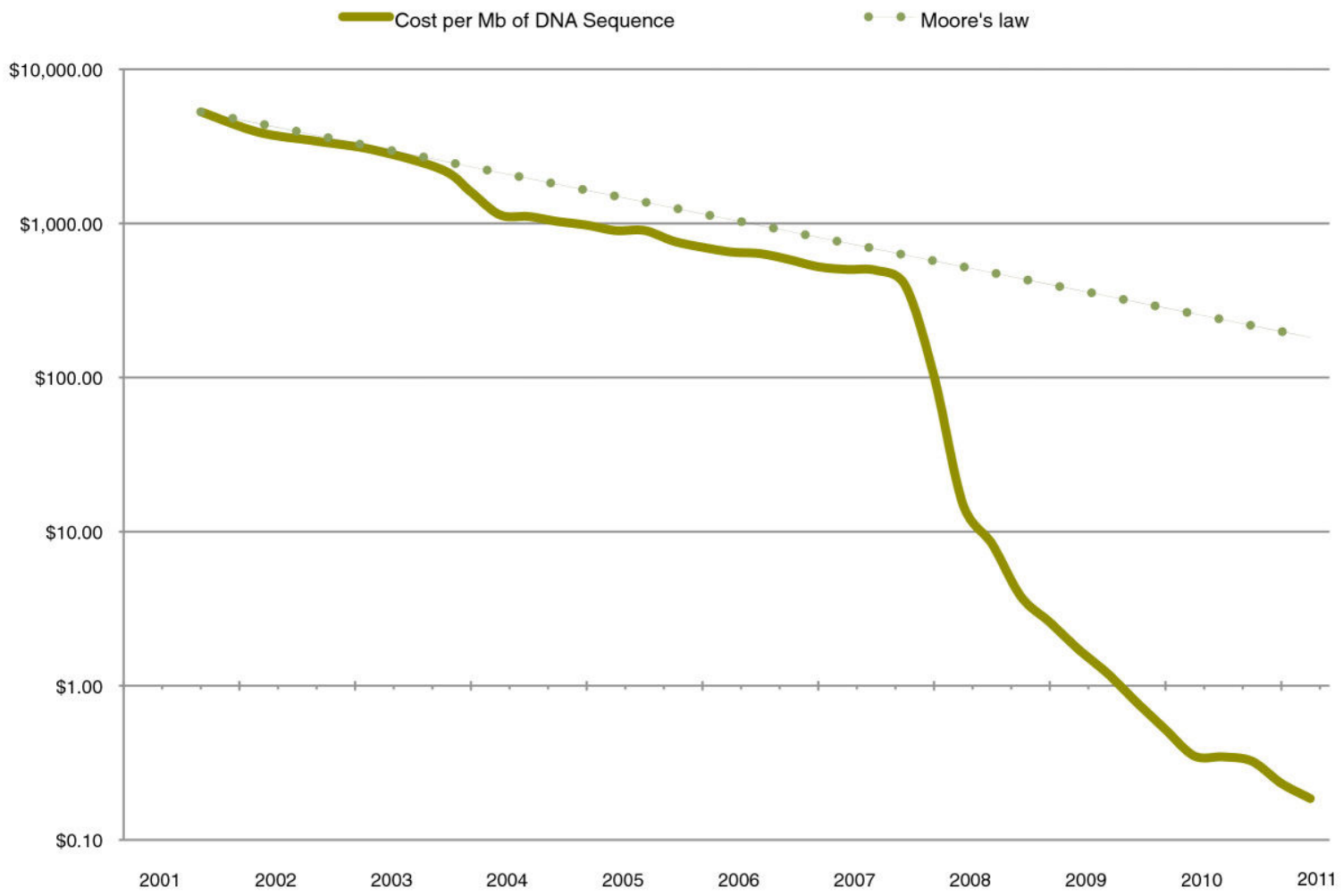
- Read (Page)
  - Read any page in device
  - Fast - 10s of microseconds, irrespective of page number
- Erase (Block)
  - To write a page, the entire block must be *erased*
  - Implication – all data needs to be copied safely before erase
  - Slow – few milliseconds to complete
- Program (Page)
  - Can write to a page after block erase
  - Fast(er) - 100s of microseconds

Device	Read (μs)	Program (μs)	Erase (μs)
SLC	25	200-300	1500-2000
MLC	50	600-900	~3000
TLC	~75	~900-1350	~4500

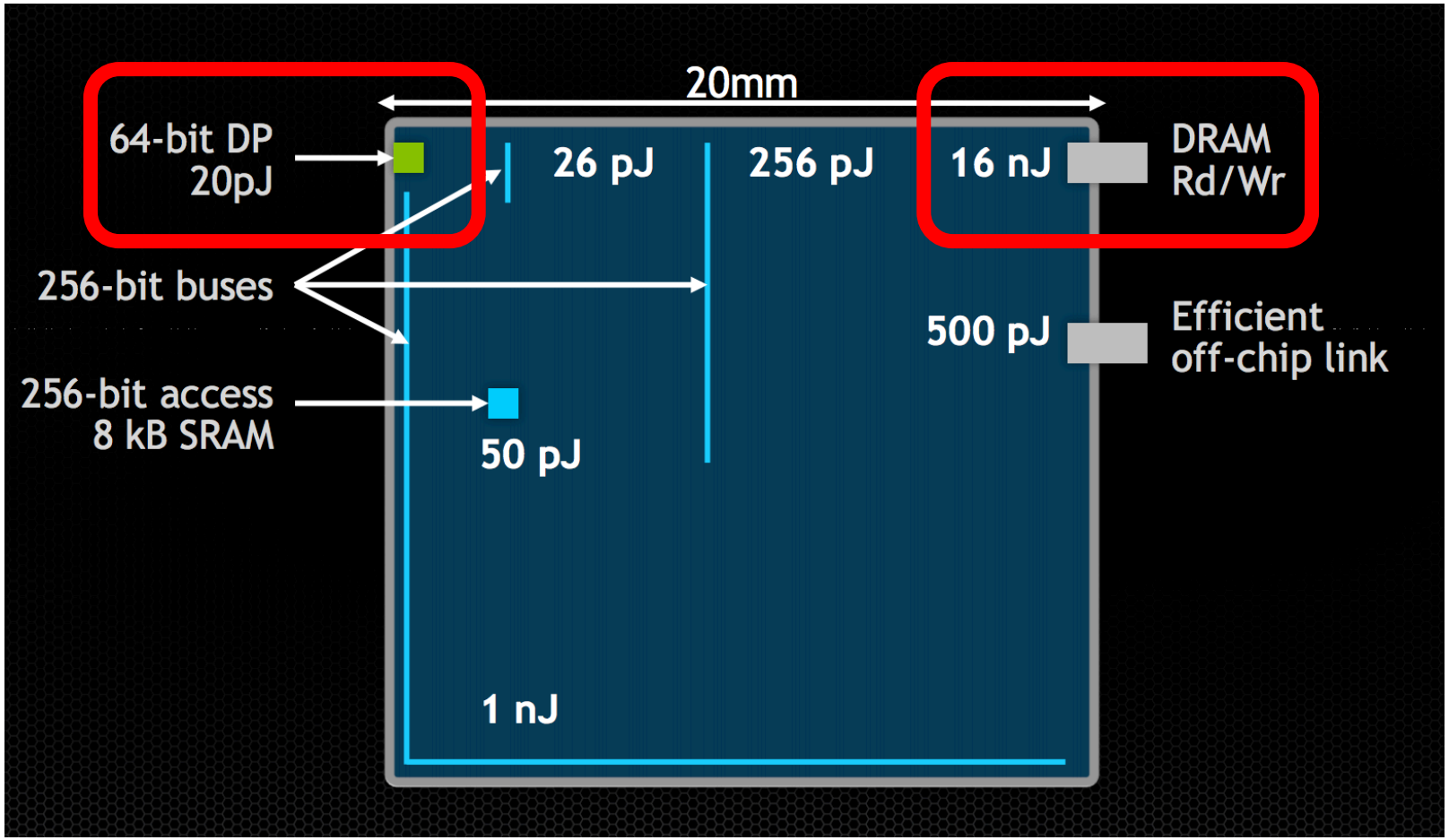
The Solid State (Storage) Device



The Genomics Use Case : Need for Compute



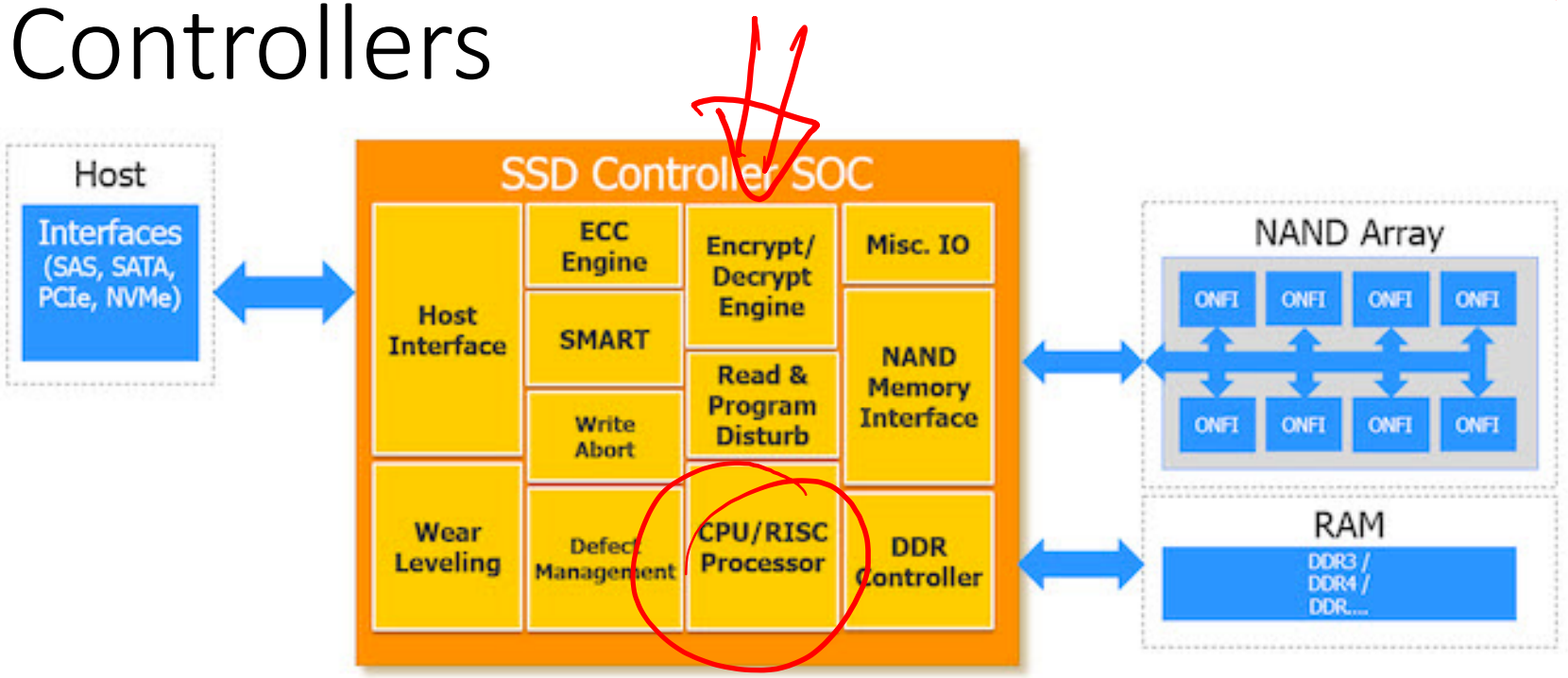
Data Movement: Energy Argument



Moving Data is **3 orders of magnitude** costlier than computation

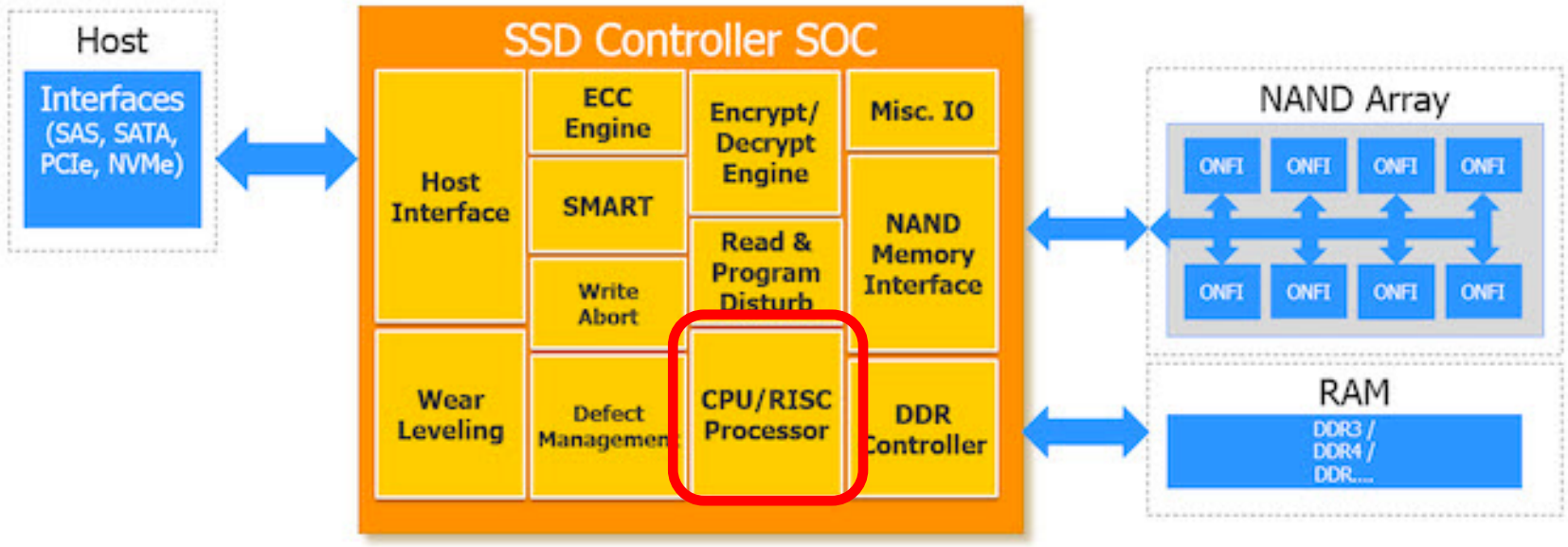
Bill Dally, SC 2015

SSD Controllers



- "Complete" subsystems, with non-trivial compute capabilities
- Currently carrying out SSD management related tasks
- What if we could add, more specialized compute for multiple use cases?

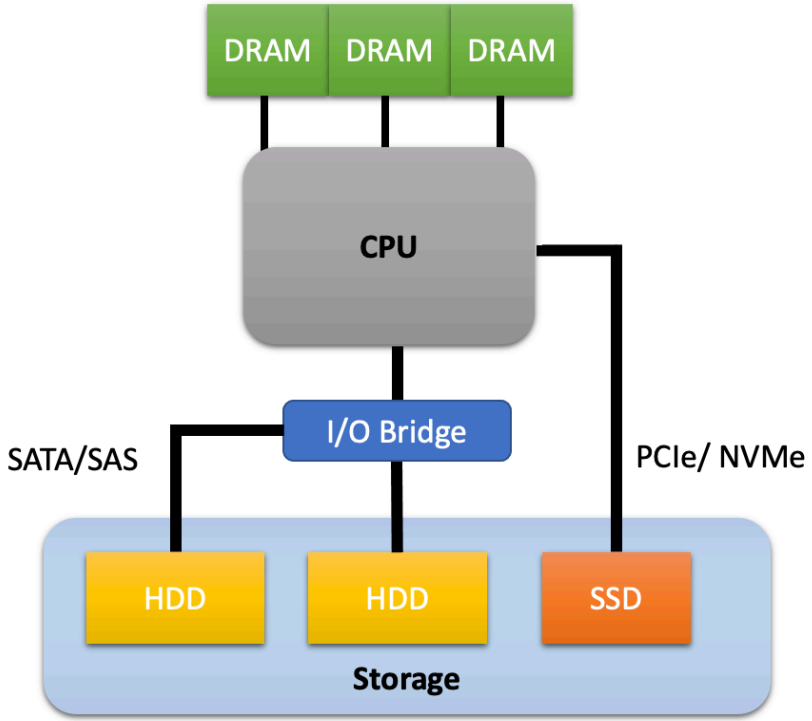
Near Storage Computing : Challenges



- What type of ***additional*** compute should be added?
- Design of APIs to ship off work to specialized compute structures
- How do you divide work between the host and compute in the controller?

Key Research Directions

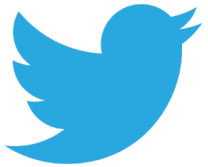
- The right place for storage: memory bus or the I/O bus?
- Near Storage Compute
- Workload Characterization and Visualization
- Mechanisms to “ship off” work



Acknowledgements

Varun Gohil, Shreyas Singh, Tom Glint Issac, Sneha Ved, Subisha V,
Nisarg Ujjainkar, Chandan Kr. Jha


manu.awasthi@ashoka.edu.in


@mnwsth