# CS310 Operating Systems

## Lecture 40 : File System Design – 1
## Storage Devices and FAT

Ravi Mittal

IIT Goa

# References

- CS162, Operating Systems and Systems Programming, University of California, Berkeley
- Various websites on the Internet

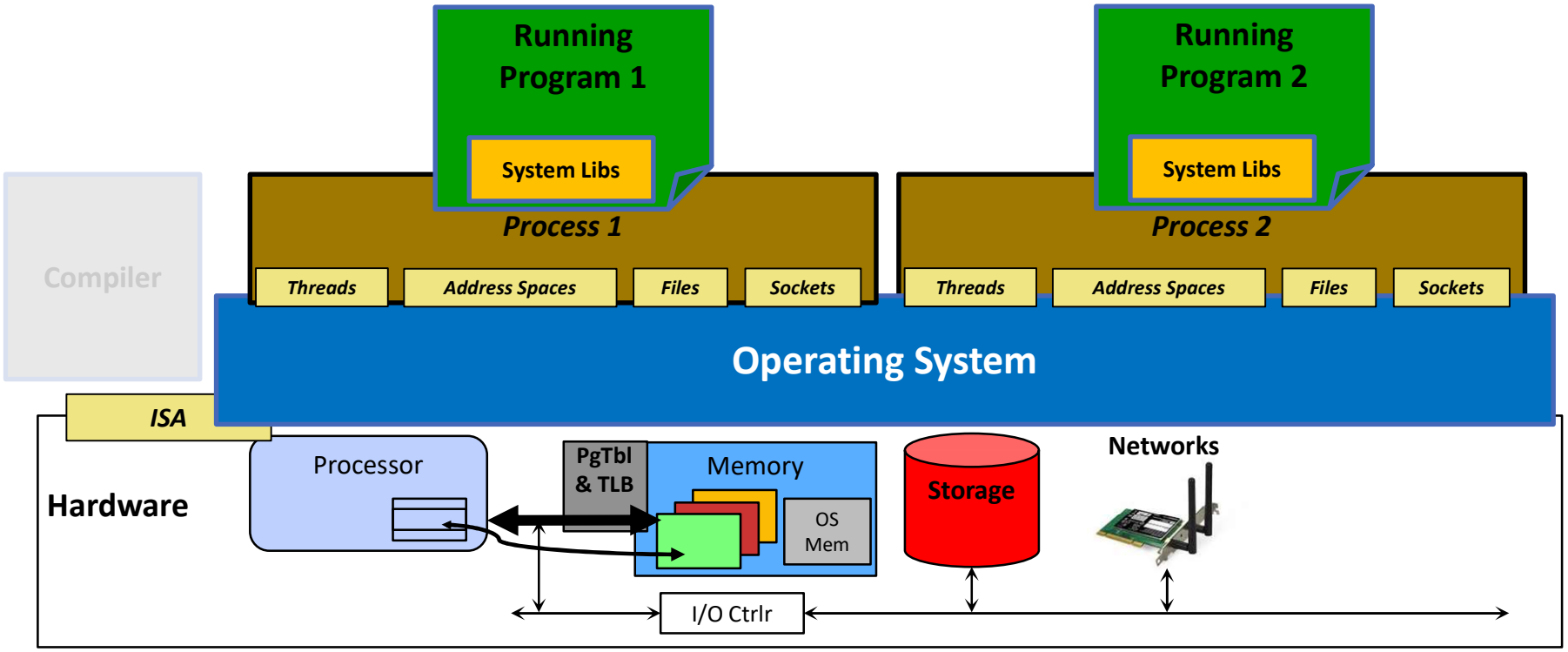
# Reading

- CS162, Operating Systems and Systems Programming, University of California, Berkeley
- Book: Operating System Concepts, 10th Edition, by Silberschatz, Galvin, and Gagne

# Lecture Contents

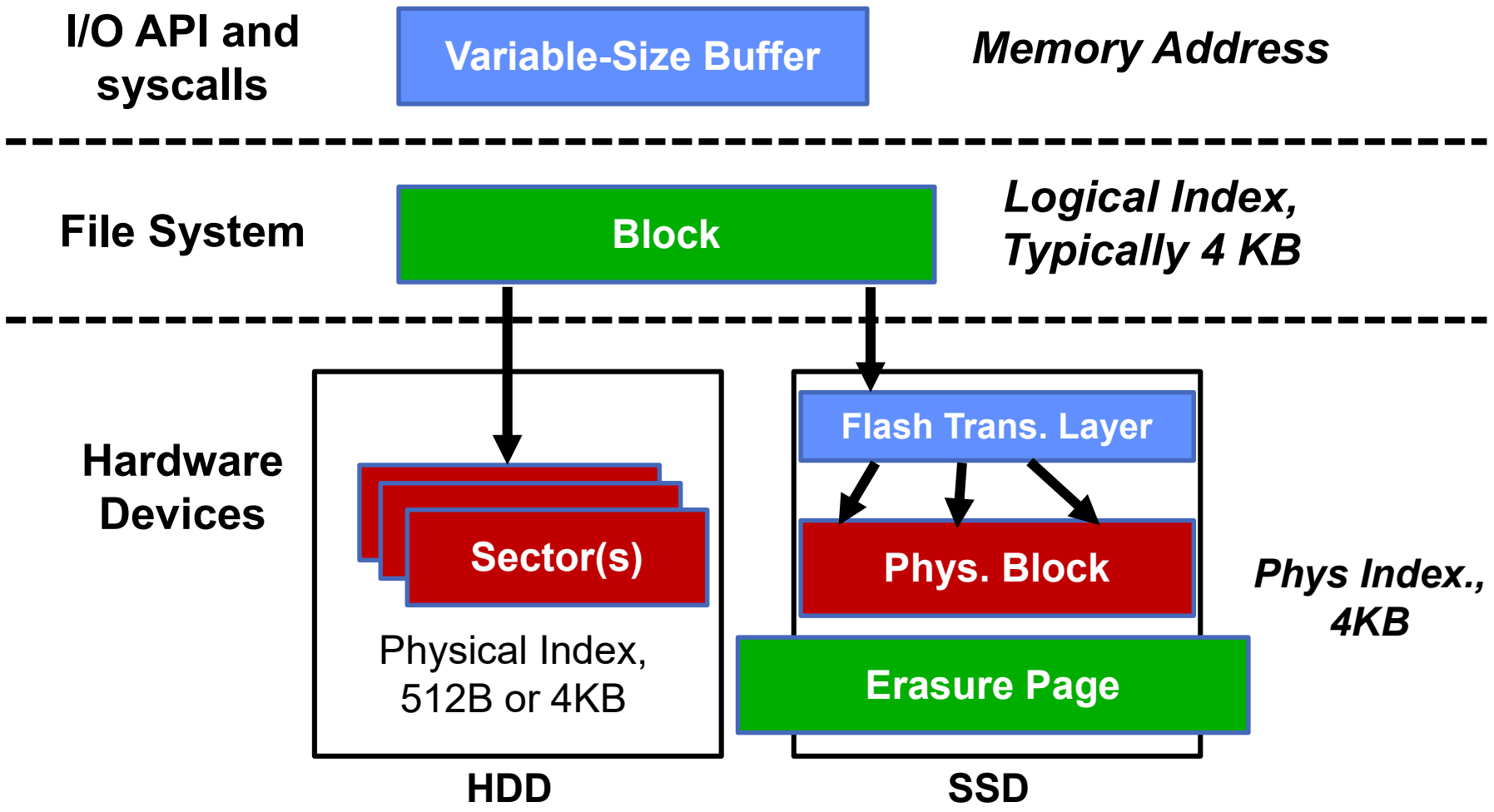
- Persistent Storage Technologies
- Hard Disk Drives
- Solid State Drive (SDD)

# Where are we?


Running Program 1
System Libs
Process 1
Threads
Address Spaces
Files
Sockets
Running Program 2
System Libs
Process 2
Threads
Address Spaces
Files
Sockets
Compiler
Operating System
ISA
Hardware
Processor
PgTbl & TLB
Memory
OS Mem
Storage
Networks
I/O Ctrlr

# Persistent Storage Technologies

# Persistent Storage for File System

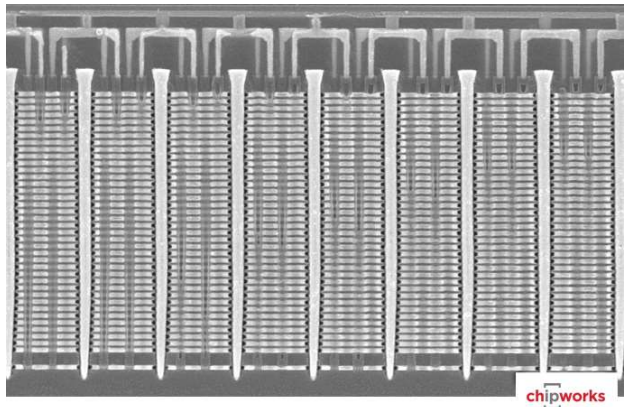


# Storage Technologies

## Magnetic Disks

- Store on magnetic medium
- Electromechanical access

## Nonvolatile (Flash) Memory




Close-up image of V-NAND flash array

- Store as persistent charge
- Implemented with 3-D structure
  - 100+ levels of cells
  - 3 bits data per cell

# RAM vs Hard Disk vs SSD - 2018

| | RAM | HDD | SSD |
|---|---|---|---|
| **Typical Size** | 8 GB | 1 TB | 256 GB |
| **Cost** | $10 per GB | $0.05 per GB | $0.32 per GB |
| **Power** | 3 W | 2.5 W | 1.5 W |
| **Read Latency** | 15 ns | 15 ms | 30 µs |
| **Read Speed (Seq.)** | 8000 MB/s | 175 MB/s | 550 MB/s |
| **Read/Write Granularity** | word | sector | page* |
| **Power Reliance** | volatile | non-volatile | non-volatile |

**In SSD Each cell has limited program/erase lifetime (thousands, for modern devices) – Cells become slowly less reliable**

# Popular Storage Devices

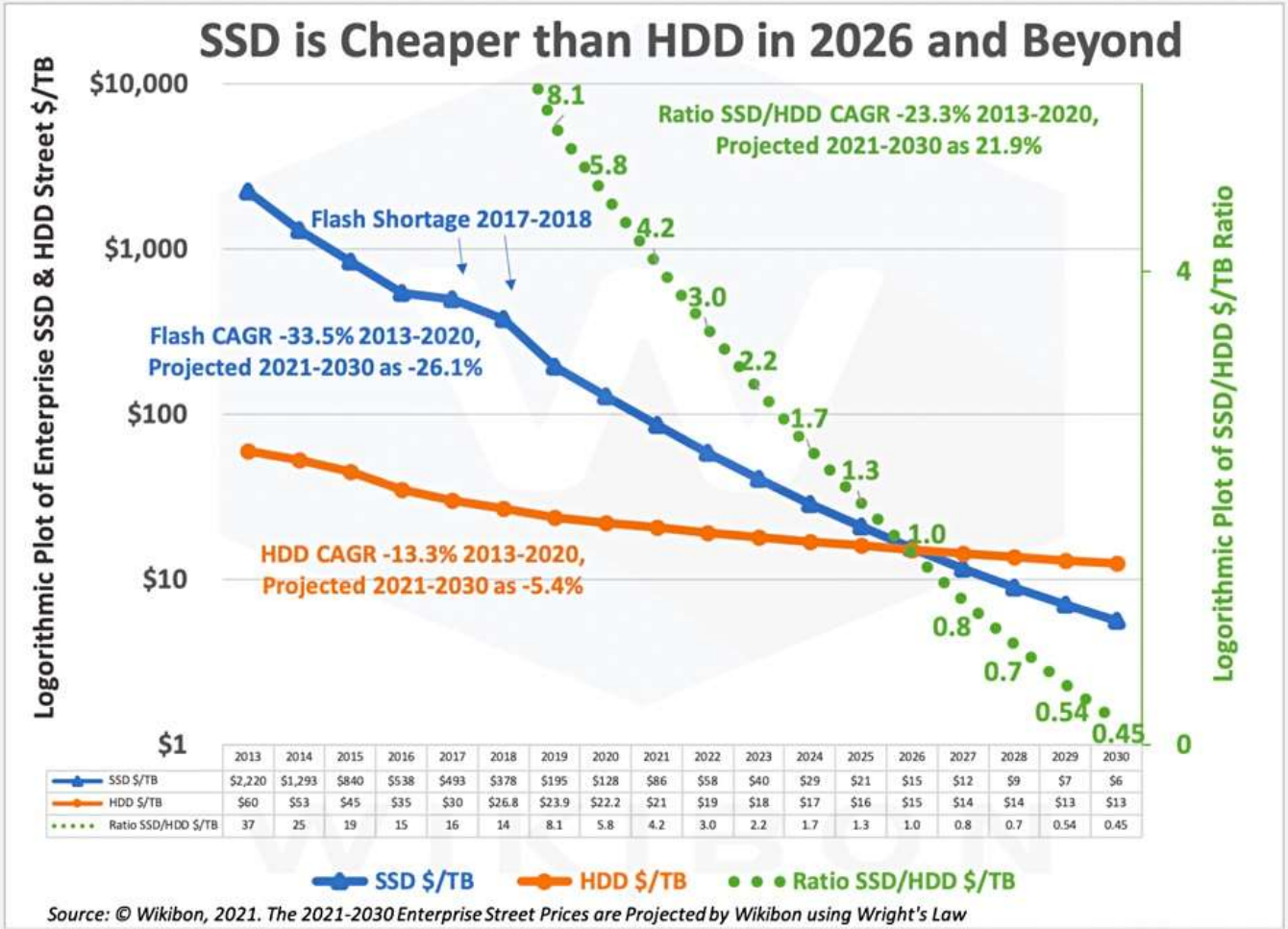
**Magnetic Disks**

- Rarely becomes corrupted
- Traditionally: large capacity at low cost
- Block level random access
- Slow performance for random access
- Better performance for sequential access

**Flash Memory**

- Rarely becomes corrupted
- Increasingly larger and cheaper
- Block level random access
- Good performance for reads, worse for random writes
- Have to erase data in large blocks
- Challenge: Wear Levelling

# Emergence of SSDs




| | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SSD $/TB | $2,220 | $1,293 | $840 | $538 | $493 | $378 | $195 | $128 | $86 | $58 | $40 | $29 | $21 | $15 | $12 | $9 | $7 | $6 |
| HDD $/TB | $60 | $53 | $45 | $35 | $30 | $26.8 | $23.9 | $22.2 | $21 | $19 | $18 | $17 | $16 | $15 | $14 | $14 | $13 | $13 |
| Ratio SSD/HDD $/TB | 37 | 25 | 19 | 15 | 16 | 14 | 8.1 | 5.8 | 4.2 | 3.0 | 2.2 | 1.7 | 1.3 | 1.0 | 0.8 | 0.7 | 0.54 | 0.45 |

**Figure 4 - SSD/HDD Pricing Ratio 2013 - 2030**

*Source: © Wikibon, 2021.*

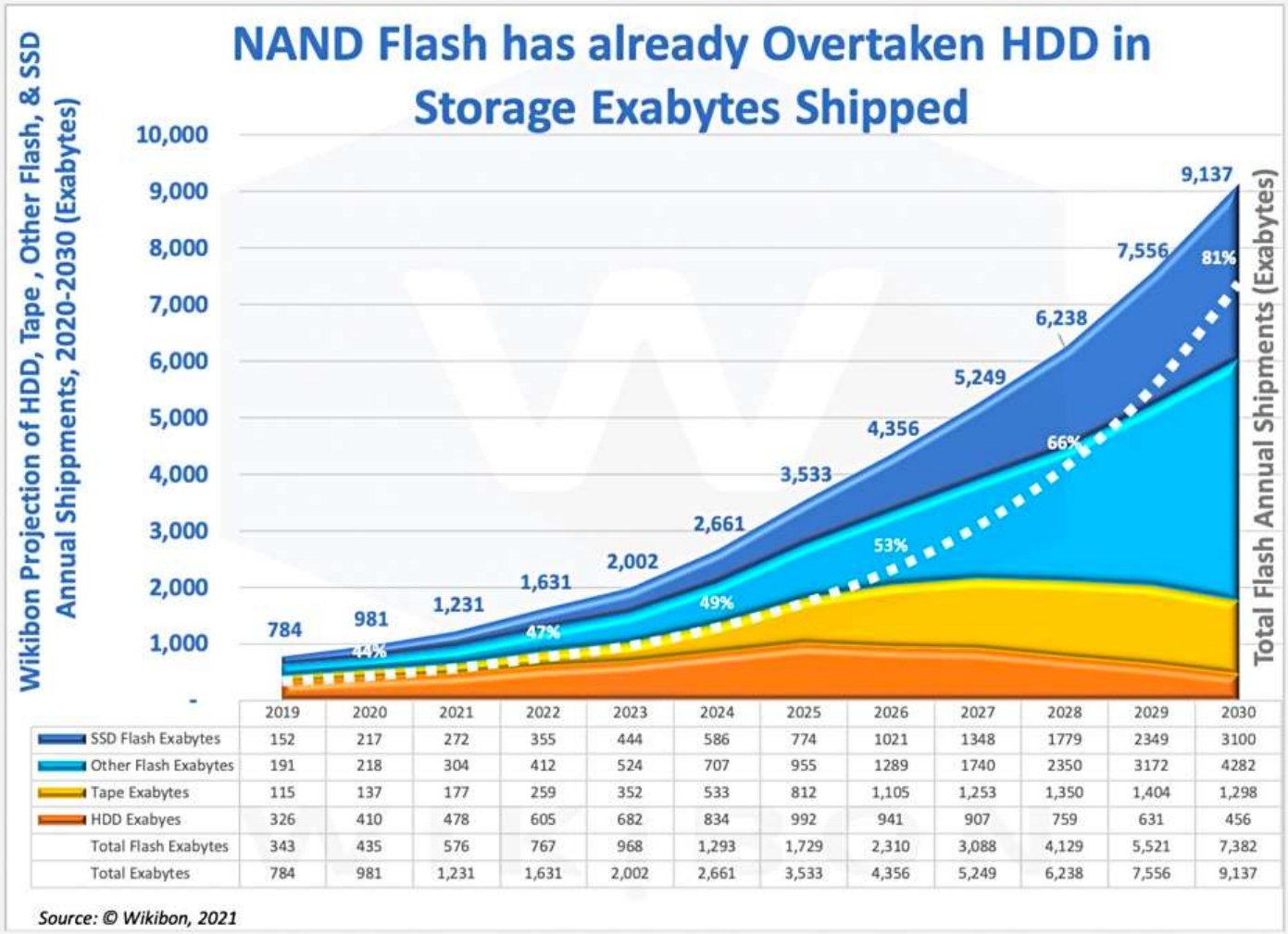


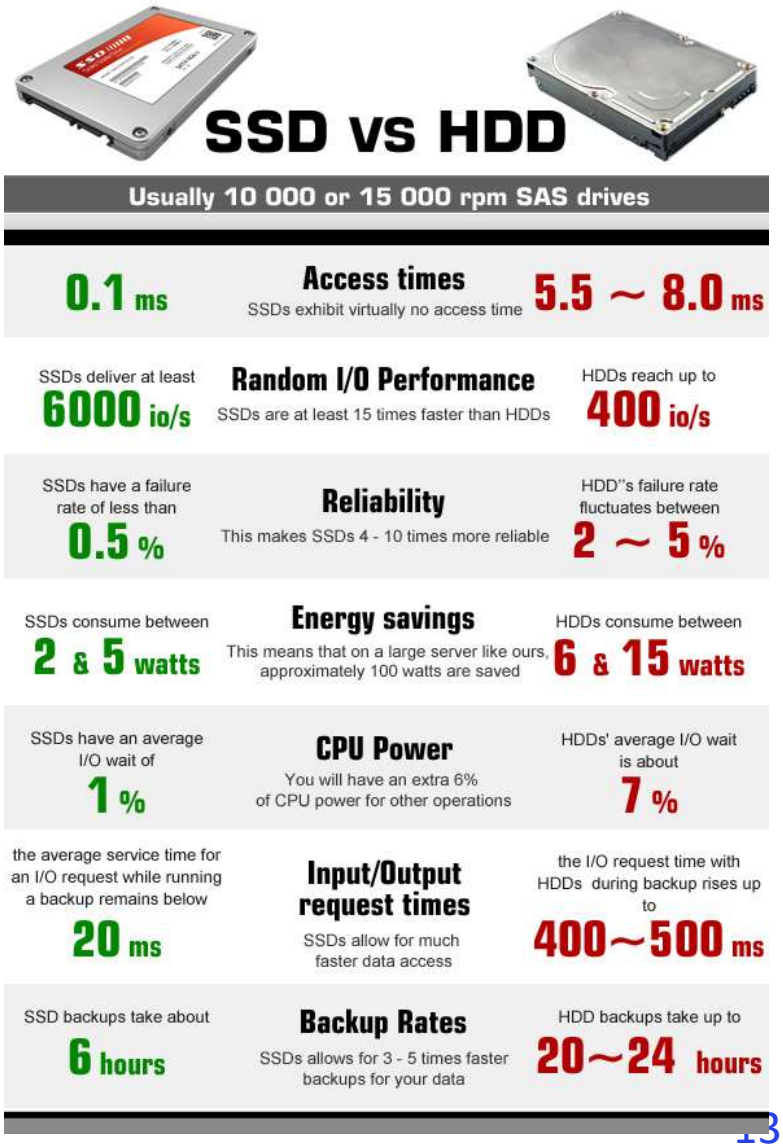

**Figure 9 – Exabyte Storage Shipments Split by SSD, Other Flash, HDDs, and Tape**

*Source: © Wikibon, 2021.*

# The Emergence of SSDs

- Faster
- Lower power
- No moving parts

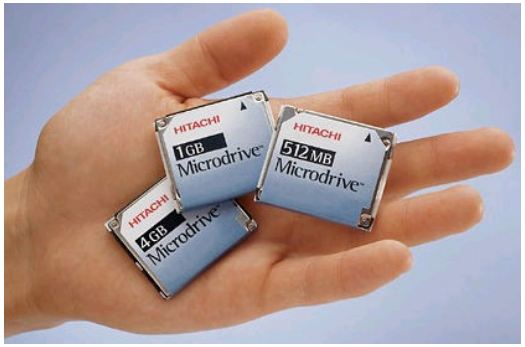
- But HDDs have their place
  - Cheapest online storage per byte
  - Application: Archival storage



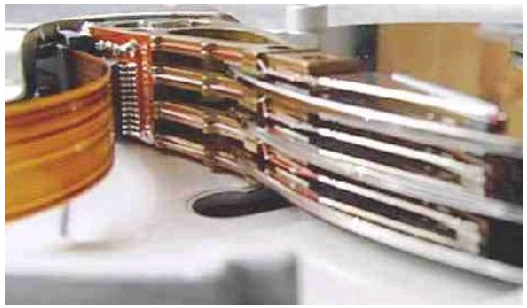
# Hard Disk Drives (HDD)

# Hard Disk Drivers (HDDs)


Western Digital Drive
http://www.storagereview.com/guide/




**IBM/Hitachi Microdrive**

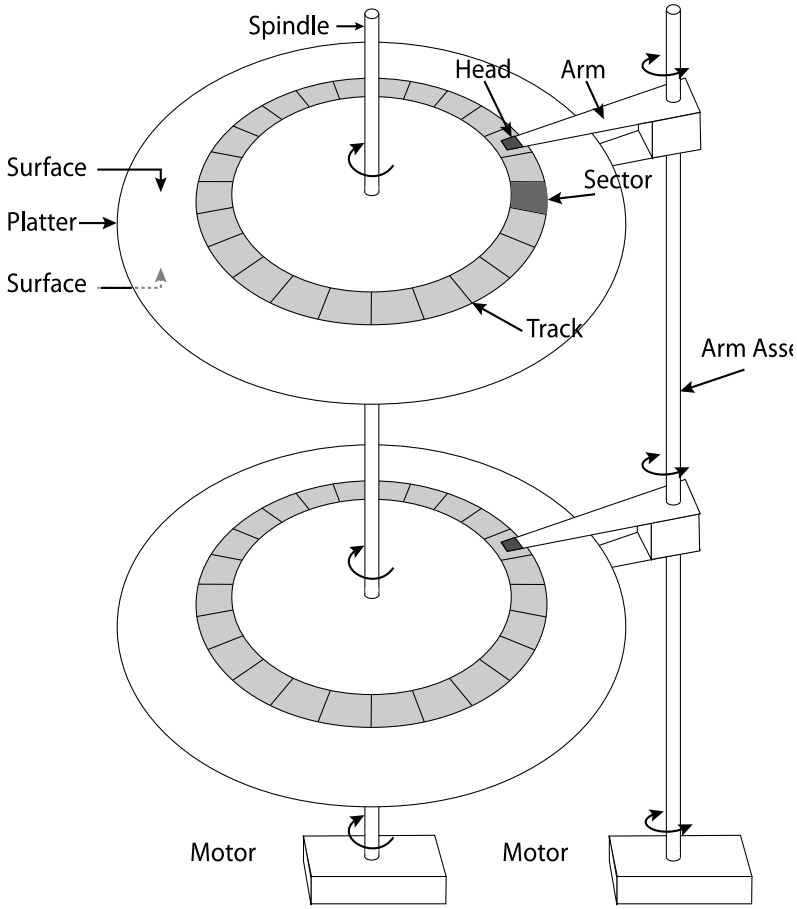


**Read/Write Head Side View**

IBM Personal Computer/AT (1986)
- 30 MB hard disk - $500
- 30-40ms seek time
- 0.7-1 MB/s (est.)

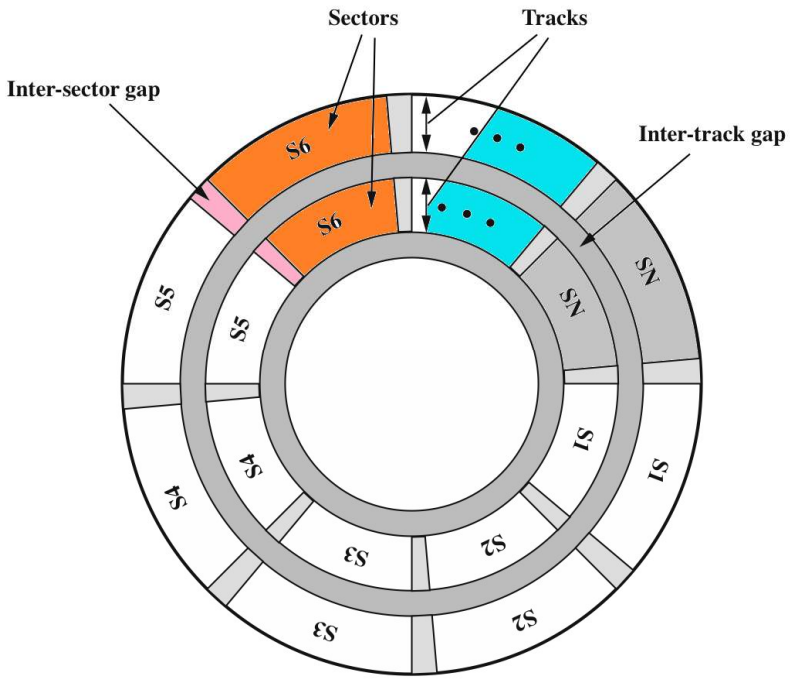
# The Amazing Magnetic Disk

- Unit of Transfer: Sector
  - Ring of sectors form a track
  - Stack of tracks form a cylinder
  - Heads position on cylinders
- Disk Tracks ~ 1µm (micron) wide
  - Wavelength of light is ~ 0.5µm
  - Resolution of human eye: 50µm
  - 100K tracks on a typical 2.5” disk
- Separated by unused guard regions
  - Reduces likelihood neighboring tracks are corrupted during writes (still a small non-zero chance)



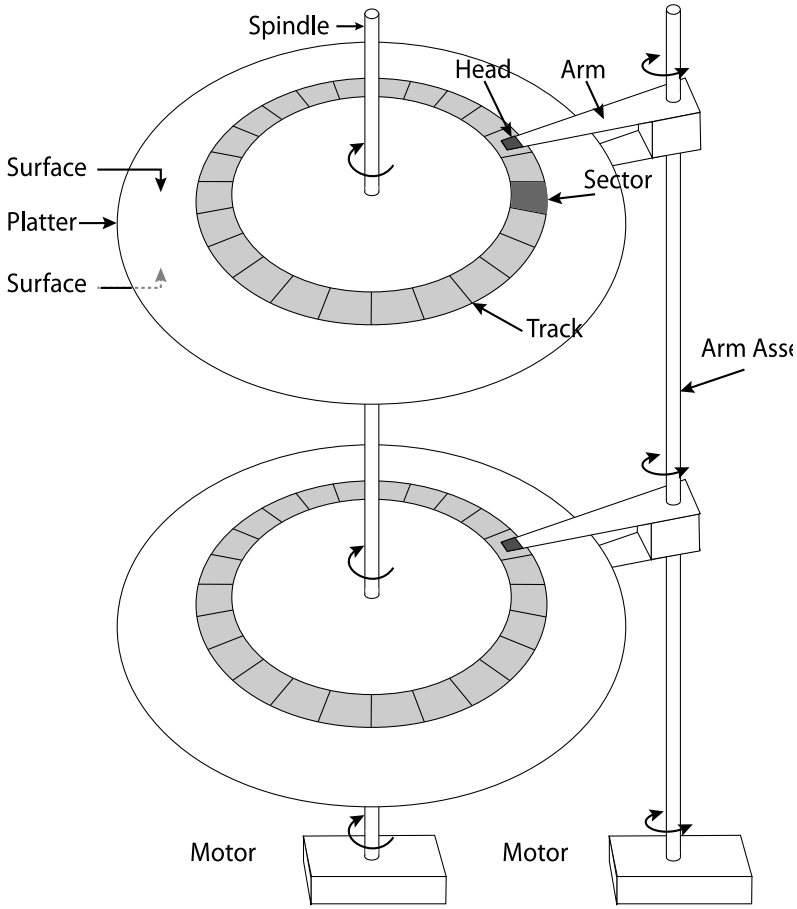
# Disk platter – top view

- Typically sector = 512 bytes
- Minimum reading on disk: 1 sector
  - Can’t read individual byte or word



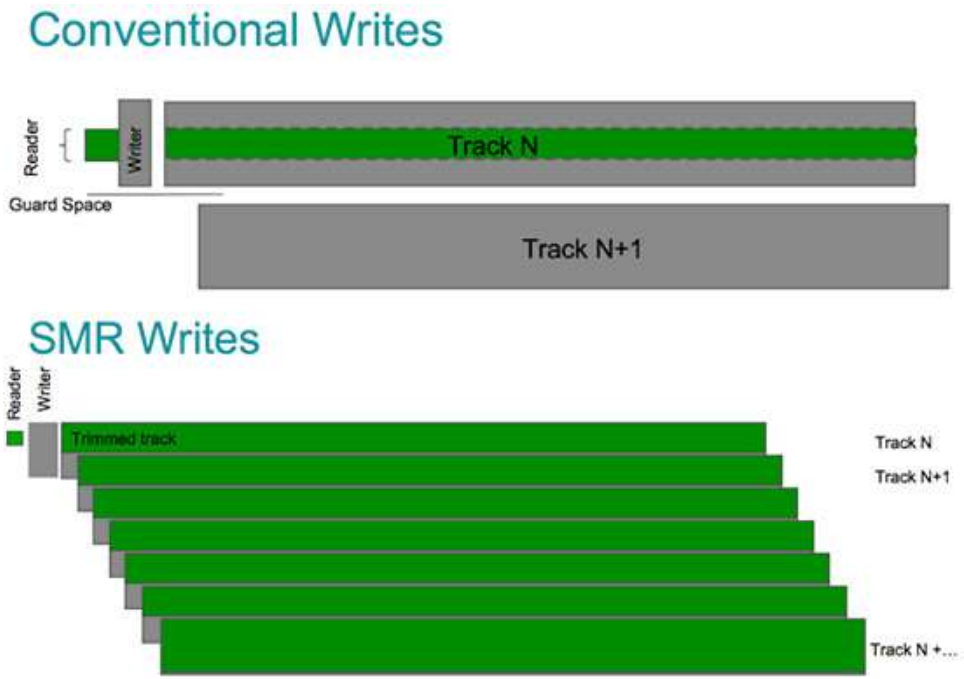
# The Amazing Magnetic Disk

- Track length varies across disk
  - Outside: More sectors per track, higher bandwidth
  - Disk is organized into regions of tracks with same # of sectors/track
  - Only outer half of radius is used
    - Most of the disk area in the outer regions of the disk



# Shingled Magnetic Recording (SMR)

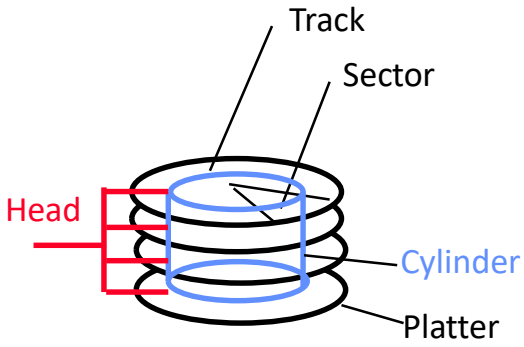
- Shingled magnetic recording is a magnetic storage data recording technology used in hard disk drives to increase storage density and overall per-drive storage capacity



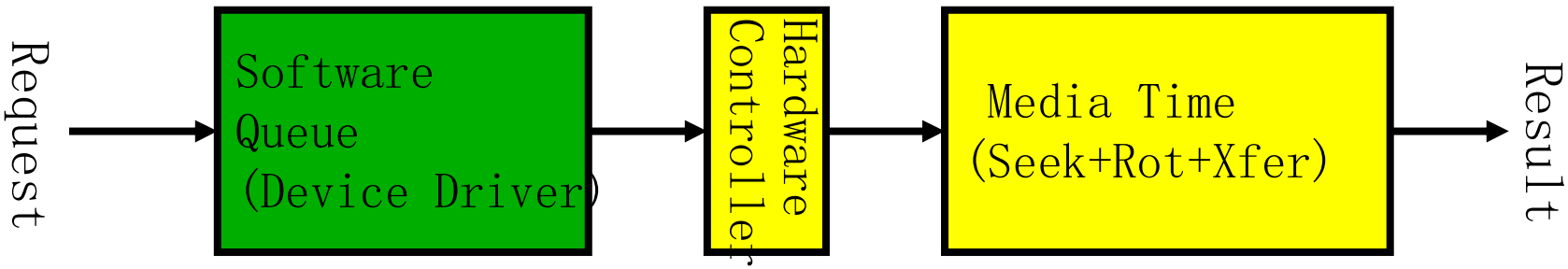

- Overlapping tracks yields greater density, capacity
- Restrictions on writing, complex DSP for reading

# Review: Magnetic Disks

- Cylinders: all the tracks under the head at a given point on all surfaces
- Read/write data is a three-stage process:
  - Seek time: position the head/arm over the proper track –Average of 5-10 ms
  - Rotational latency: wait for desired sector to rotate under r/w head
    - 4-8 ms (3600-7200 rpm), 2-4 ms (15000ms)
  - Transfer time: Time to actually read sectors
    - 50-100 MB/sec



# Bigger Picture




Latency = Queue Time + Controller Time +
Seek Time + Rotational Latency +
Transfer Time

To Achieve Best Bandwidth: Large Transfers of Physically Adjacent Sectors from one track

# Disk Performance Example – Self reading

- Assumptions:
  - Ignoring queuing and controller times for now
  - Avg seek time of 5ms,
  - 7200RPM ⇒ Time for rotation: 60000 (ms/min) / 7200(rev/min) ~= 8ms
  - Transfer rate of 50MByte/s, block size of 4Kbyte ⇒ 4096 bytes/$50\times10^6$ (bytes/s) = $81.92 \times 10^{-6}$ sec ≅ 0.082 ms per block
- Read block from random place on disk:
  - Seek (5ms) + Rot. Delay (4ms) + Transfer (0.082ms) = 9.082ms
  - Approx 9ms to fetch/put data: 4096 bytes/$9.082\times10^{-3}$ s ≅ 451KB/s
- Read block from random place in same cylinder:
  - Rot. Delay (4ms) + Transfer (0.082ms) = 4.082ms
  - Approx 4ms to fetch/put data: 4096 bytes/$4.082\times10^{-3}$ s ≅ 1.03MB/s
- Read next block on same track:
  - Transfer (0.082ms): 4096 bytes/$0.082\times10^{-3}$ s ≅ 50MB/sec
- Key to using disk effectively (especially for file systems) is to minimize seek and rotational delays

# HDD Controller

# HDD Controllers

- Old Days: Device driver would address block of data by cylinder number, head (platter) number, and sector number
- Now: Hard drive is just an array of sectors
  - Sector number mapped internally to physical location
  - Numerically close sectors are probably physically close
- Lots of other intelligent features
  - Error Correcting Codes
  - Sector Sparing
  - Slip Sparing
  - Track Skewing

# Intelligence in the Controller

- Error Correcting Codes
  - Disk encodes each sector with additional error correcting code data – allowing it to fix imperfectly read or written data
- Sector sparing
  - Disks are made with tracks and sectors as thin and small as possible
  - If there are imperfections in a sector, then the sector can not be used for reliably store data
  - Manufacturers include spare sectors distributed across each surface
  - Disk firmware or formatting can remap bad sectors transparently to spare sectors on the same surface
- Slip sparing
  - Remap all sectors (when there is a bad sector) to preserve sequential behavior
  - Helps in retaining good sequential performance by remapping all sectors from the bad sector to the next spare

# Intelligence in the Controller

- Track Skewing
  - Sector numbers offset from one track to the next, to allow for disk head movement for sequential operations
  - Logical sector zero on each track is staggered from sector zero on the previous track
    - By an amount corresponding to time it takes the disk to move the head from one track to another

# Typical Numbers for Magnetic Disk (self reading)

| Parameter | Info/Range |
|---|---|
| Space/Density | Space: 14TB (Seagate), 8 platters, in 3½ inch form factor!<br>**Areal Density: ≥ 1 Terabit/square inch! (PMR, Helium, …)** |
| Average Seek Time | Typically 4-6 milliseconds |
| Average Rotational Latency | Most laptop/desktop disks rotate at 3600-7200 RPM<br>(16-8 ms/rotation). Server disks up to 15,000 RPM.<br>Average latency is halfway around disk so 4-8 milliseconds |
| Controller Time | Depends on controller hardware |
| Transfer Time | Typically 50 to 250 MB/s. Depends on:<br>• Transfer size (usually a sector): 512B – 1KB per sector<br>• Rotation speed: 3600 RPM to 15000 RPM<br>• Recording density: bits per inch on a track<br>• Diameter: ranges from 1 in to 5.25 in |
| Cost | Used to drop by a factor of two every 1.5 years (or faster), now slowing down |

# Example of Current HDDs (Self reading)

- Seagate Exos X14 (2018)
  - 14 TB hard disk
    - 8 platters, 16 heads
    - Helium filled: reduce friction and power
  - 4.16ms average seek time
  - 4096 byte physical sectors
  - 7200 RPMs
  - 6 Gbps SATA /12Gbps SAS interface
    - 261MB/s MAX transfer rate
    - Cache size: 256MB
  - Price: $615 (< $0.05/GB)


- IBM Personal Computer/AT (1986)
  - 30 MB hard disk
  - 30-40ms seek time
  - 0.7-1 MB/s (est.)
  - Price: $500 ($17K/GB, 340,000x more expensive !!)

# Lecture Summary

- Hard Disk Drives – had magical run
- Tremendous improvement in HDD technology
  - Smaller sizes
  - More information
  - Less power
  - Low cost
  - Improved Performance
  - Higher reliability
  - A lot of intelligence
- SSD are now becoming main stream