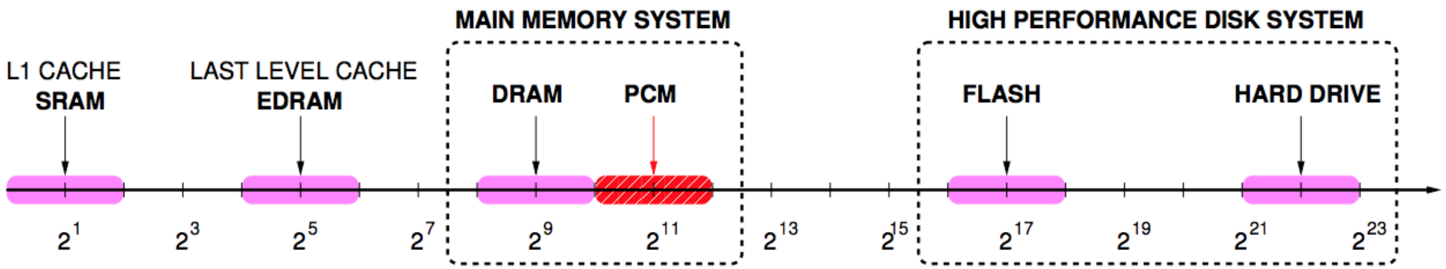
Storage Systems : Disks and SSDs

Manu Awasthi

July 6th 2018

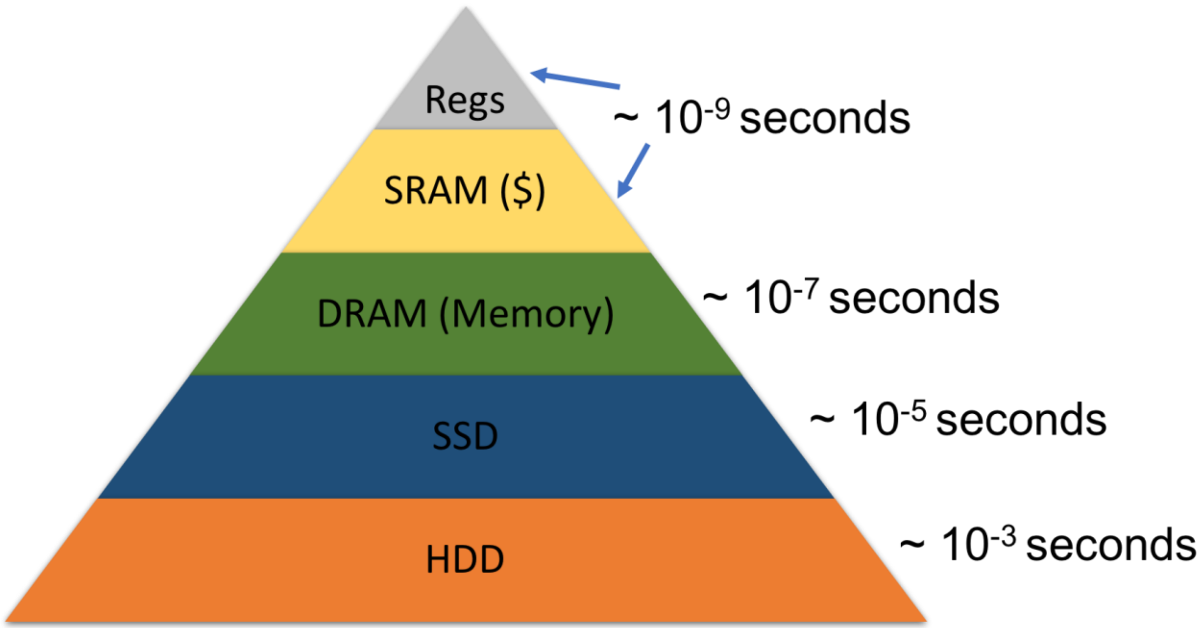
Computer Architecture Summer School 2018

Why study storage?

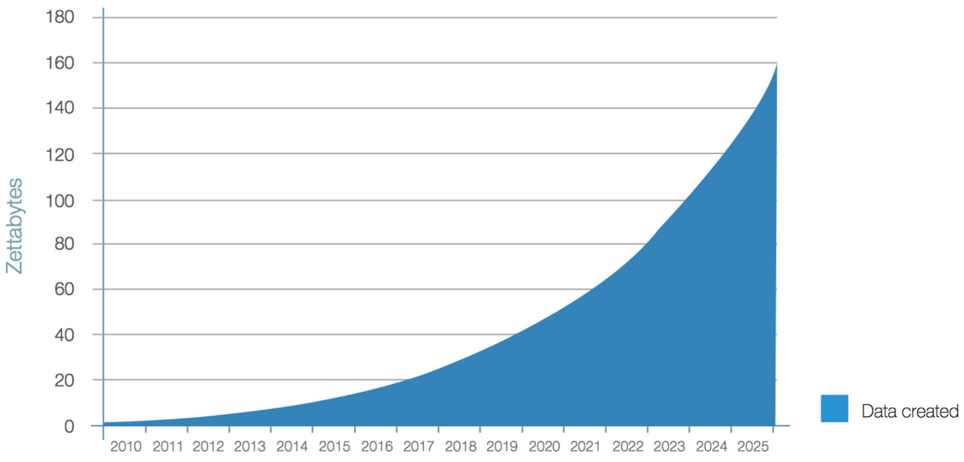


Typical Access Latency (in terms of processor cycles for a 4 GHz processor)

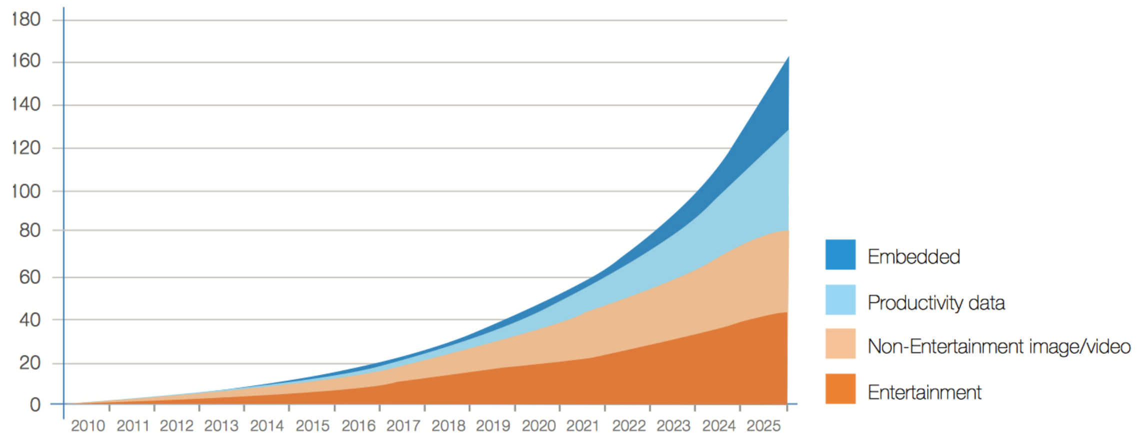
Scalable High Performance Main Memory System Using Phase-Change Memory Technology, Qureshi et al , ISCA 2009



Trends



Total amount of data created



Type of data

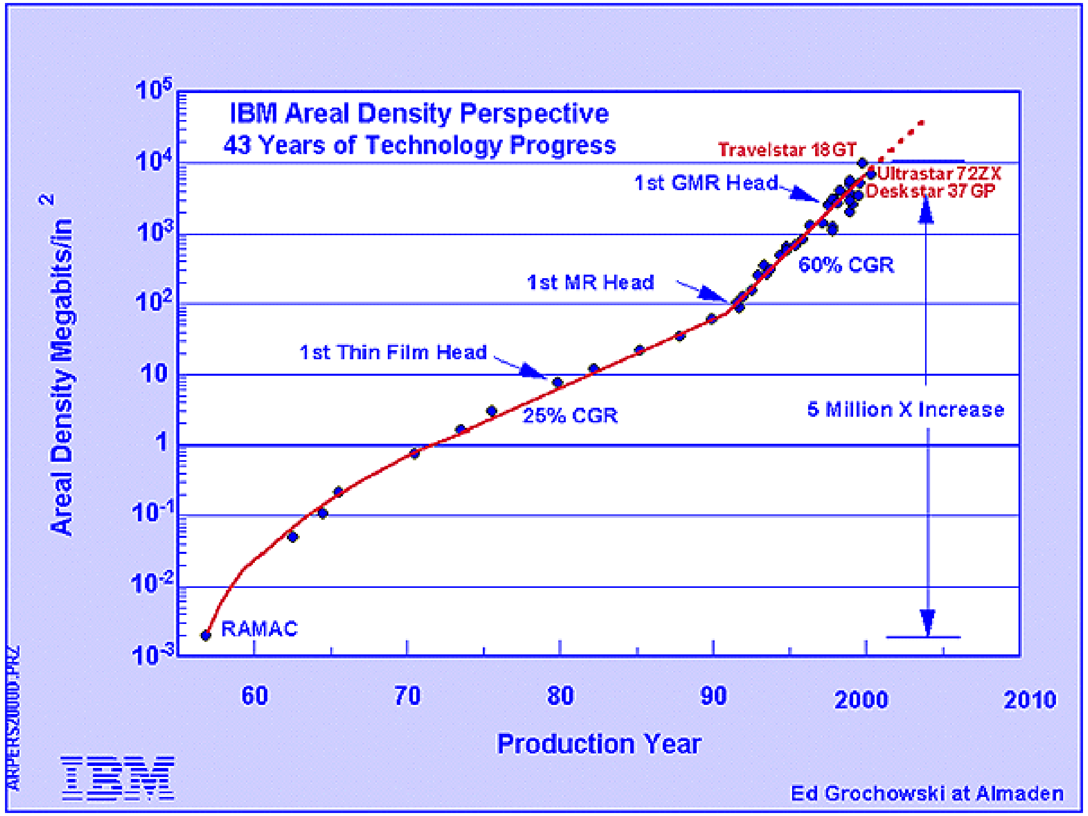
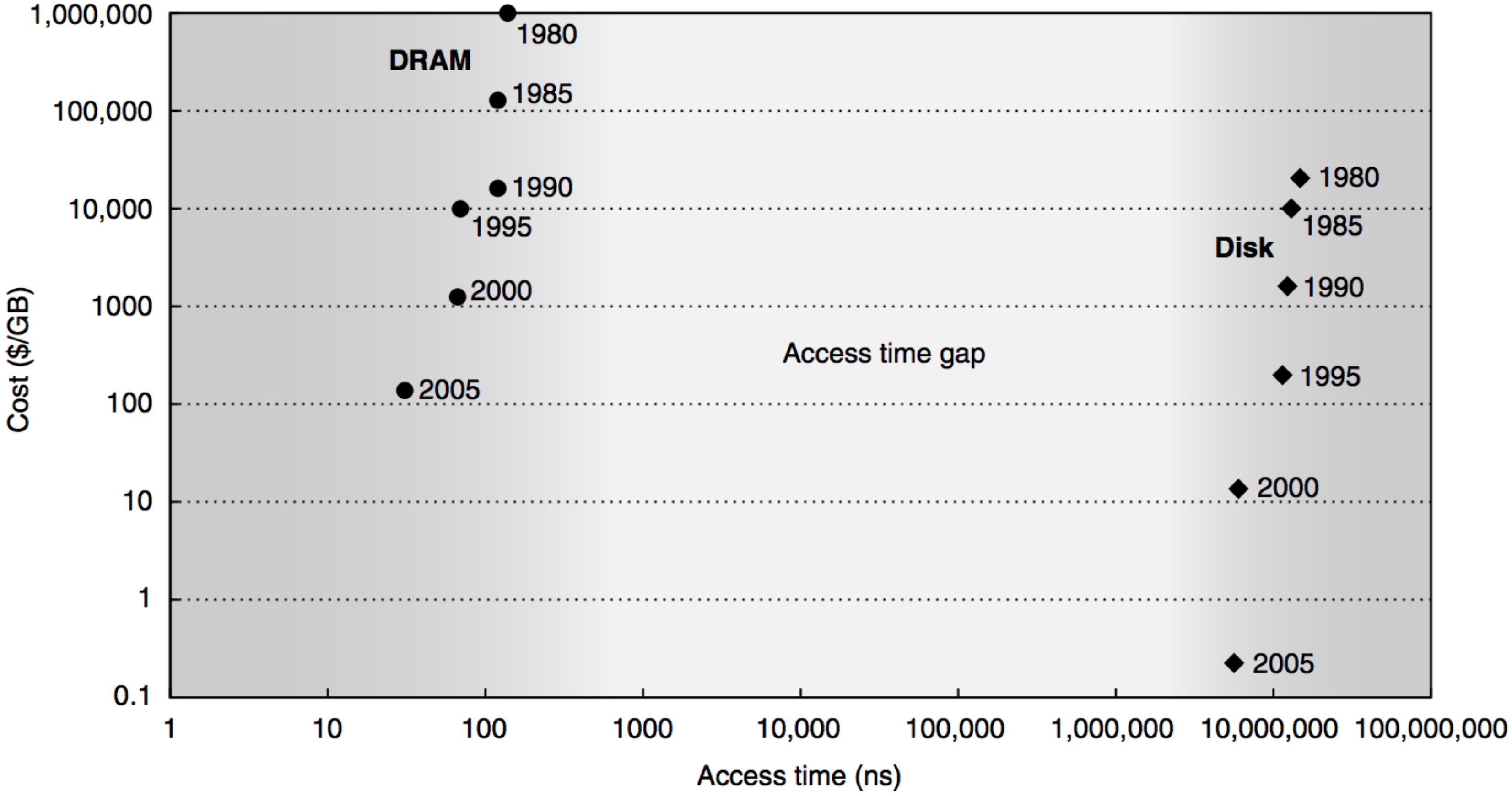
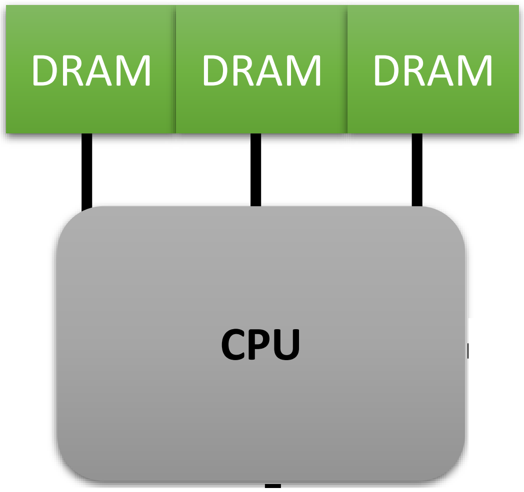


Image courtesy - https://www.storagenewsletter.com/2017/04/05/total-ww-data-to-reach-163-zettabytes-by-2025-idc/

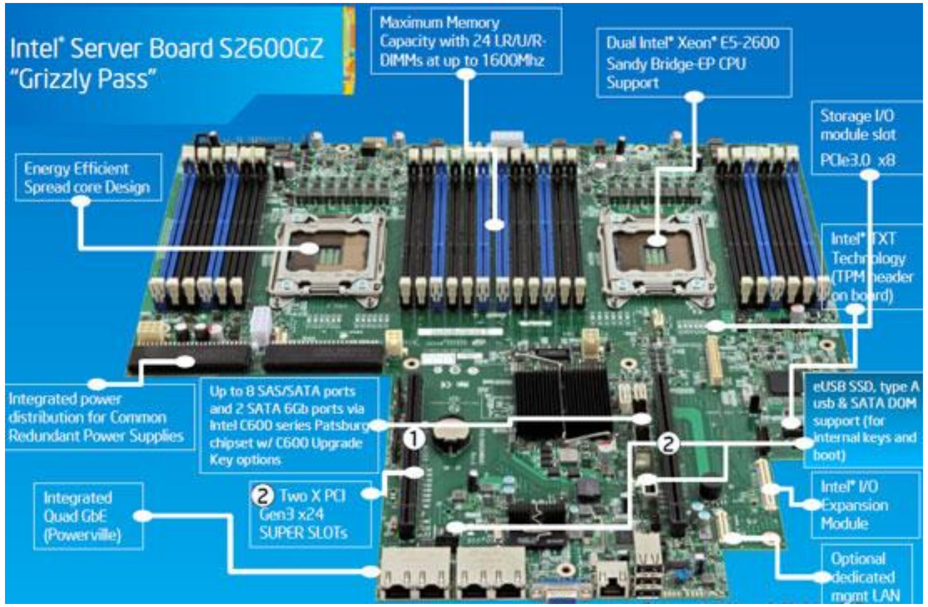
Cost versus access time for DRAM and HDD

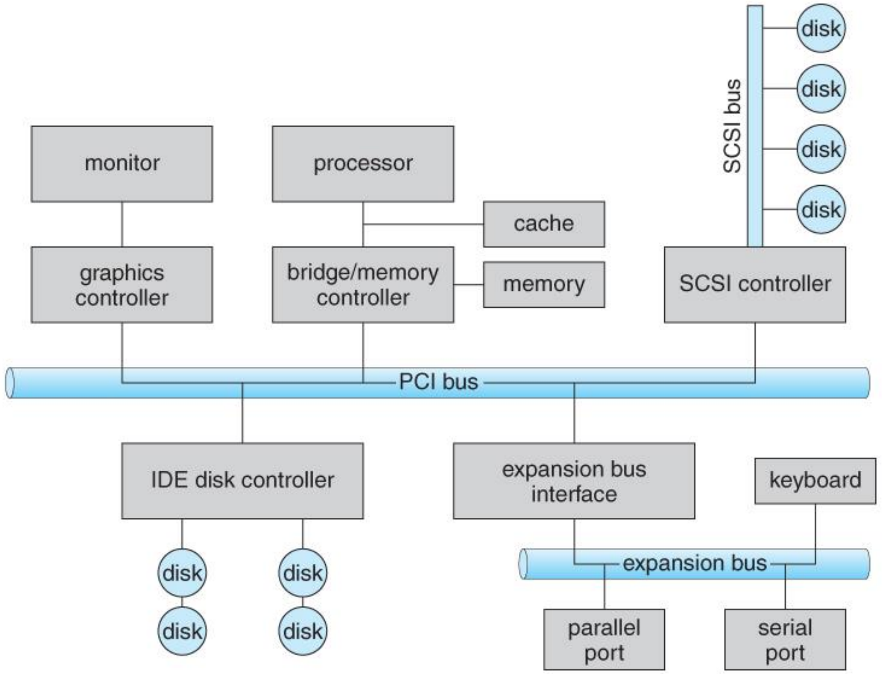


Storage Systems : Today



System Organization





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

I/O Workloads Types

- Random and Sequential

What Does “Random” Mean?

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?

1 2 3 4 5 6

IMAGINE THAT THE KEYBOARD
IS A DISK DRIVE

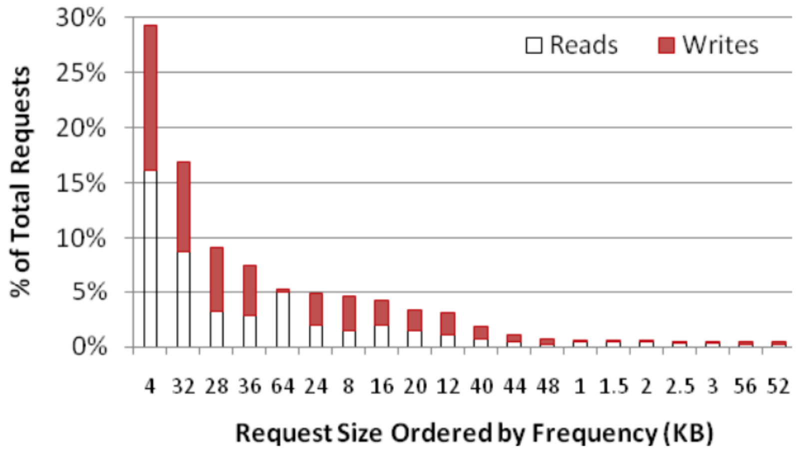
Slide courtesy https://www.snia.org/sites/default/files/ESF/Performance_Benchmarking_3_Block_Components_FinalPPT.pptx

“Sequential Read” Example

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

I/O Workloads Types

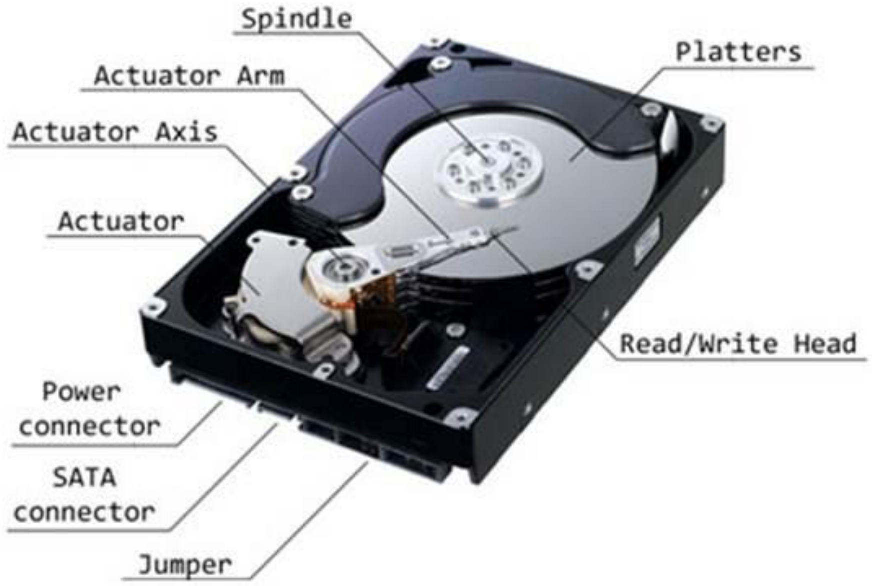
- Random and Sequential
- Most workloads are a combination of the above
  - X% sequential; (100-x)% random, or some combination thereof
  - Varies over time as well
- Most workloads also have varying granularity of request sizes



Picture courtesy - Characterization of Storage Workload Traces from Production Windows Servers Swaroop Kavalanekar et al., IISWC 2008

Magnetic Disks/ Hard Disk Drives/ HDDs

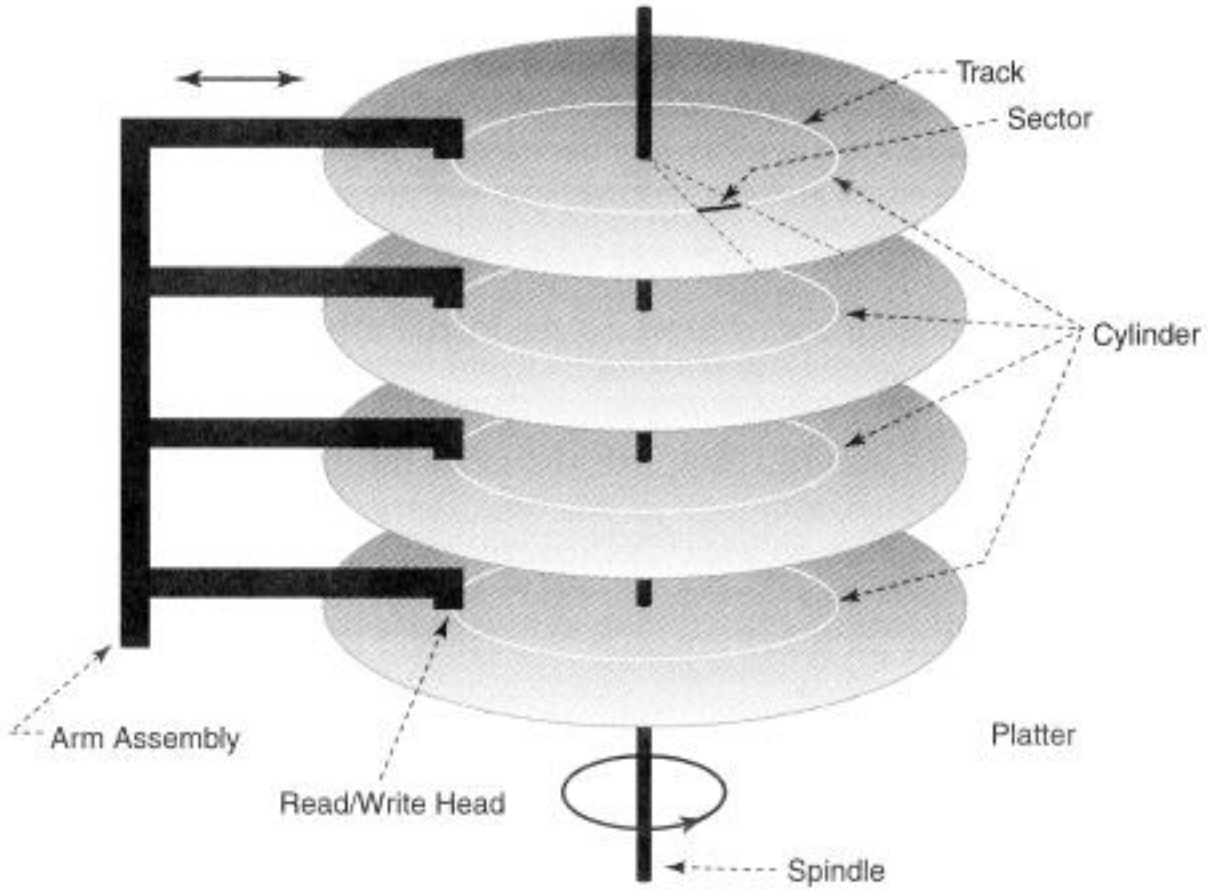
- A magnetic disk consists of 1-12 *platters* (metal or glass disk covered with magnetic recording material on both sides), with 1-3.5 inch diameter
- Each platter is comprised of concentric *tracks* (5 - 30K) and each track is divided into *sectors* (100 – 500 per track, each about 512 bytes)
- A movable arm holds the read/write heads for each disk surface and moves them all in tandem – a *cylinder* of data is accessible at a time



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

Slide courtesy: Rajeev Balasubramonian

Hard Disk Drive

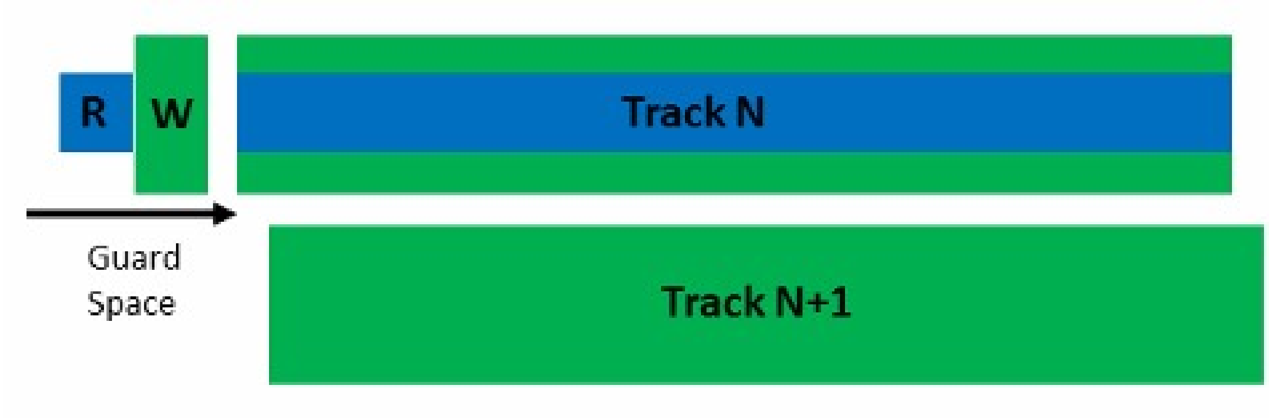


http://jcsites.juniata.edu/faculty/rhodes/dbms/diskorg.htm

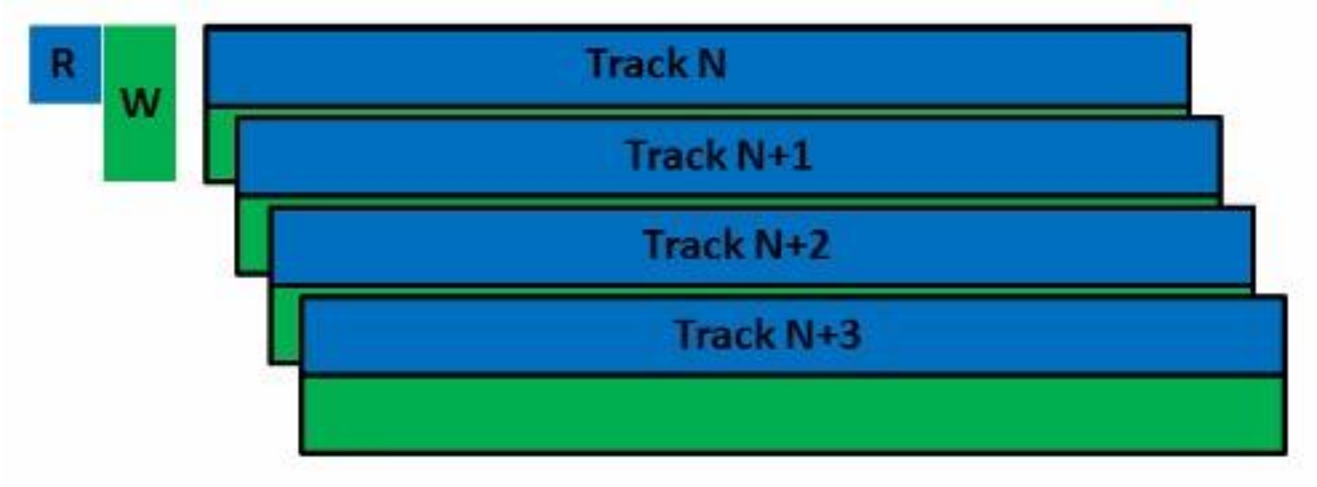
Disk Latency

- To read/write data, the arm has to be placed on the correct track – this *seek time* usually takes 5 to 12 ms on average – can take less if there is spatial locality
- *Rotational latency* is the time taken to rotate the correct sector under the head – average is typically more than 2 ms (15,000 RPM)
- *Transfer time* is the time taken to transfer a block of bits out of the disk and is typically 3 – 65 MB/second
- A disk controller maintains a disk cache (spatial locality can be exploited) and sets up the transfer on the bus (*controller overhead*)

Shingled Magnetic Recording HDDs



Typical Drive



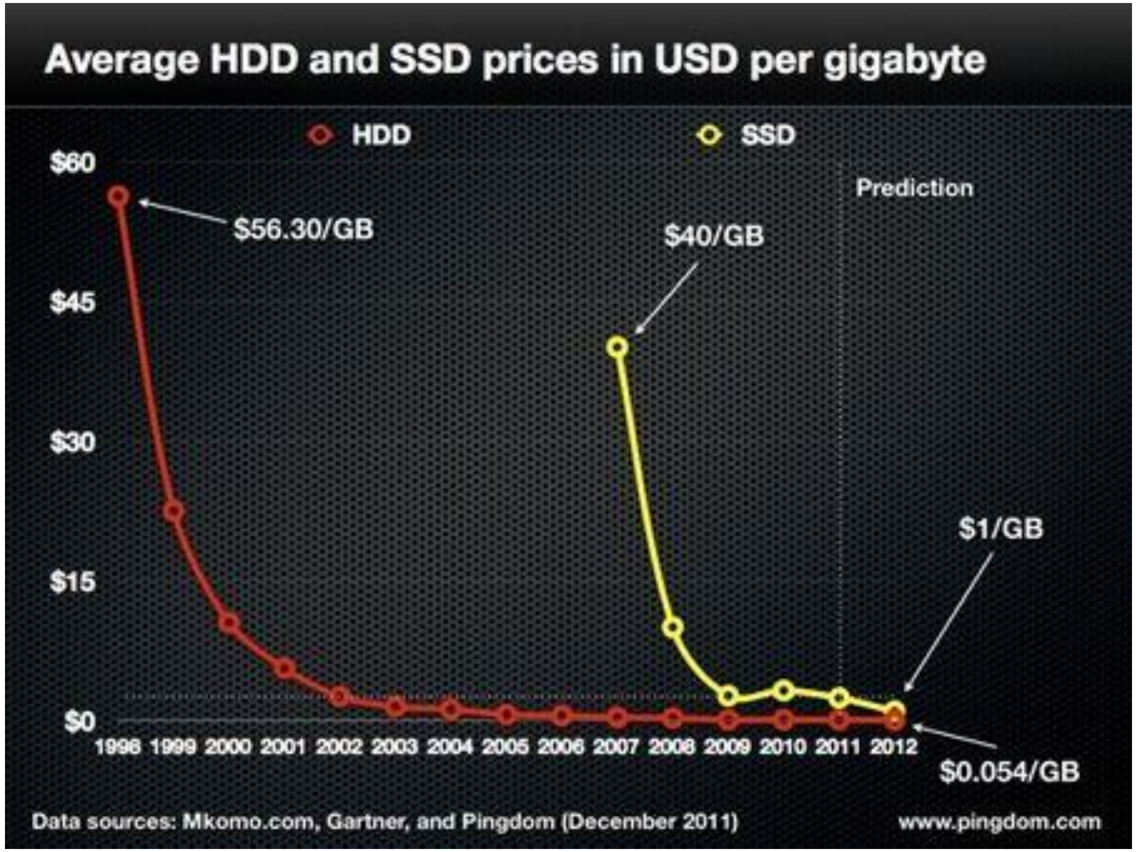
SMR Hard Disk Drive

- 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?

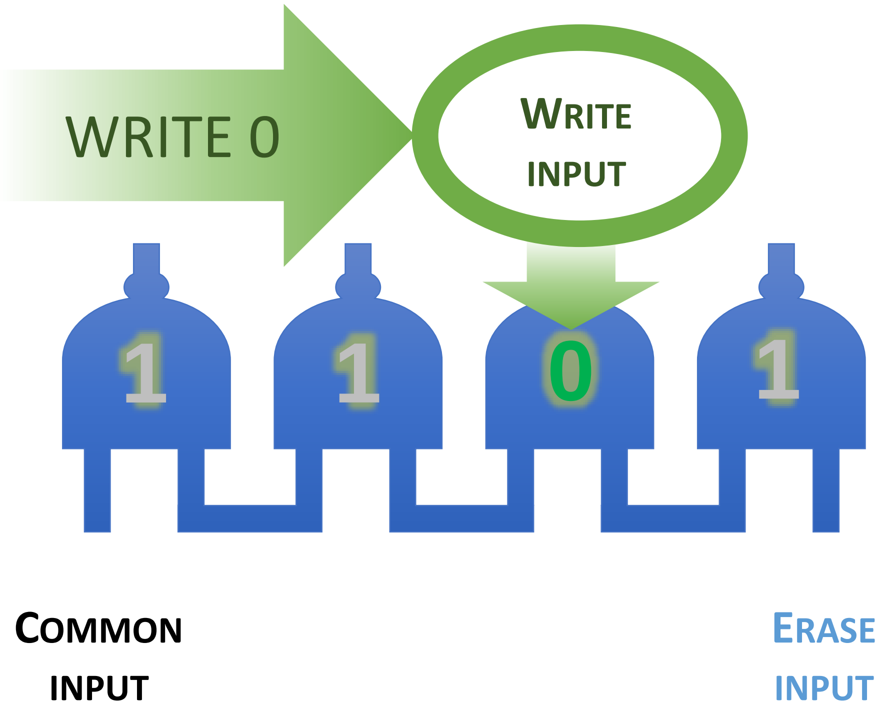
- NAND Flash has been around since 1980s – why the sudden increase in interest?



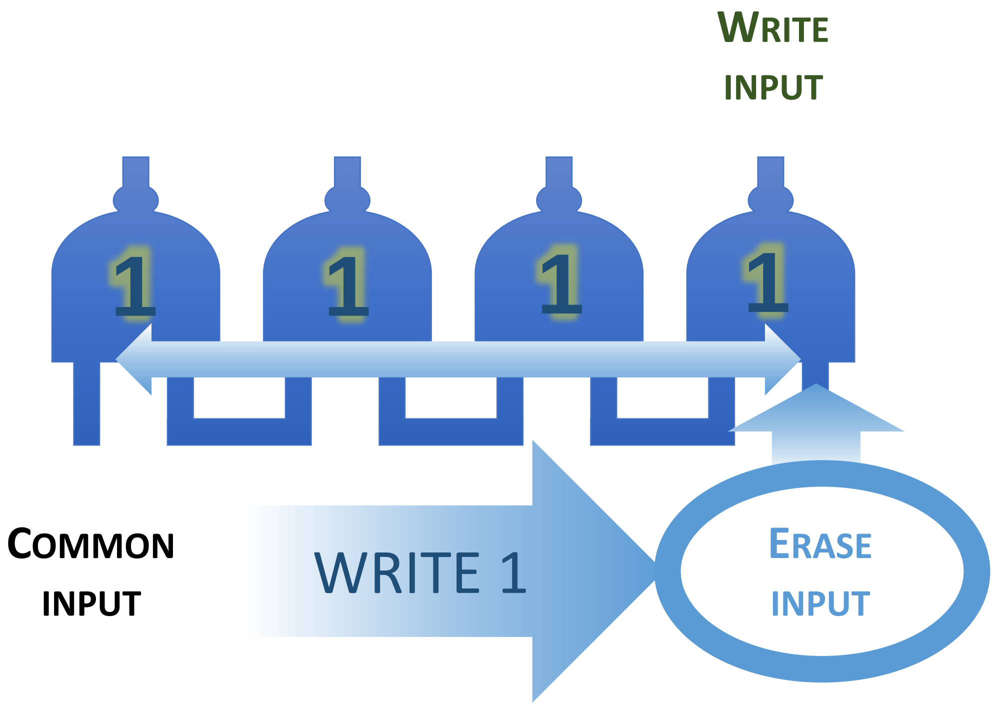
Slide courtesy : Michael Freedman, COS 518: *Advanced Computer Systems* Lecture 8, Harvard University

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

- **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)

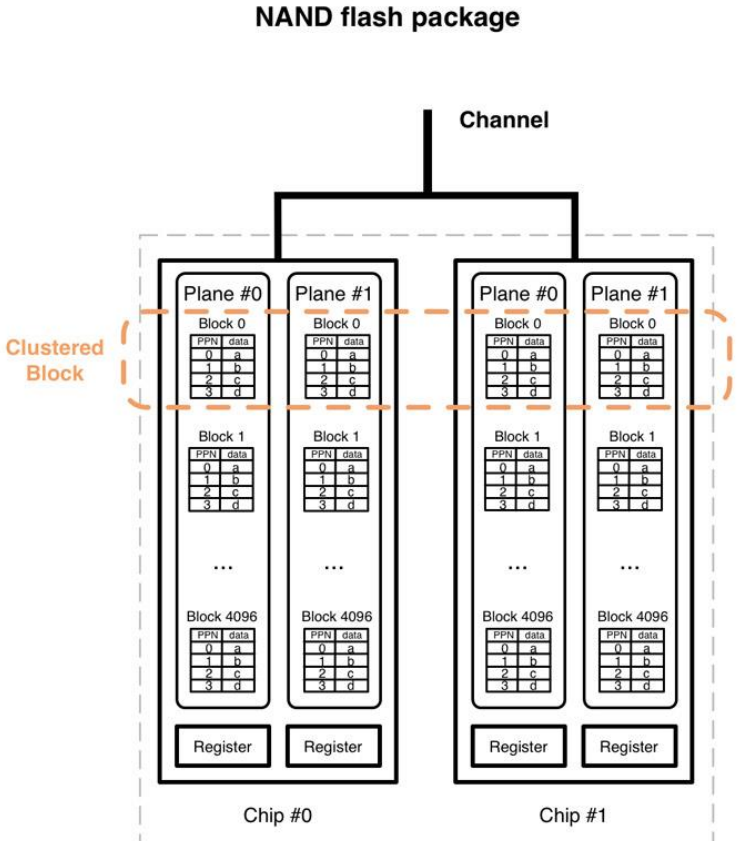
PPN	data
0	
1	
2	
3	

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

NAND Flash Basics

http://images.anandtech.com/reviews/storage/MidRange2011/DSC_3756.jpg

- Storing bits
  - Single Level Cell (SLC) – encodes single bit
  - Multi-level Cell (MLC) – 2 bits
  - Triple Level Cell (TLC) – three bits
- Levels of Organization
  - Banks/planes
  - Blocks/Erase Blocks (128-256 KB)
  - Page (4 KB)



http://codecapsule.com/2014/02/12/coding-for-ssds-part-4-advanced-functionalities-and-internal-parallelism/

NAND Flash: Reads/ Writes

- **Always read an entire page:**
    - Can only read entire aligned page from SSD
- **Always write an entire page:**
    - To change single byte, need to write entire page
- **Pages cannot be overwritten**
    - Page can be written only if the “free”/”erased” state.
    - **Updating**: **Read** page to internal register, **modify**, then **write** to free page
- **Erases are aligned on block size**
    - To make a page “free”, need to erase it
    - Erasures can only occur at block boundary

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 has to be *erased* – content destroyed
    - every bit changed to 1.
  - Implication – all data needs to be copied safely before erase
  - Slow – few millisecs 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

Example

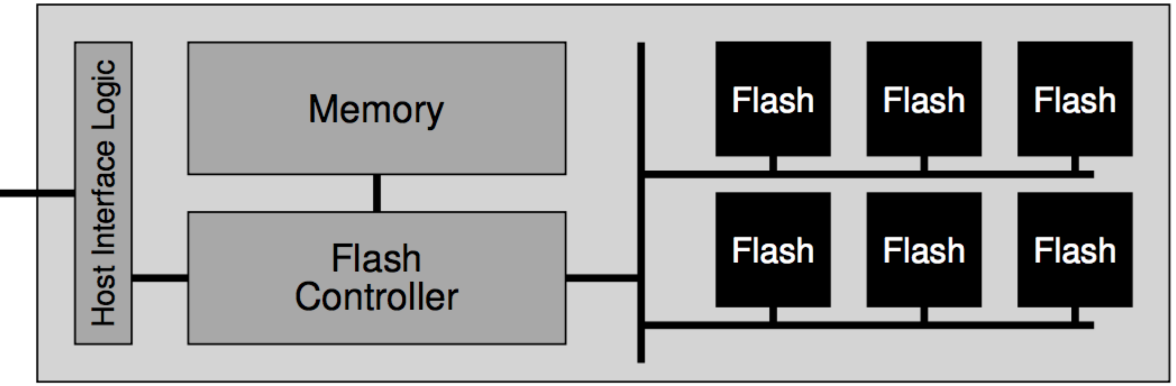
Page 0	Page 1	Page 2	Page 3
00011000	11001110	00000001	00111111
VALID	VALID	VALID	VALID

Page 0	Page 1	Page 2	Page 3
11111111	11111111	11111111	11111111
ERASED	ERASED	ERASED	ERASED

Page 0	Page 1	Page 2	Page 3
00000011	11111111	11111111	11111111
VALID	ERASED	ERASED	ERASED

Figure courtesy : Operating Systems: Three Easy Pieces

Flash Based SSDs



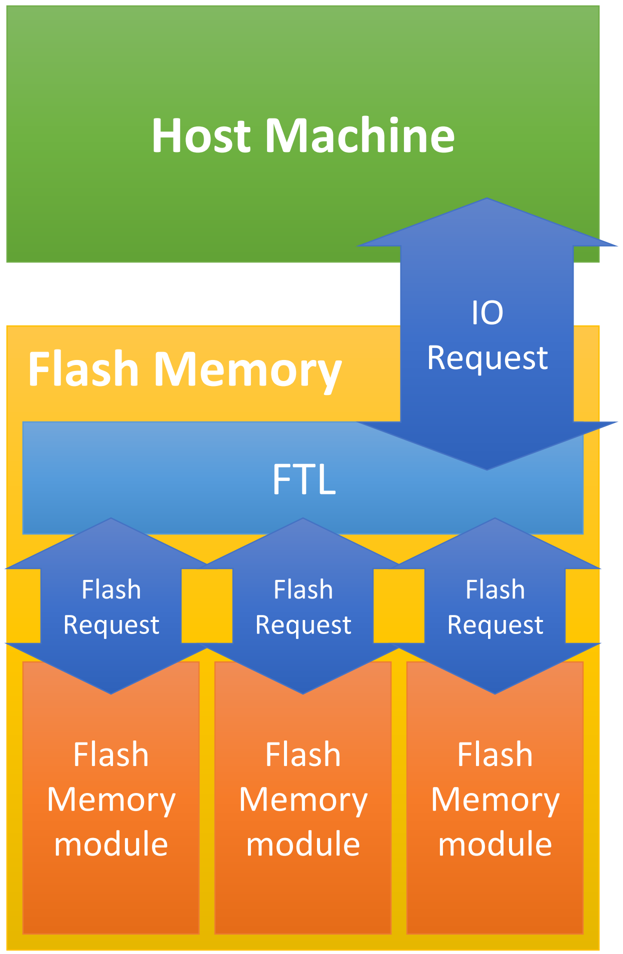
- Turning Raw Flash to a storage device requires
  - provide the standard block interface to applications/OS
- SSD contains
  - Flash chips (of course)
  - Volatile memory (SRAM/DRAM)
  - Control logic for operations
- FTL – Flash Translation Layer
  - Takes logical rd/wr ops; converts them into ops for device

Figure courtesy : Operating Systems: Three Easy Pieces

Flash Translation Layer

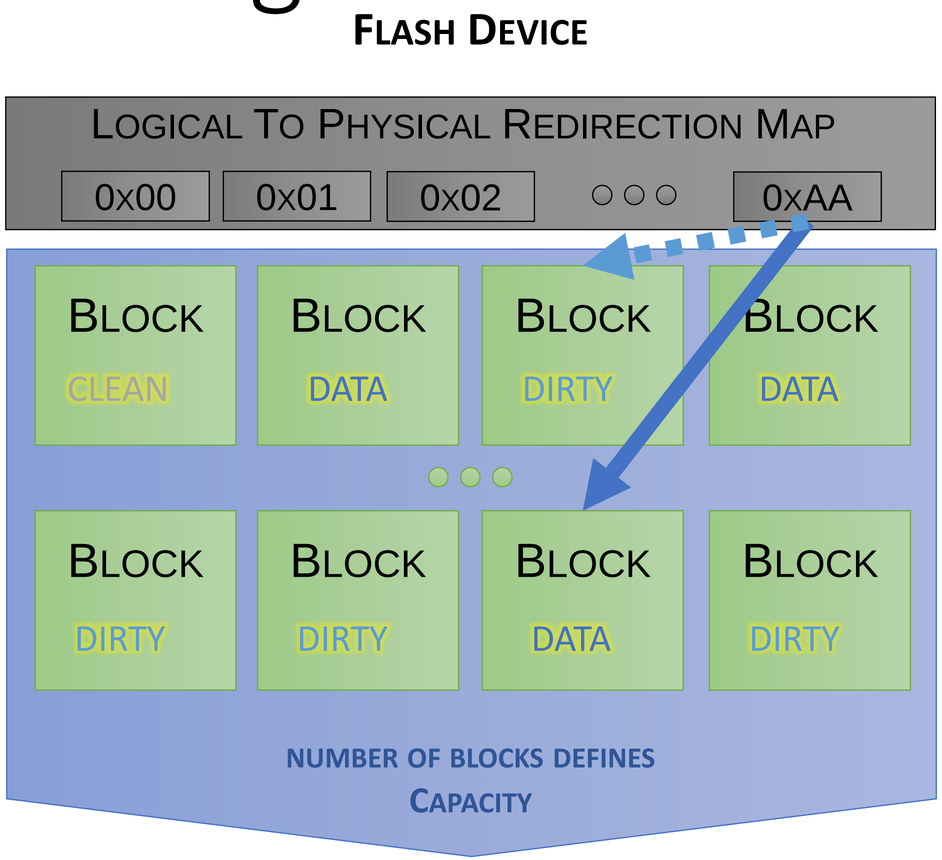
- Flash translation layer (FTL)
  - Provides abstraction of flash memory characteristics
  - Maintains logical to physical address mapping
  - Carries out cleaning operations
  - Conducts wear leveling

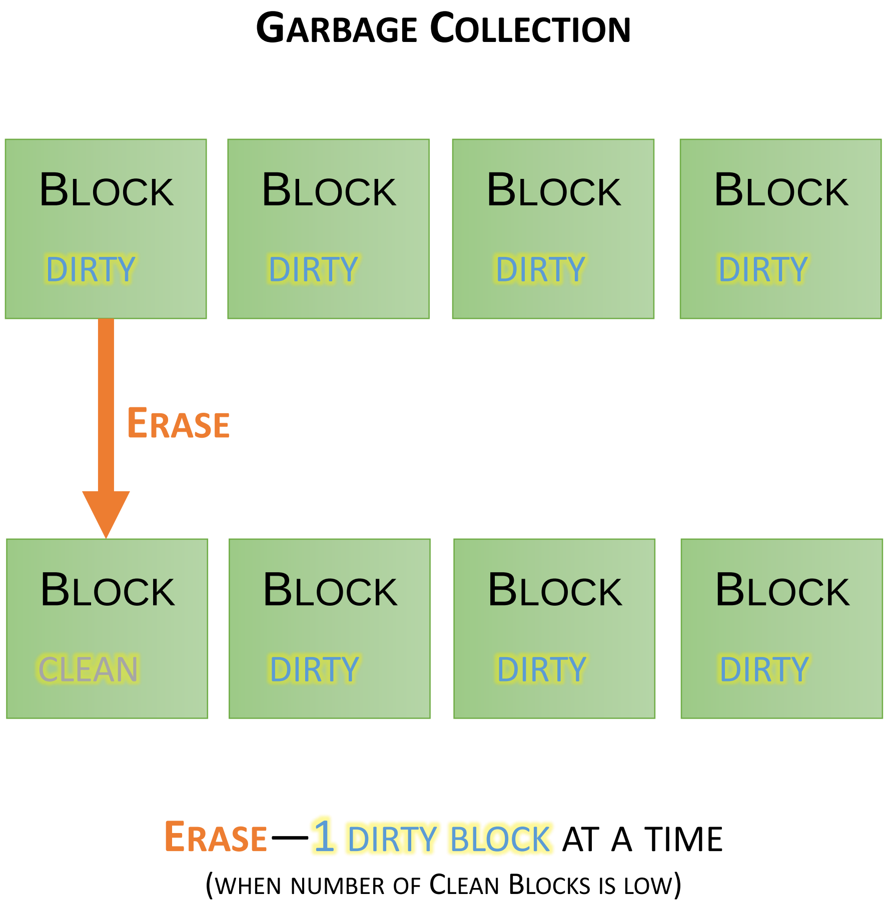
- FTL in multiple flash chip environment
  - Manages parallelism and wear level among chips



Slide coutresy : FTL Design Exploration in Reconfigurable High-Performance SSD (RHPSSD) for Server Applications, ICS 2009

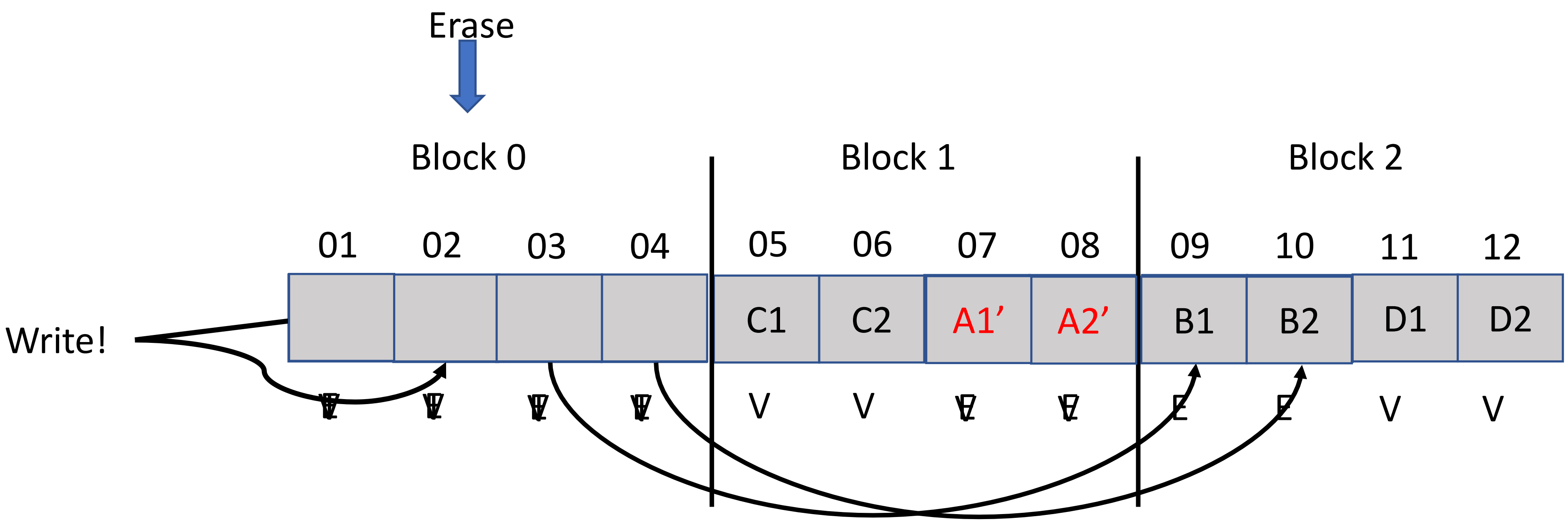
Garbage Collection



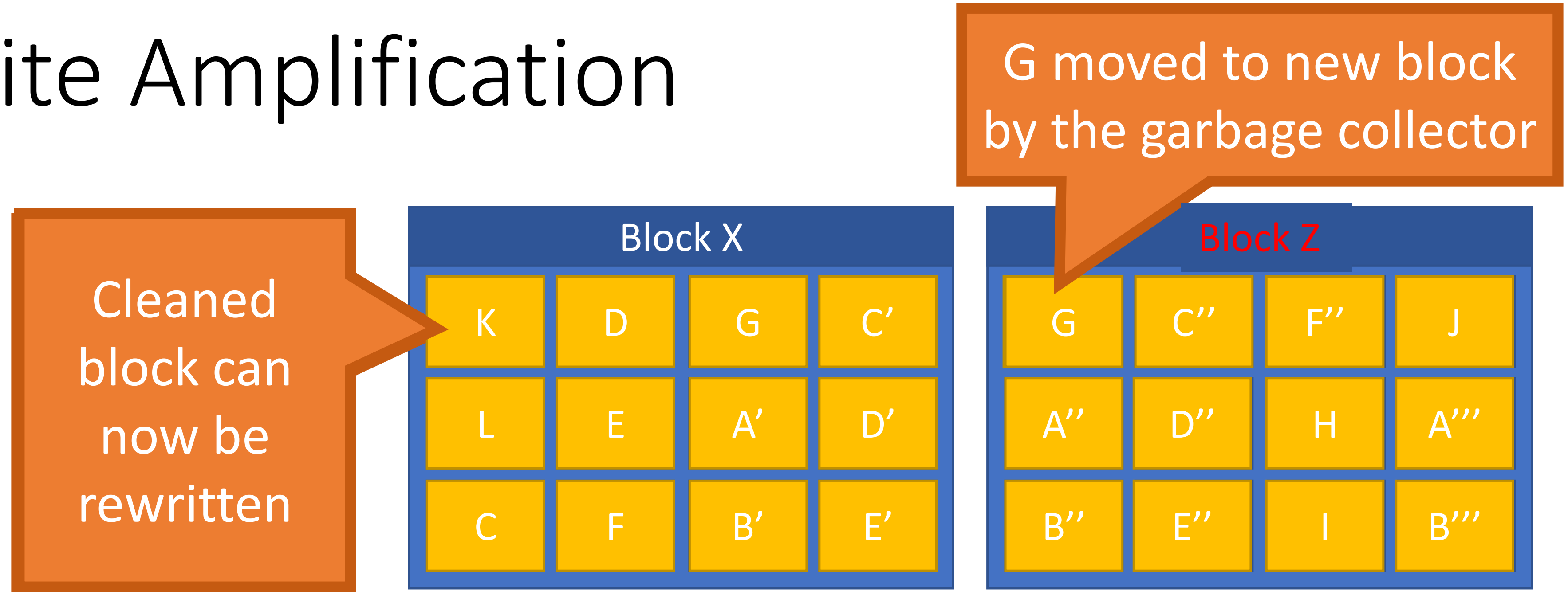


Slide courtesy https://www.snia.org/sites/default/files/ESF/Performance_Benchmarking_3_Block_Components_FinalPPT.pptx

Garbage Collection Example



Write Amplification



- Once all pages have been written, valid pages must be consolidated to free up space
- Write amplification: a write triggers garbage collection/compaction
  - One or more blocks must be read, erased, and rewritten before the write can proceed

Slide courtesy : Christo Wilson https://cbw.sh/5600/index.html

SSD Controllers

- SSDs are extremely complicated internally
- All operations handled by the SSD controller
  - Maps LBAs to physical pages
  - Keeps track of free pages, controls the GC
  - May implement background GC
  - Performs wear leveling via data rotation
- Controller performance is crucial for overall SSD performance

Slide courtesy : Christo Wilson https://cbw.sh/5600/index.html

Flavors of NAND Flash Memory

Multi-Level Cell (MLC)

- Multiple bits per flash cell
  - For two-level: 00, 01, 10, 11
  - 2, 3, and 4-bit MLC is available
- Higher capacity and cheaper than SLC flash
- Lower throughput due to the need for error correction
- 3000 – 5000 write cycles
- Consumes more power

Consumer-grade drives

Single-Level Cell (SLC)

- One bit per flash cell
  - 0 or 1
- Lower capacity and more expensive than MLC flash
- Higher throughput than MLC
- 10000 – 100000 write cycles

Expensive, enterprise drives

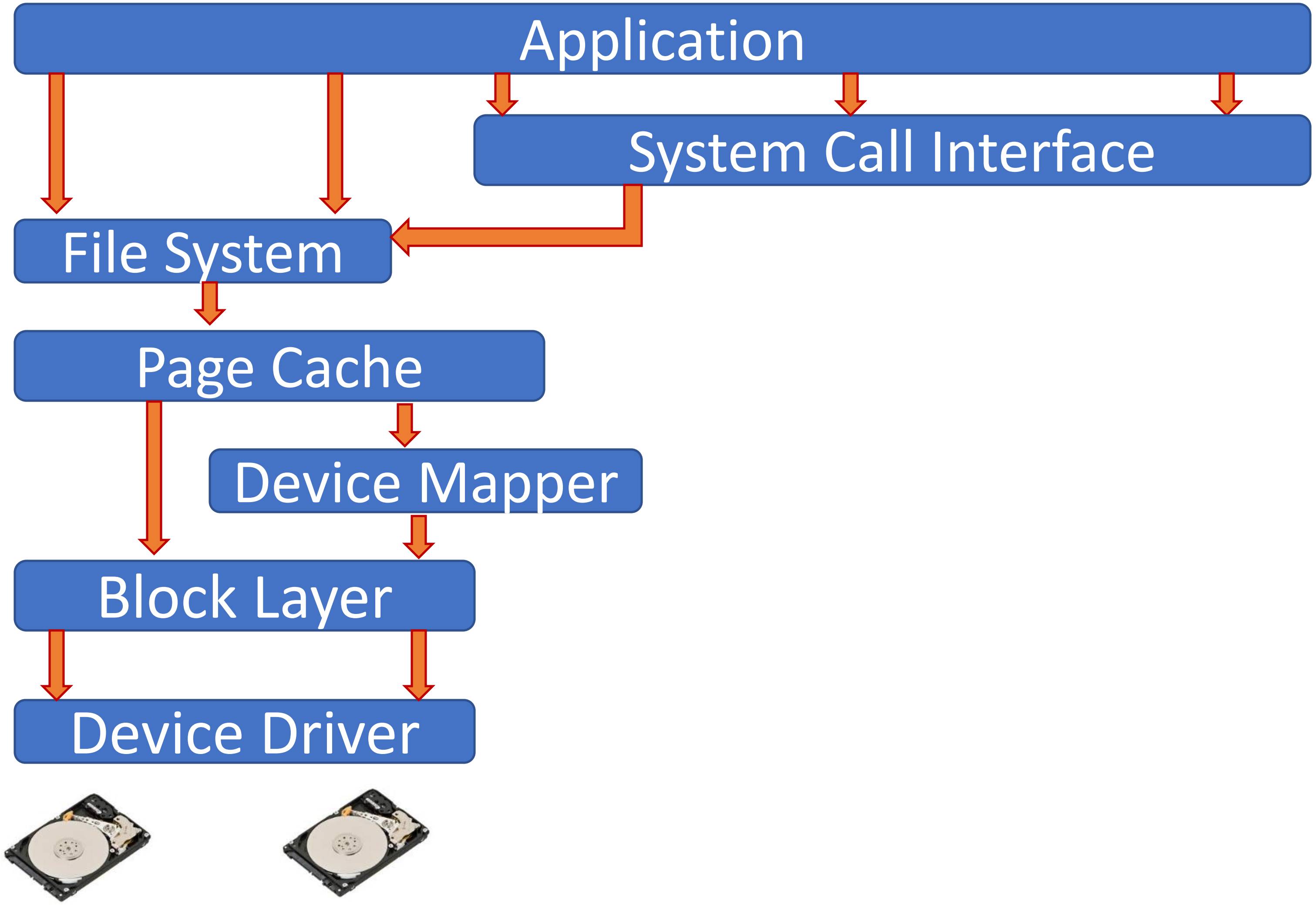
Slide courtesy : Christo Wilson https://cbw.sh/5600/index.html

Other Concerns

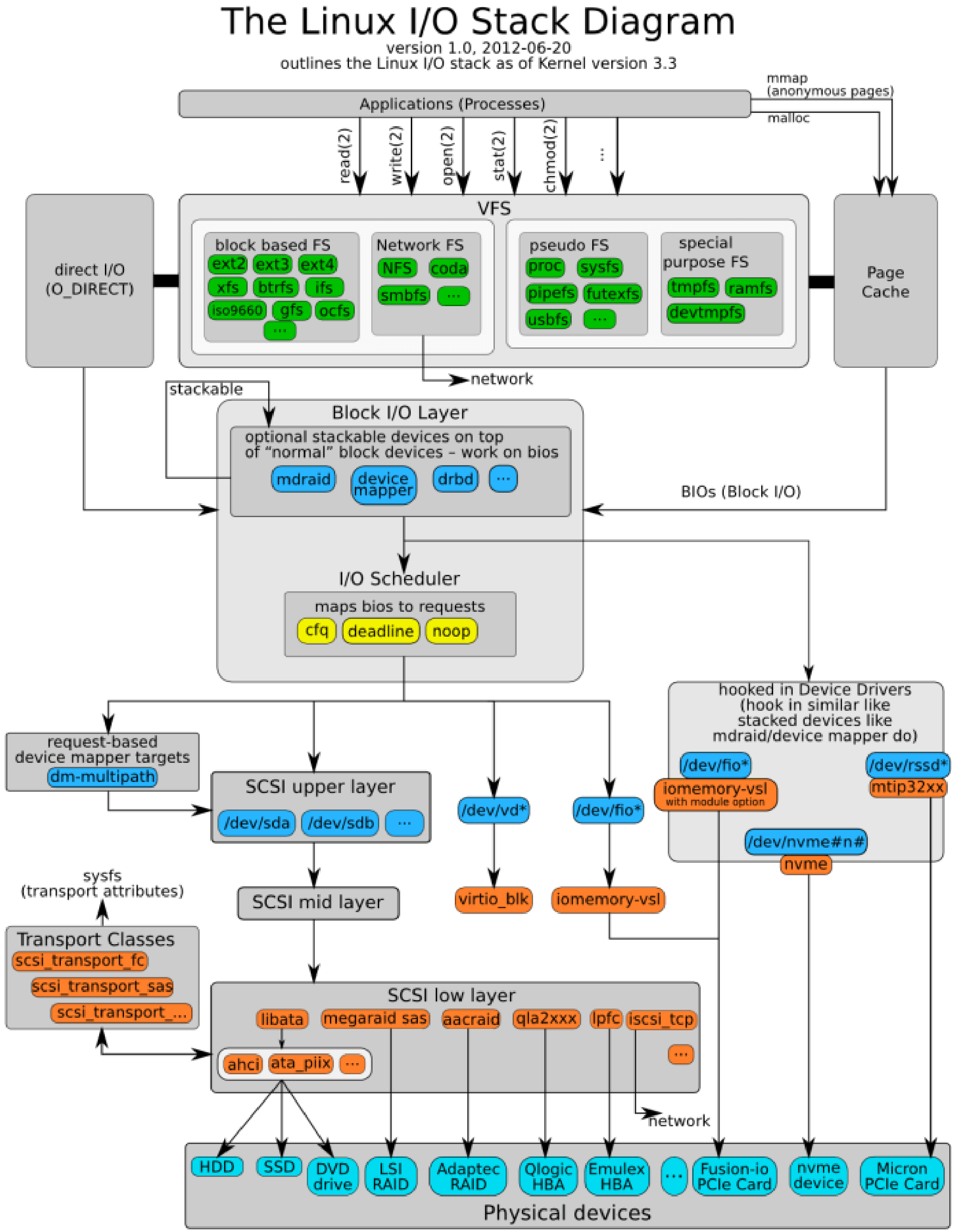
- Wear Leveling
  - Each cell has limited number of P/E cycles
  - Keep track of how many times physical pages have been erased/written
  - Try to make sure all pages have similar numbers

Device	Random Reads (MB/s)	Random Writes (MB/s)	Sequential Reads (MB/s)	Sequential Writes (MB/s)
Samsung 840 Pro SSD	103	287	421	384
Seagate 600 SSD	84	252	424	374
Intel SSD 335 SSD	39	222	344	354
Seagate Savvio 15K.3 HDD	2	2	223	223

The Simplified I/O Software Stack

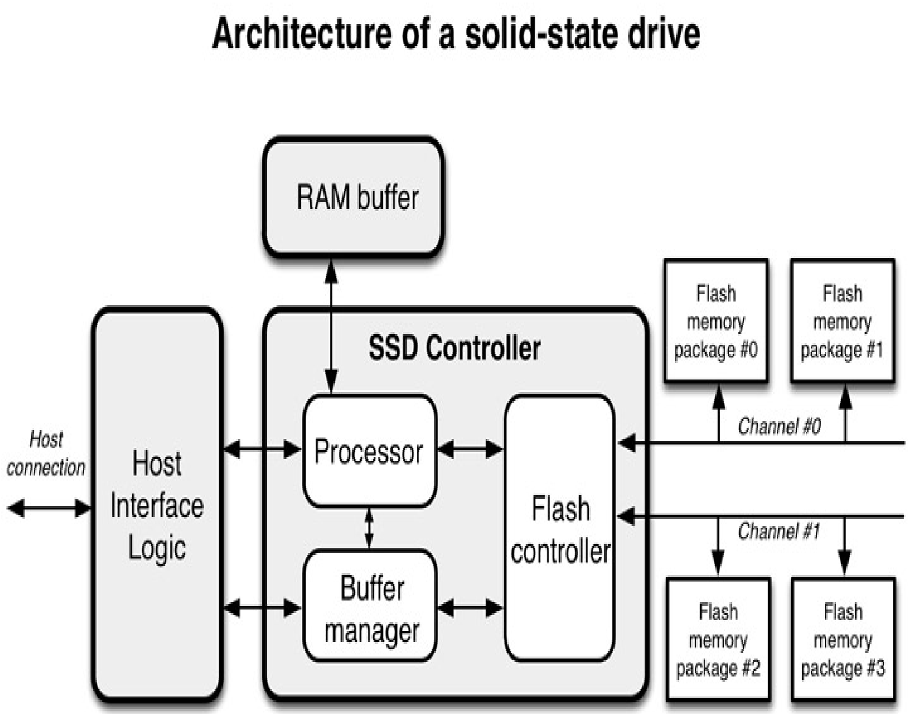


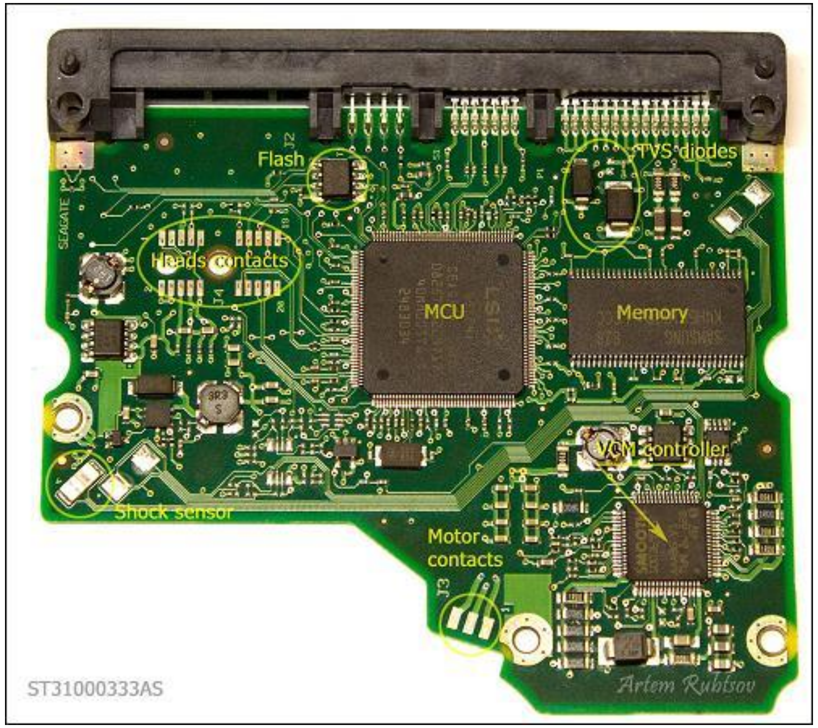
Linux I/O Stack



DRAM inside HDDs/SSDs

- DRAM can cache I/O data within an SSD
- What should you cache?

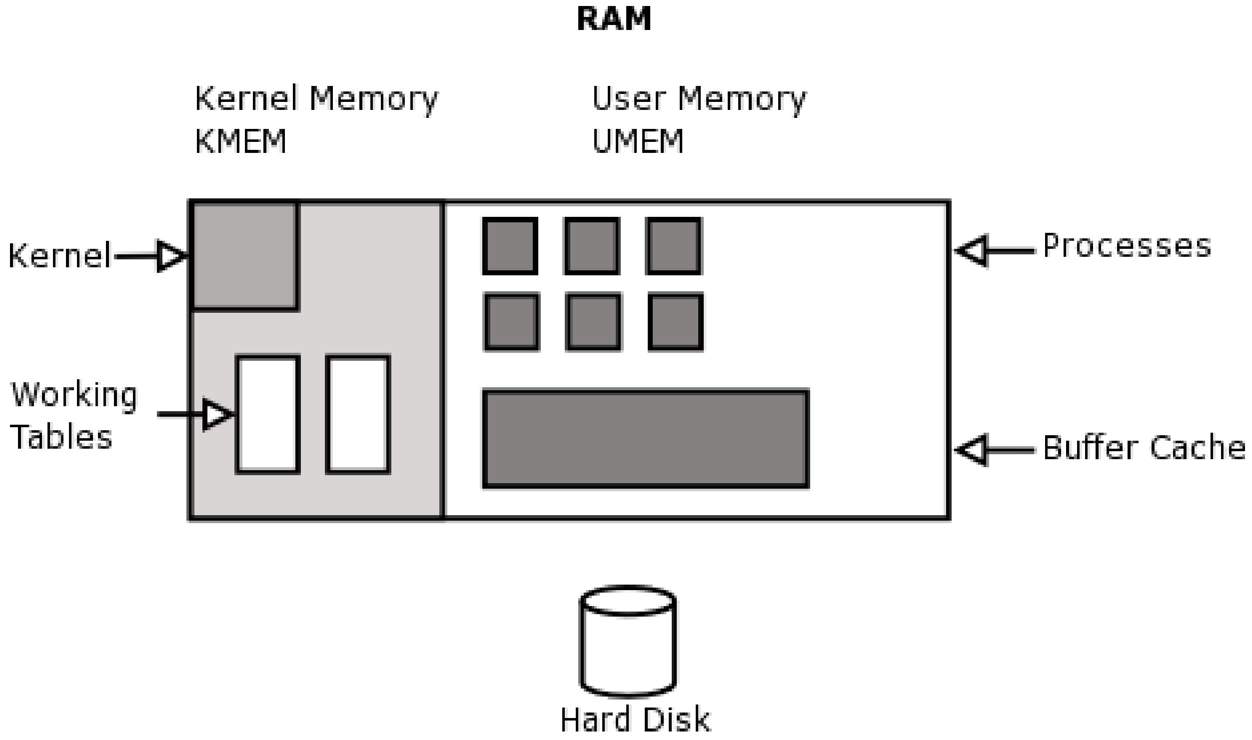




http://hddscan.com/doc/HDD_from_inside.html

Bonus Slide # 2 – Caching I/O Data in DRAM

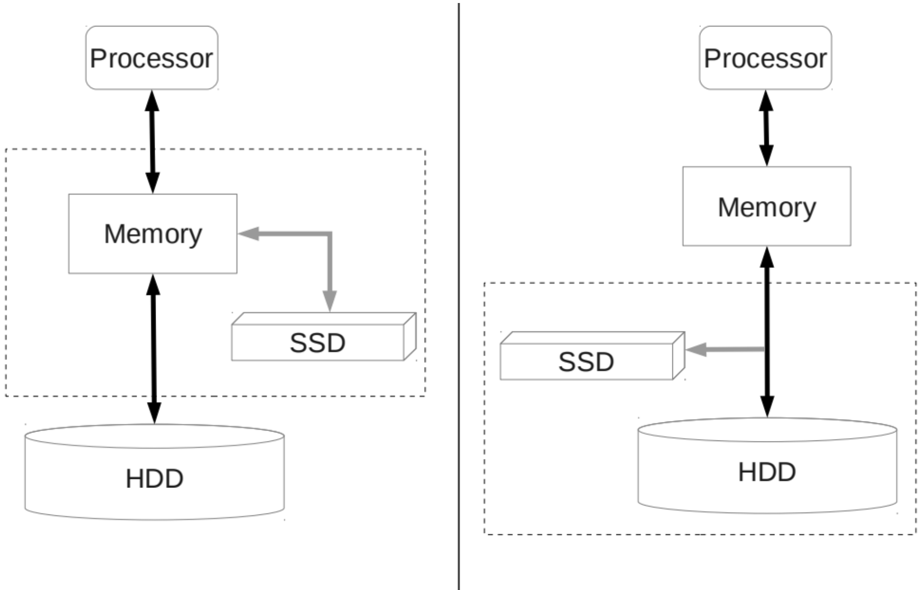
- Main memory is used as a I/O cache, among other things
- What data should be cached?
- "Who" should cache it?
  - The application?
  - The OS?



http://www.learnlinux.org.za/courses/build/internals/ch05.html

Hybrid SSD + HDD Architectures

- HDDs are cheap; have larger capacities
- SSDs are fast, are much more expensive
- Is there a mechanism to get the best of both worlds?



(a) SSD as extended memory
(b) SSD as extended disk

Improving Flash-based Disk Cache with Lazy Adaptive Replacement, Huang et al., *ACM Transactions on Storage (TOS), 2016*

Thanks!