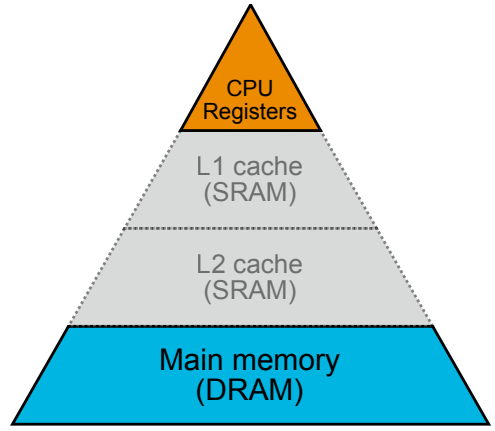
Storage

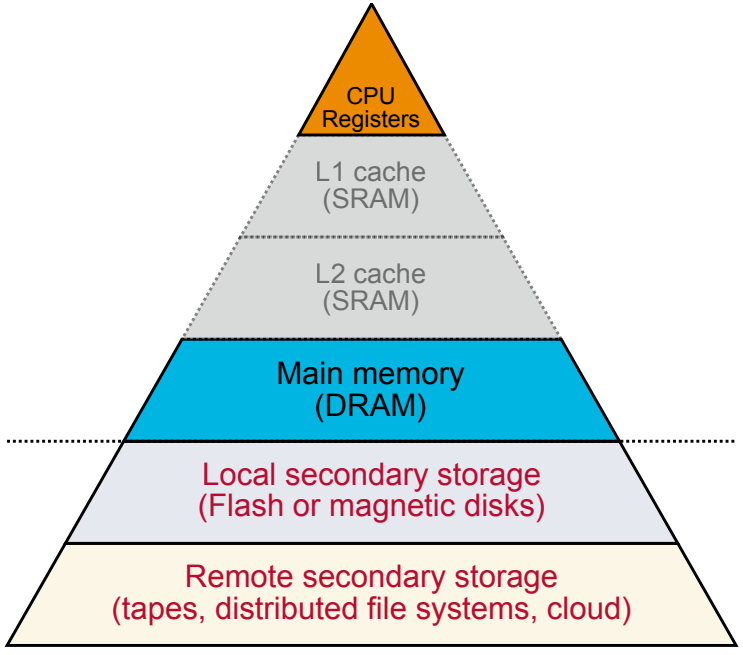
Introduction

Memory issues

- Volatile
- Small
- Expensive

Need for big and cheap persistent storage!





Memory hierarchy

- Size
- Cost
- Speed

- Addressability
- Byte vs block access
- Persistence
- Latency/throughput
- Power drain (in use/idle)
- Weight/volume

Technologies

Volatile memory

SRAM

Static random-access memory

Characteristics

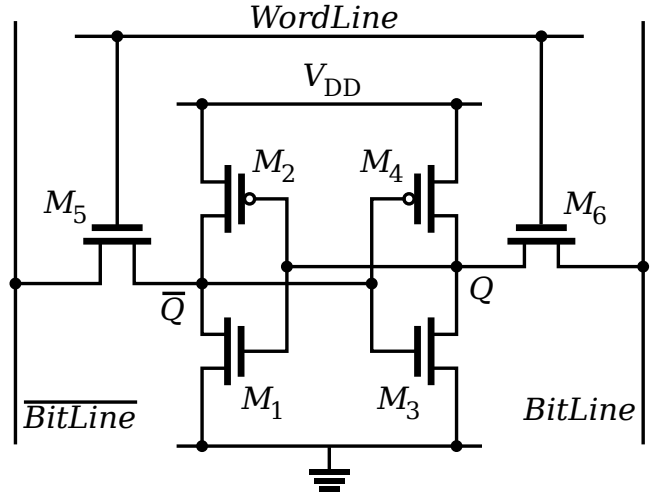
- Bits stored in transistor flip/flops
- Bits degrade on poweroff

Performance

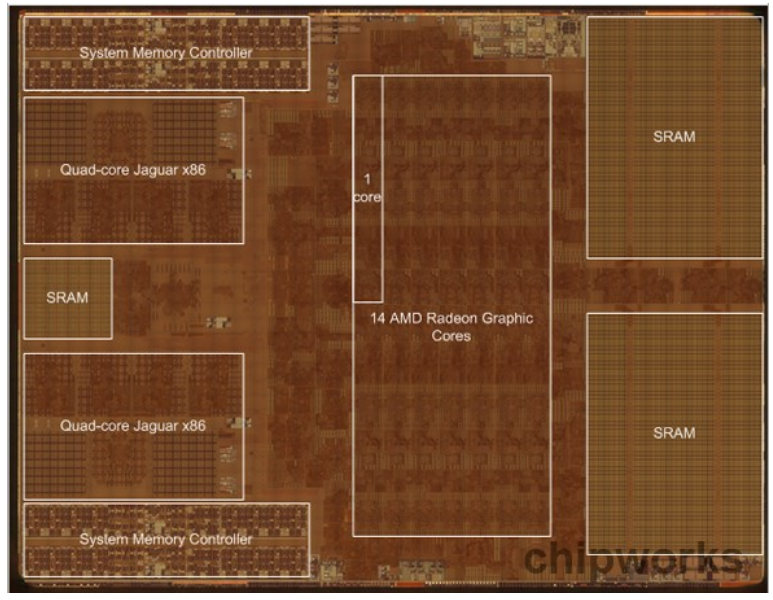
- Access time between 1 - 10 ns

Typical use

- On-chip cache



SRAM cell



Xbox One APU

Technologies

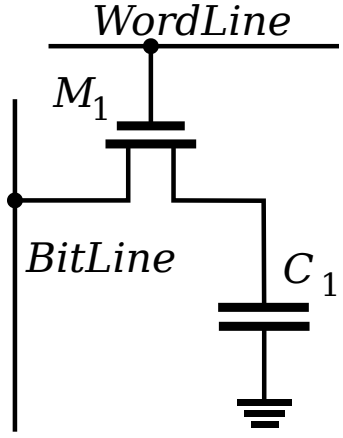
Volatile memory

DRAM

Dynamic random-access memory

Characteristics

- Bits stored in capacitors
- 2D/3D array for dense packing
- Bits degrade even when powered: need to be periodically refreshed



DRAM cell

Performance

- Access time between between 50 - 100 ns
- Transfer bandwidth up to 25GiB/s

Typical use

- Off-chip volatile memory

DRAM module

Technologies

Persistent memory

Magnetic disk

Characteristics

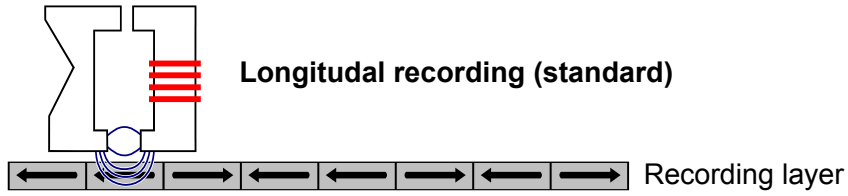
- > 1 Tbit per square inch
- Physical motion needed to read bits off surface
- Not directly addressable
- Block level random access

Performance

- 10ms random access latency
- Up to 200MiB/s streaming access

Typical use

- Desktops, data center bulk storage



Magnetic recording

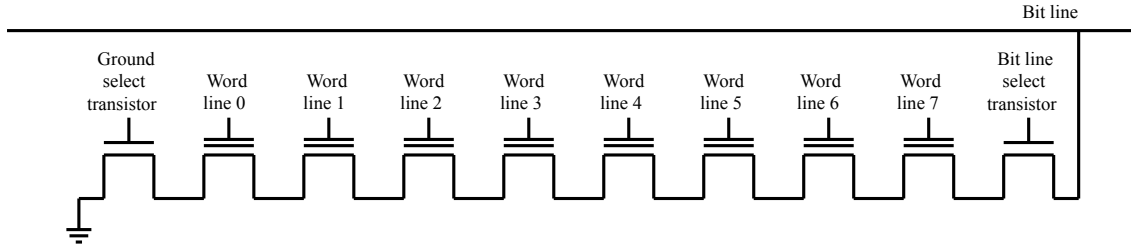
Hard drive

Technologies

Persistent memory

Flash/SSD

Solid State Drive



Characteristics

- Blocks of bits stored persistently in silicon (even when unpowered)
- Densely packed in 2D array (newly 3D)
- Electrically reprogrammable (*for a limited number of times*)
  - Writes must be to a clean page, no update in place
  - Erasing only for regions of blocks (~256 KiB)

Performance

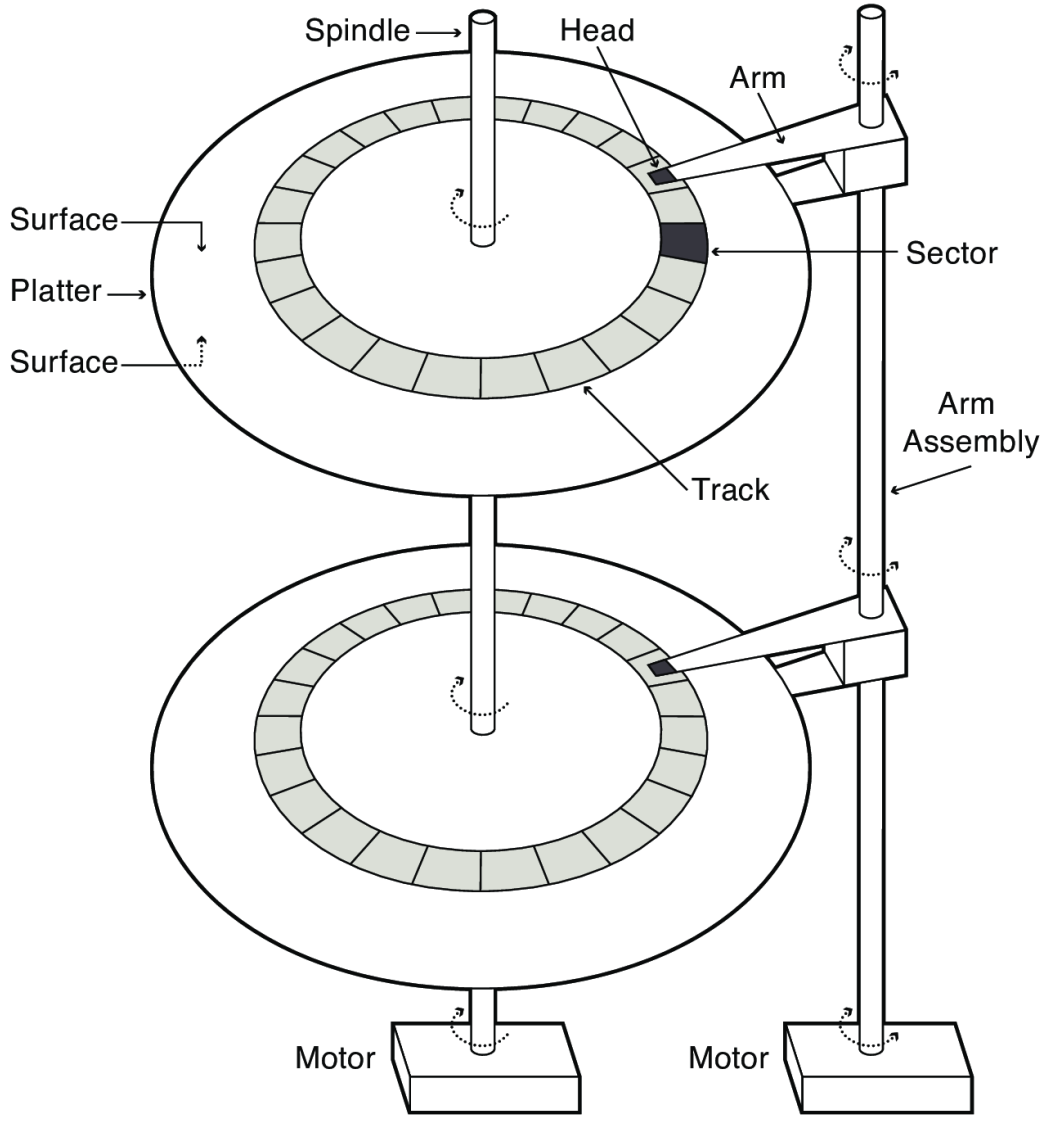
- 100µs random access latency
- 200MiB/s to +2000MiB/s

Typical use

- Smartphones, laptops, cameras

Magnetic disks

Anatomy



Magnetic disks

History

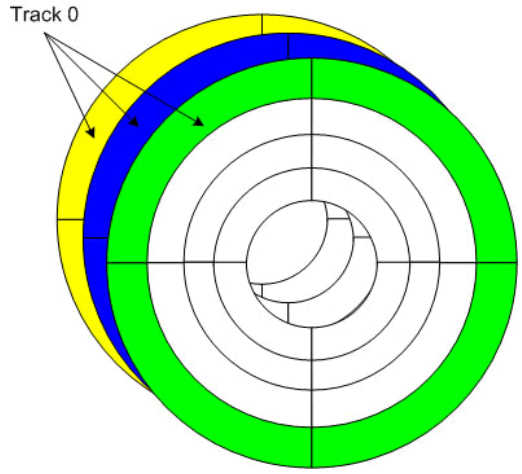
Principle hasn't really changed since the mid-1950s

IBM 305 hard drive

Magnetic disks

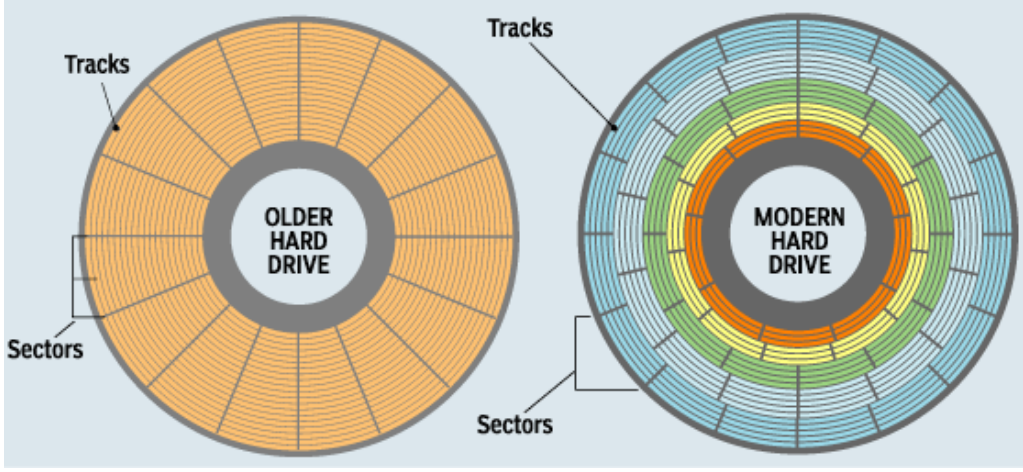
More about tracks

- ~ 1 micron wide
- Separated by unused guard regions to avoid corruptions
- Variable track length across disk



Sectoring

1. Uniform sectoring
2. ZBR (Zone Bit Recording)



Velocity

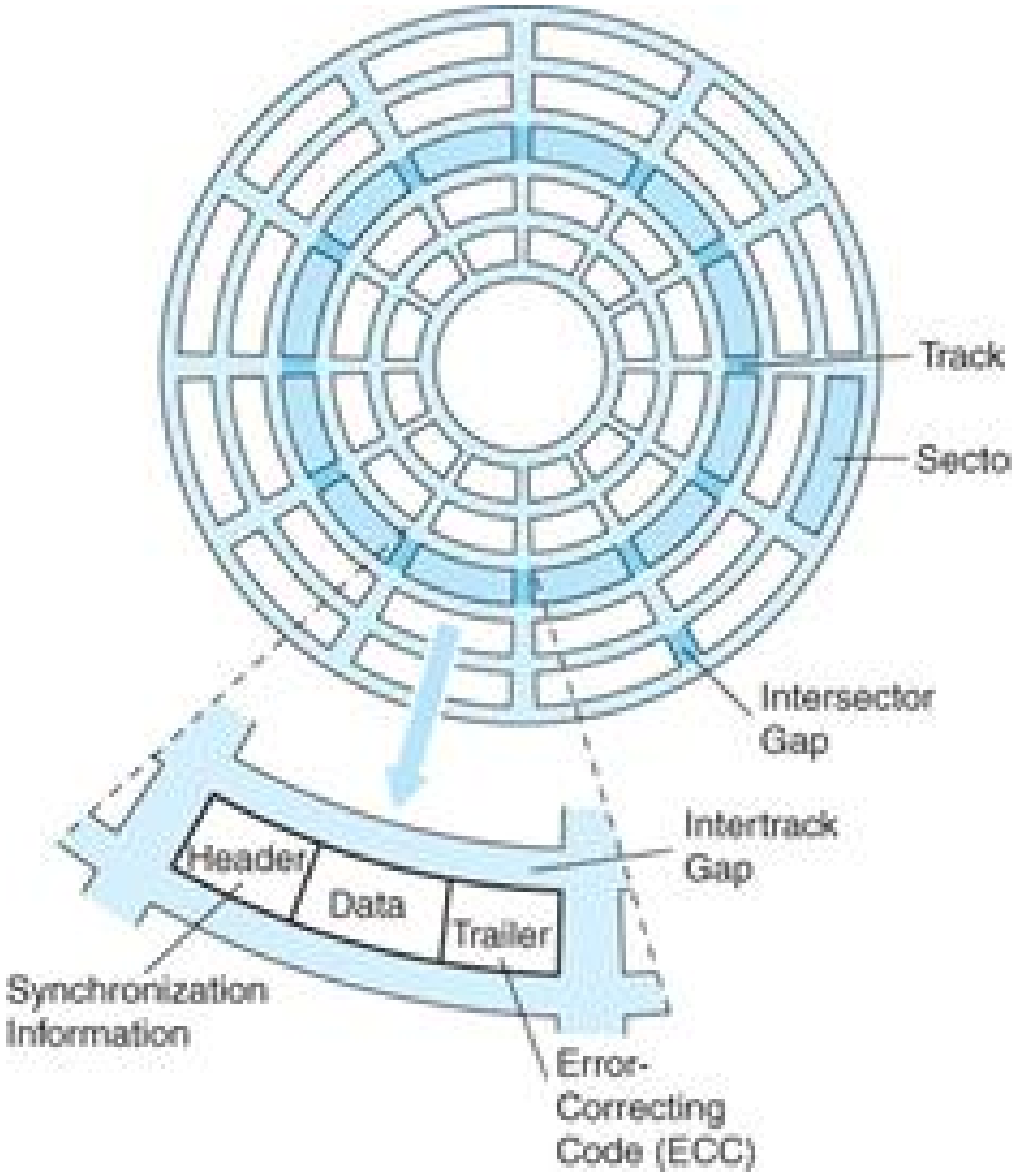
- CLV (Constant Linear Velocity): e.g. old CDROM
- CAV (Constant Angular Velocity): e.g. HDD

Magnetic disks

More about sectors

Composition

- Header
  - Sector ID, bad flag, header parity
- Data
  - Historically 512 bytes
  - 2048 bytes for CD/DVD
  - 4096 bytes for newer disks
- Error correcting codes (ECC)

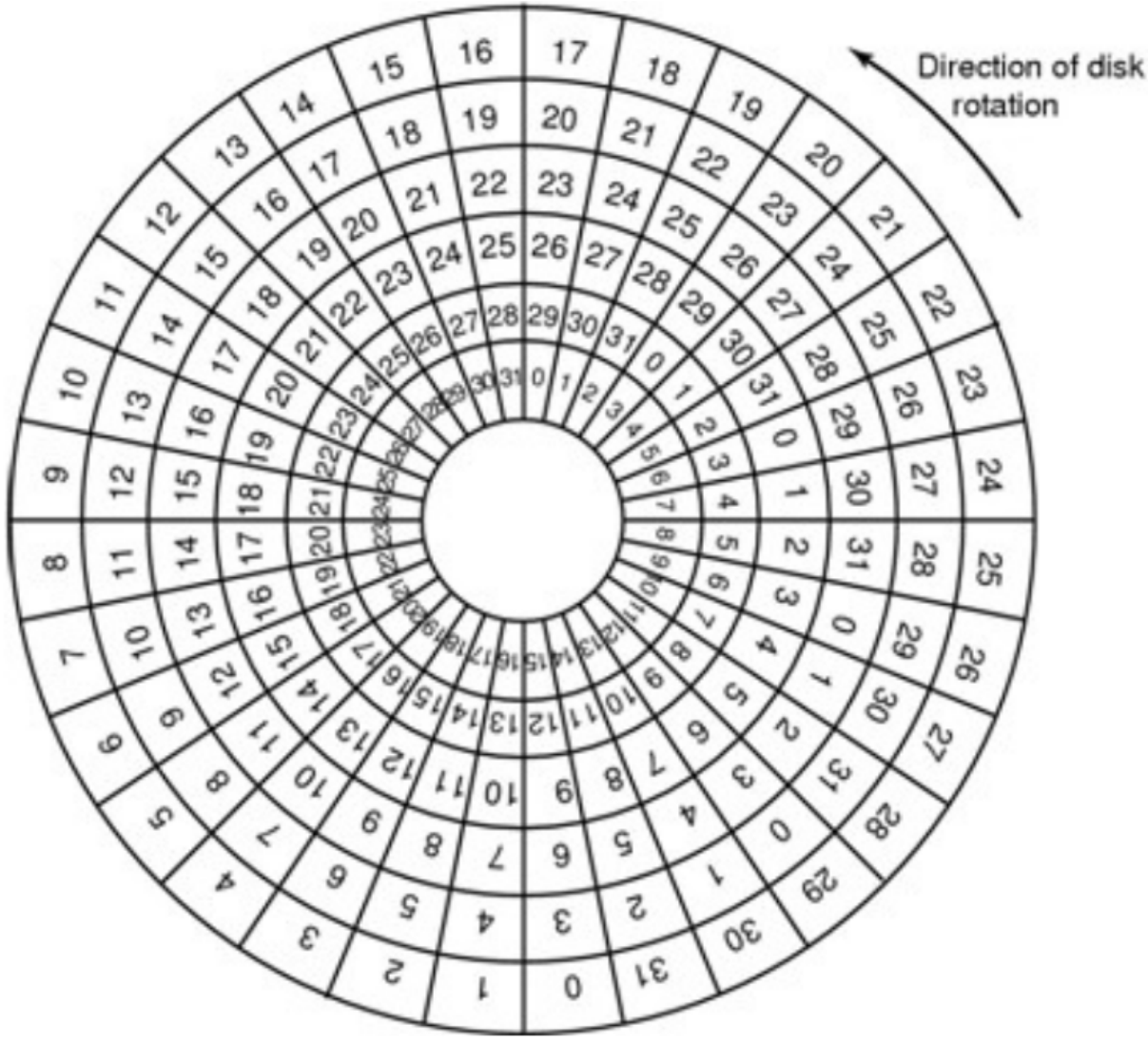


Addressing

- Old: CHS (Cylinder/Head/Sector)
- New: LBA (Logical Block Address)

Magnetic disks

Track skewing

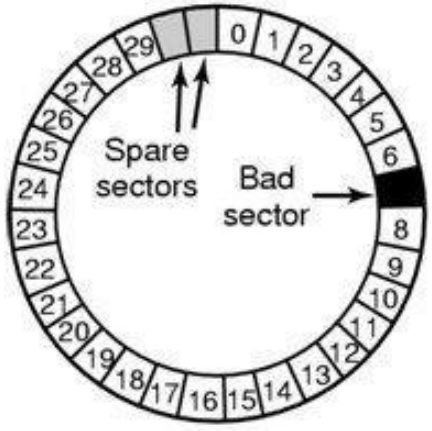


- Offset ordering between tracks to preserve sequential properties

Magnetic disks

Dealing with bad sectors

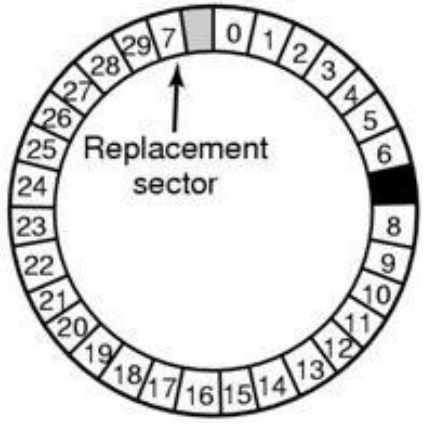
Spare sectors



- Keep provision of *spare* sectors on each track

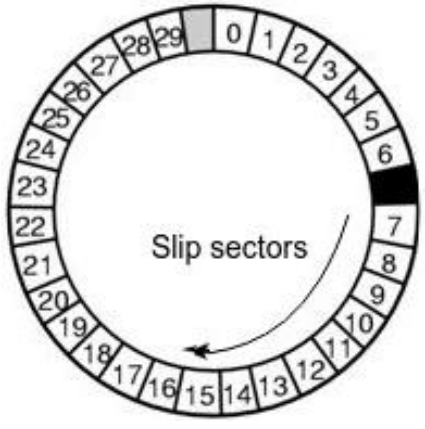
1. Sector sparing

- Remap bad sector transparently



2. Slip sparing

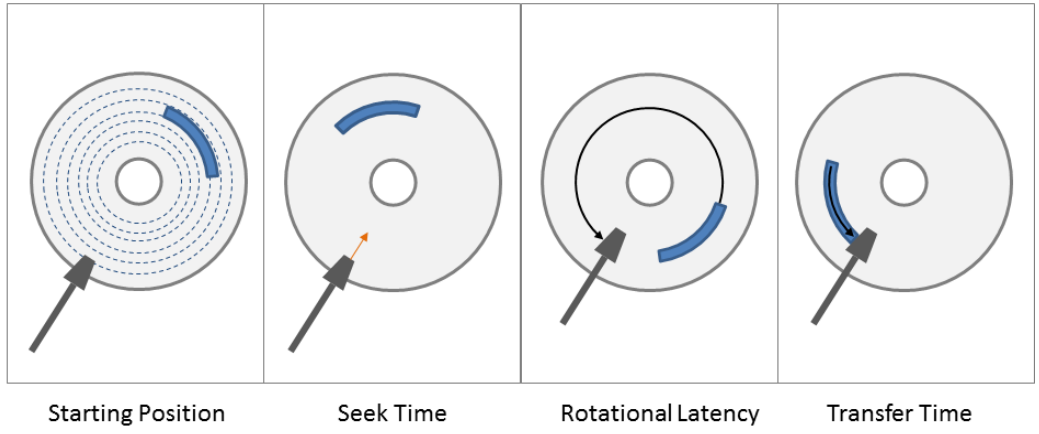
- Remap all sectors to preserve sequential properties



Magnetic disks

Disk operations

- When accessing a sector:
  1. Arm moves to correct cylinder, and proper head is enabled to reach the track containing the sector
     - *Seek time* (+ settle time)
  2. Wait for sector to appear under head
     - *Rotation time*
  3. Read/write sector as it spins by
     - *Transfer time*
- *Access time = seek time + rotation time + transfer time*

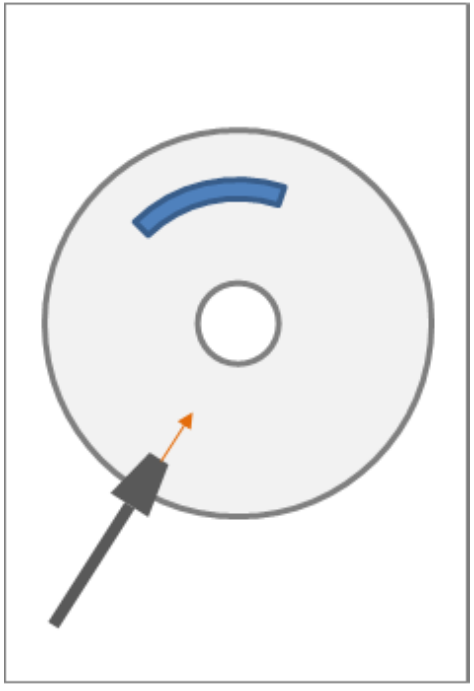


Magnetic disks

Disk performance

Seek time

- Time to position the head over a track
  - Depends on how fast the arm assembly moves the arms
- Head switch time (i.e. same cylinder, but different head/track)



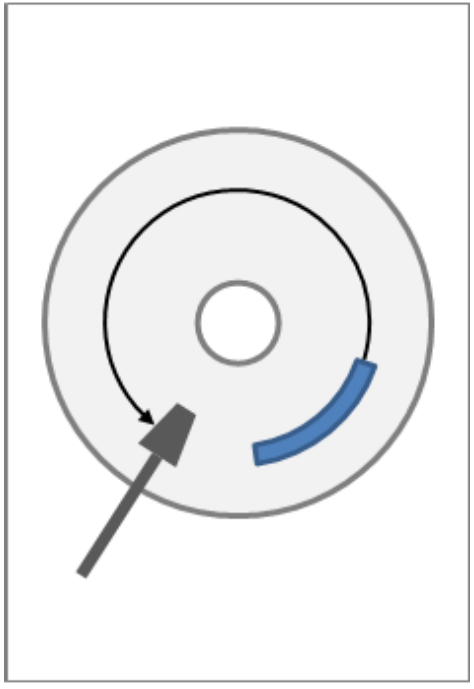
Seek Time

- Maximum seek time
  - From innermost track to outermost track
  - ~10ms to 20ms
- Minimum seek time
  - From one track to the next one
  - ~1ms
- Average seek time
  - Average between each possible pairs of tracks
  - 1/3 maximum time

Magnetic disks

Rotation time

- Time for the sector to appear underneath the head
  - Depends on how fast the disk spins (e.g. 4200/5400/7200/10k/15k RPM)



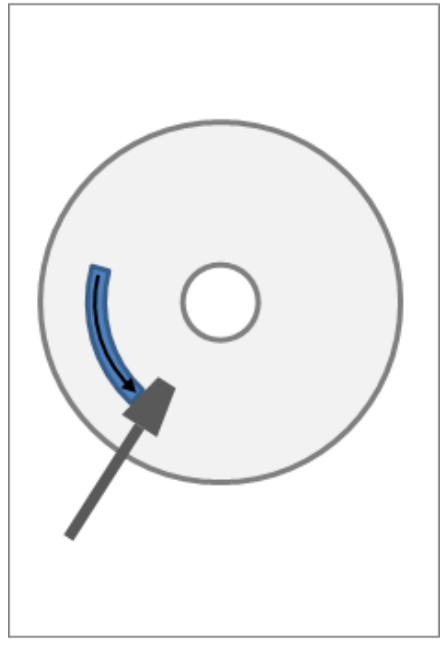
Rotational Latency

- Rotation latency is typically half of full rotation
  - ~15ms to 4ms

Magnetic disks

Transfer time

- Time to move the bytes from disk to memory
- Surface transfer time *(from surface to disk buffer)*
- Host transfer time *(from disk buffer to main memory)*



Transfer Time

Magnetic disks

Example: Toshiba MK3254GSY (2009)

Specifications	
Platters/Heads	2/4
Capacity	320 GiB
Spindle speed	7200 RPM
Average seek time R/W	10.5/12 ms
Track-to-track	1 ms
Surface transfer time	54-128 MiB/s
Host transfer time	375 MiB/s
Buffer	16 MiB

Magnetic disks

Example: 500 random reads

Specifications	
Platters/Heads	2/4
Capacity	320 GiB
Spindle speed	7200 RPM
Average seek time R/W	10.5/12 ms
Track-to-track	1 ms
Surface transfer time	54-128 MiB/s
Host transfer time	375 MiB/s
Buffer	16 MiB

Description

- Workload
  - 500 read requests
  - Randomly chosen sectors
  - Served in FIFO order
- How long to service them?
  - seek time: 10.5 ms
  - rotation time: 4.15 ms
  - transfer time: at least 54 MiB/s

Result

- Seek time: 10.5 ms
- Rotation time: 4.15 ms
  - 7200RPM => 120 RPS => 8.3 ms/rotation
- Transfer time: 9 µs
  - 512 bytes at 54 MiB/s
- 500 * (10.5 ms + 4.15 ms + 9 µs) = 7.3 s!

Magnetic disks

Example: 500 sequential reads

Specifications	
Platters/Heads	2/4
Capacity	320 GiB
Spindle speed	7200 RPM
Average seek time R/W	10.5/12 ms
Track-to-track	1 ms
Surface transfer time	54-128 MiB/s
Host transfer time	375 MiB/s
Buffer	16 MiB

Description

- Workload
  - 500 read requests
  - Sequential sectors on same track
- How long to service them?
  - seek time: 10.5 ms
  - rotation time: 4.15 ms
  - transfer time: 54-128 MiB/s

Result

- Seek time: 10.5 ms
- Rotation time: 4.15 ms
  - 7200RPM => 120 RPS => 8.3 ms/rotation
- Transfer time:
  - outer track: 4 µs (512 bytes at 128 MiB/s)
  - inner track: 9 µs (512 bytes at 54 MiB/s)
- 10.5 + 4.15 + 500 * 4 µs = 16.65 ms
- 10.5 + 4.15 + 500 * 9 µs = 19.15 ms

Magnetic disks

Disk scheduling

Rationale

- Seek and rotation times dominate the cost of small accesses
- Disk transfer bandwidth is wasted
- Need algorithms to reduce seek time

ECS 150 - Storage

Prof. Joël Porquet-Lupine

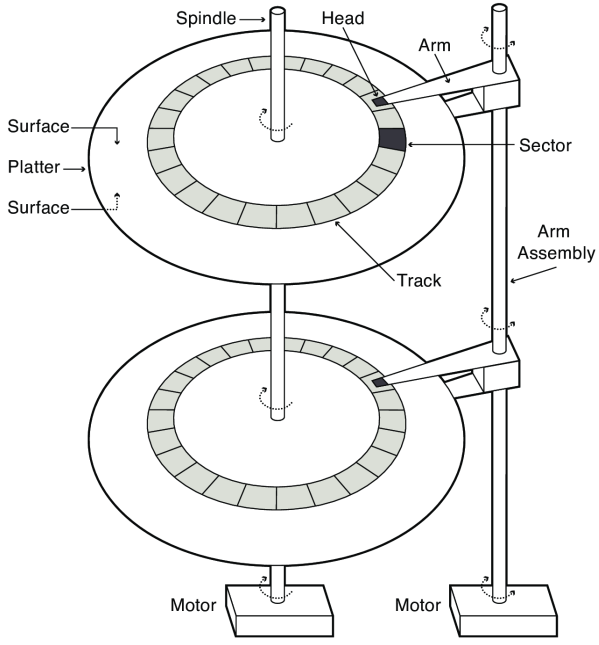
UC Davis - 2020/2021

Recap

Technologies

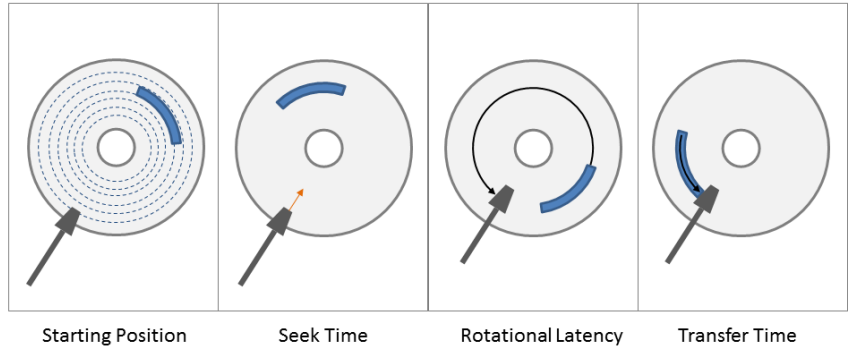
- Memory
  - SRAM
  - DRAM
- Secondary storage
  - Magnetic disk
  - Flash memory

Magnetic disks



Disk performance

- *Access time = seek time + rotation time + transfer time*



Random vs sequential reads

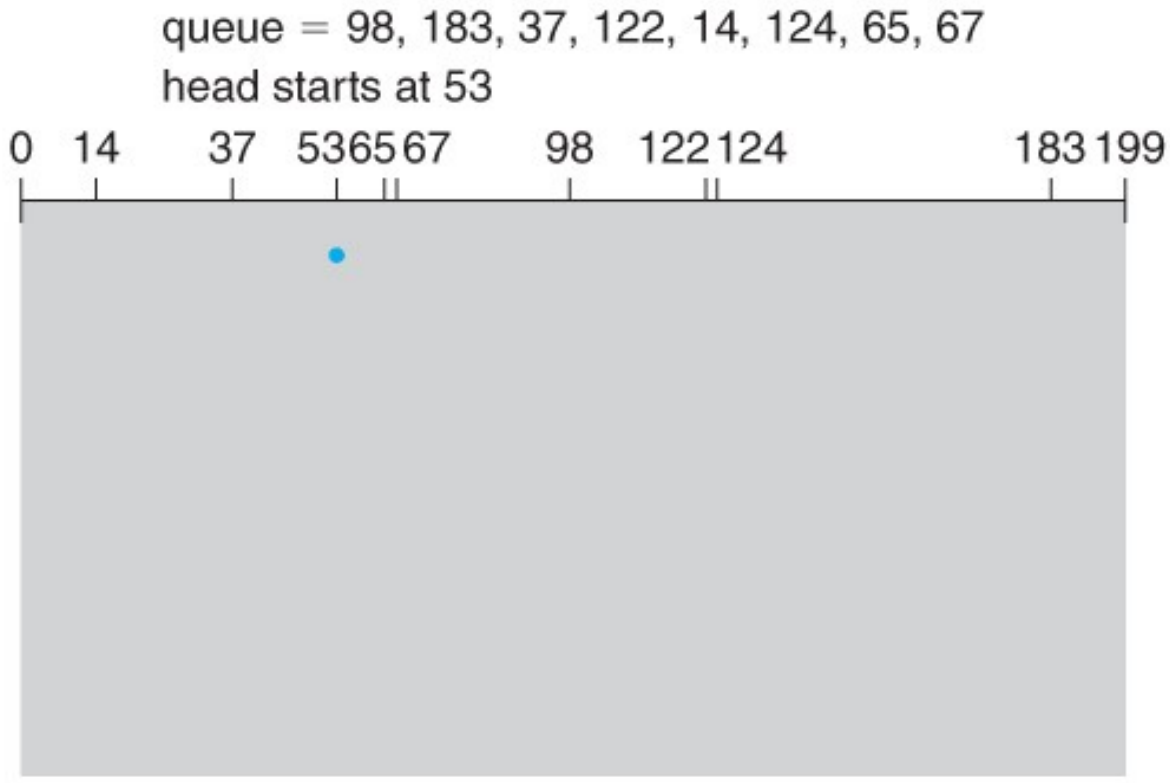
- Example Toshiba MK3254GSY
  - 500 random reads: ~7s
  - 500 sequential reads: ~20ms
- Seek time dominates access time
  - Need algorithms to reduce it

Magnetic disks

Disk scheduling

Scheduling benchmark

- Queue of disk I/O requests

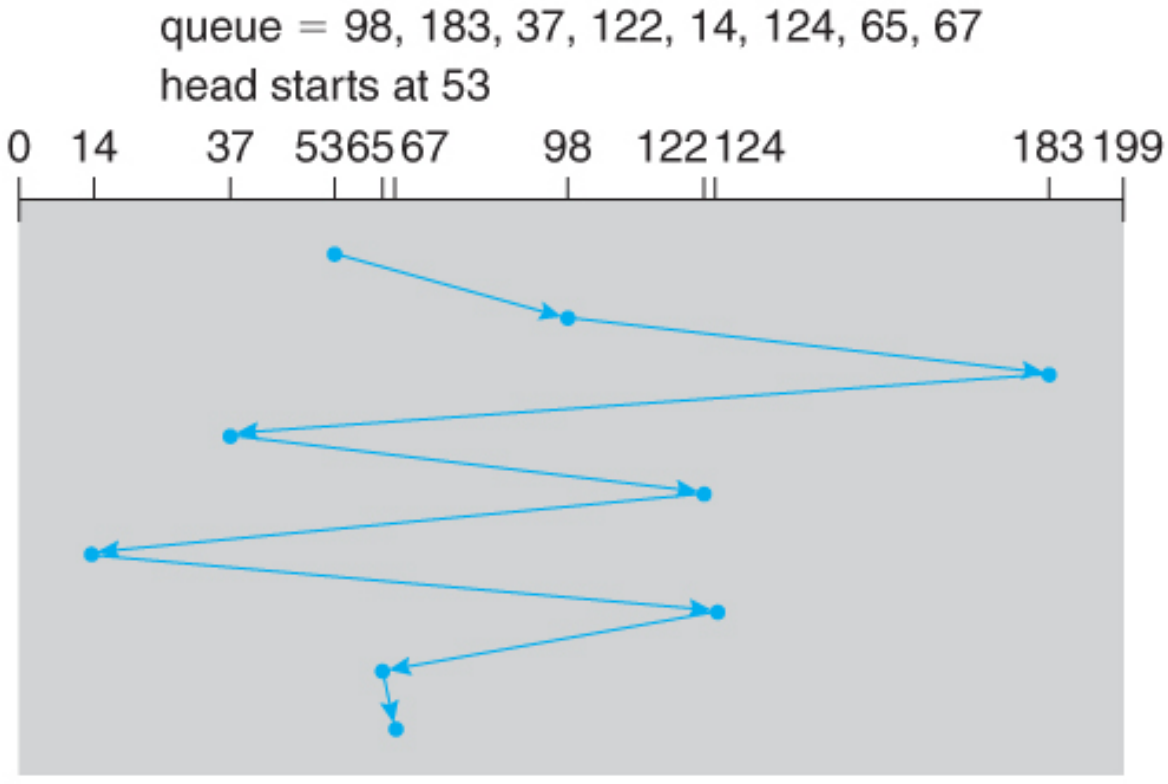


- Objective: (re-)schedule requests to minimize seek time
- Metric: total head movement (in number of tracks)

Magnetic disks

Scheduling: FCFS

- *First come, first server* (aka FIFO)

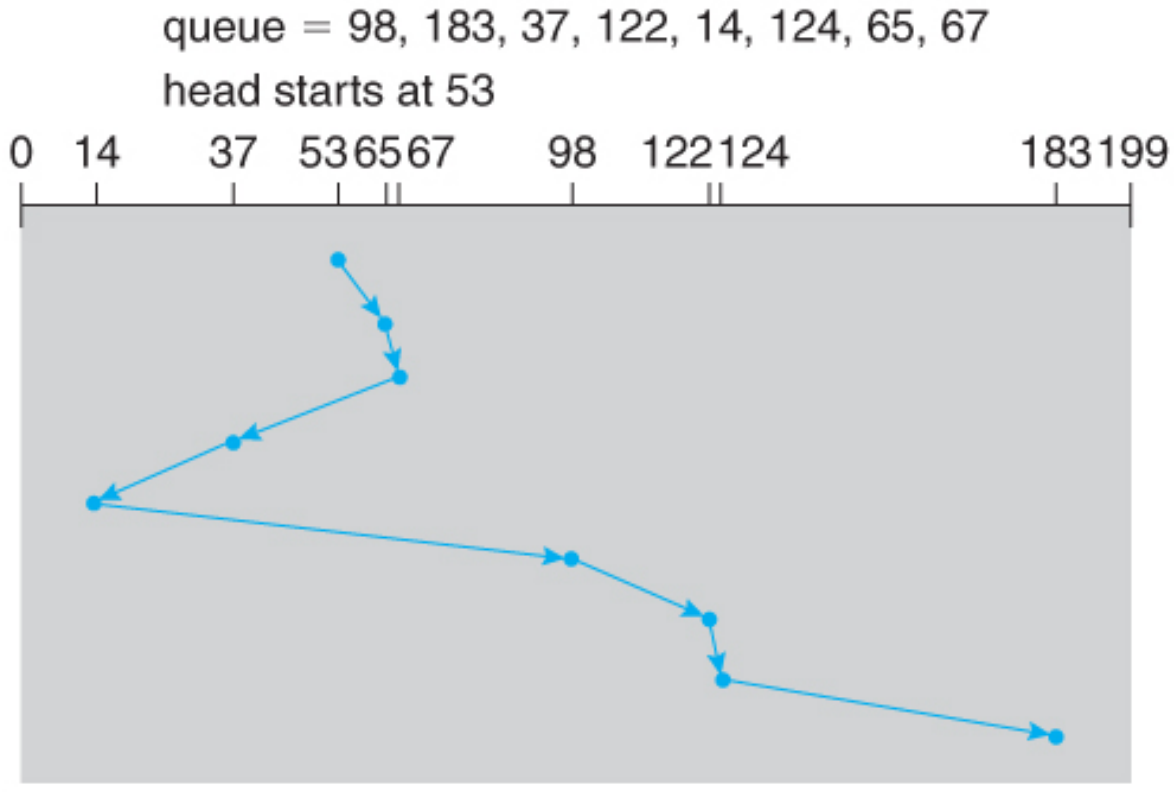


- Total head movement: 640 tracks

Magnetic disks

Scheduling: SSTF

- *Shortest seek time first*

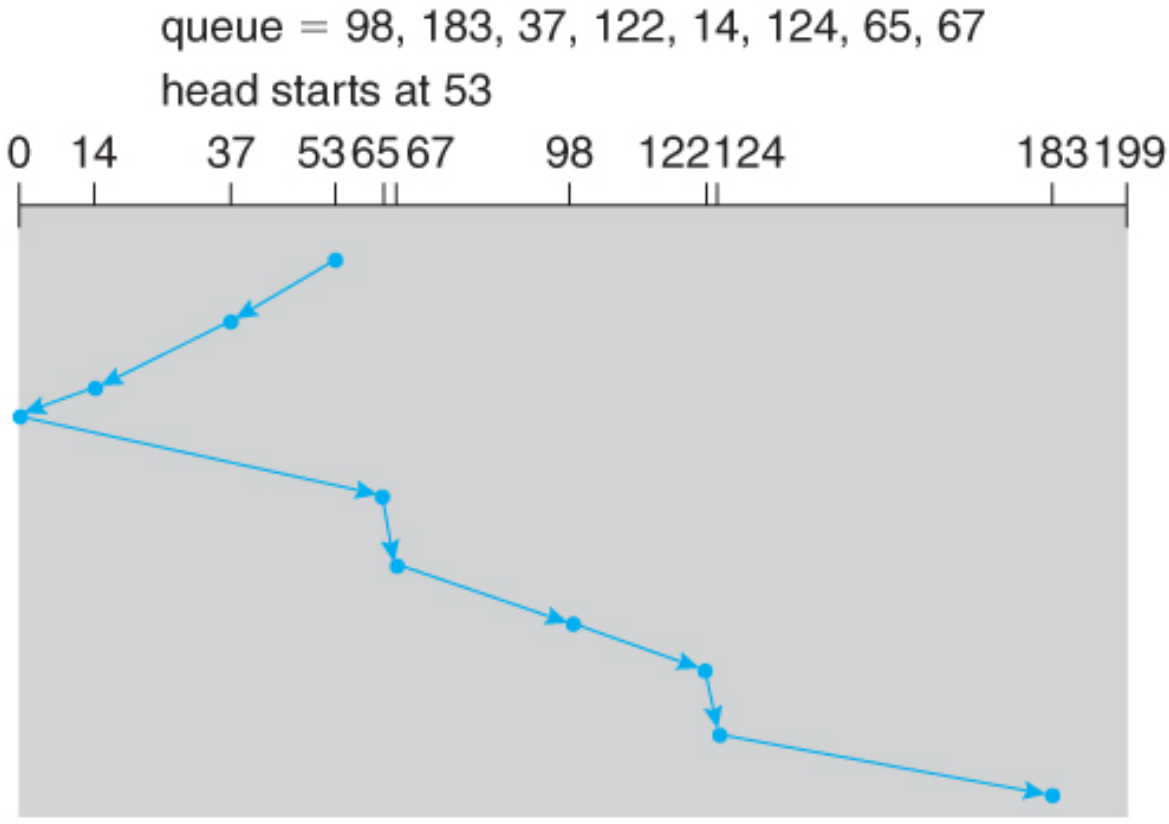


- Total head movement: 236 tracks

Magnetic disks

Scheduling: SCAN

- The *elevator* algorithm

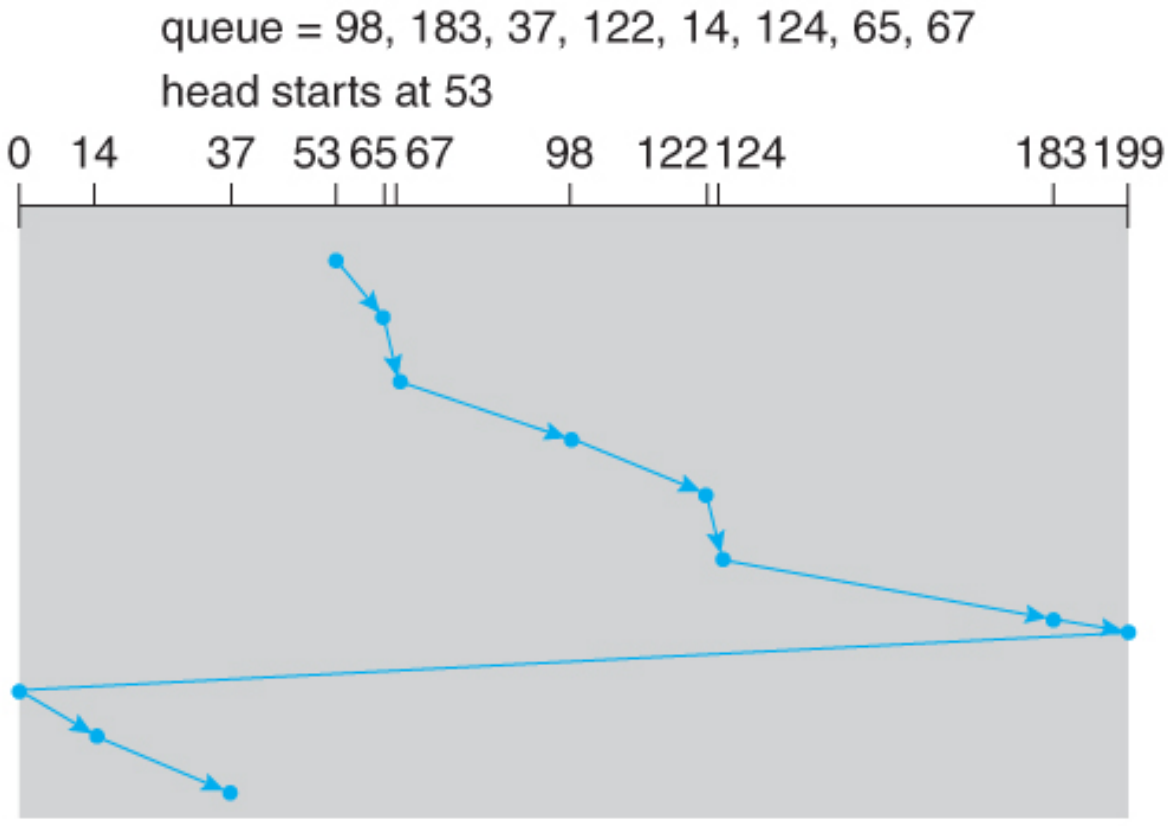


- Total head movement: 208 tracks

Magnetic disks

Scheduling: C-SCAN

- The *circular* elevator algorithm
  - Goes back directly to beginning after scanning

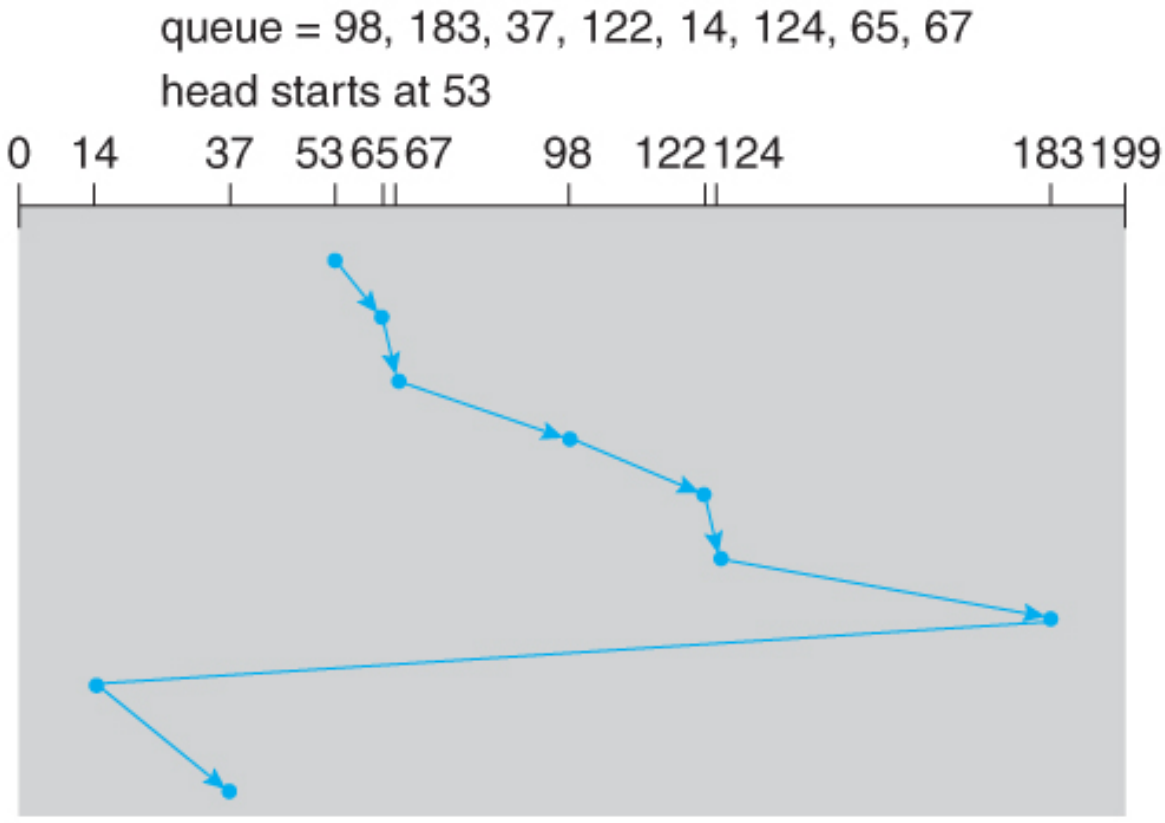


- Total head movement: 183 tracks (+200 for return trip)

Magnetic disks

Scheduling: C-LOOK

- Optimized C-SCAN
  - Goes only as far as last request in each direction



- Total head movement: 153 tracks (+169 for return trip)

Magnetic disks

Scheduling

Other algorithms

- R-CSCAN
  - Account for rotation time
  - Allow small steps back and forth during scanning
- F-SCAN
  - Two I/O request queues to prevent arm "stickiness"
  - Service one queue, while new requests are enqueued in other queue
  - At the end of scan, swap queues
- N-SCAN
  - Same as F-SCAN but multiple queues

Summary

- FCFS
- SSTF
- Elevator algorithms (e.g., SCAN, C-SCAN, C-LOOK)

Magnetic disks

Effects of disk scheduling (C-LOOK)

Specifications	
Platters/Heads	2/4
Capacity	320 GiB
Spindle speed	7200 RPM
Average seek time R/W	10.5/12 ms
Track-to-track	1 ms
Surface transfer time	54-128 MiB/s
Host transfer time	375 MiB/s
Buffer	16 MiB

Description

- Workload
  - 500 read requests
  - Randomly chosen sectors
  - Disk head on outside track
  - Served in C-LOOK order
- How long to service them?
  - seek time: estimated as 1-track seek + 0.2% seek
  - rotation time: 4.15 ms
  - transfer time: at least 54 MiB/s

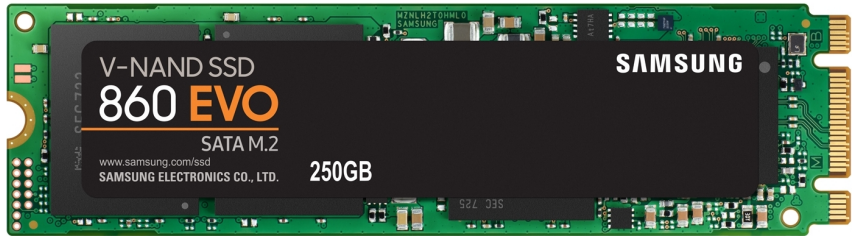
Result

- Seek time: 1.06 ms
  - Estimated 0.2% seek: 1ms + (0.2/33.3) * 10.5 ms
- Rotation time: 4.15 ms
  - 7200RPM => 120 RPS => 8.3 ms/rotation
- Transfer time: 9 µs
  - 512 bytes at 54 MiB/s
- 500 * (1.06ms + 4.15ms + 9 µs) = 2.61 s!

Flash storage

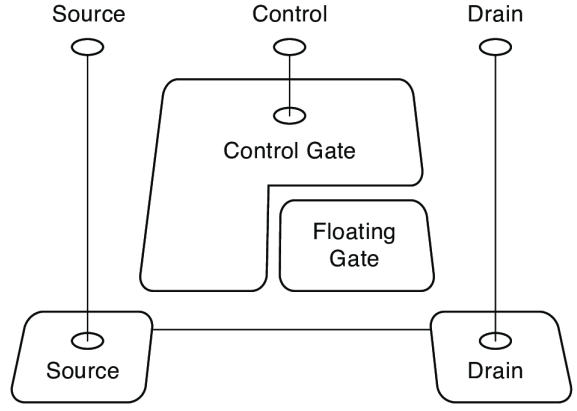
Characteristics

- No moving parts
- Better random access performance
- Less power
- More resistant to physical damage
- But also, more expensive…



Technologies

- NOR vs **NAND**
- Single- vs Multi-level



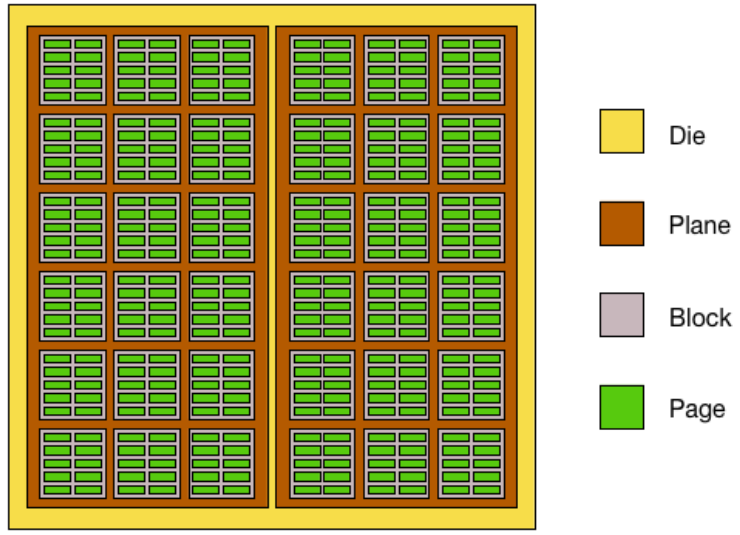
Flash storage

Organization

Typical sizes:

- Page: 4 KiB
- Block: 128 pages (512 KiB)
- Plane: 1024 blocks (512 MiB)

- Multiple independent data paths accessible in parallel



Operations

- Read and writes only occur in page units

- Read page: ~10 µs
- Write page: ~100 µs
  - Can only write an empty page (and not update existing page)
  - But pages can only be emptied at block level
- Erase block: > 1 ms

Flash storage

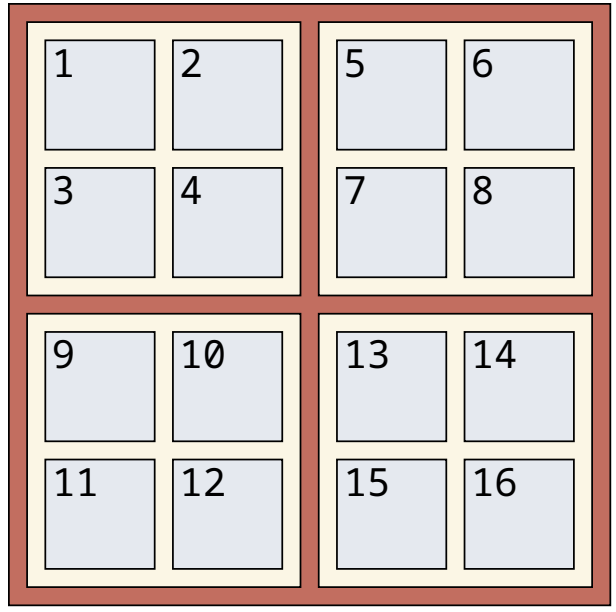
Page writing

- How long does it take to write to a single page?
- Example flash drive specifications
  - 4 KiB page
  - 3 ms block erasure time
  - 512 KiB block (128 pages)
  - 50 µs read/write page

Naive approach

- Read block (except new page)
- Erase block
- Rewrite block + new page

- $Total = 127 * 50\mu s + 3ms + 128 * 50\mu s = 16ms$



Flash storage

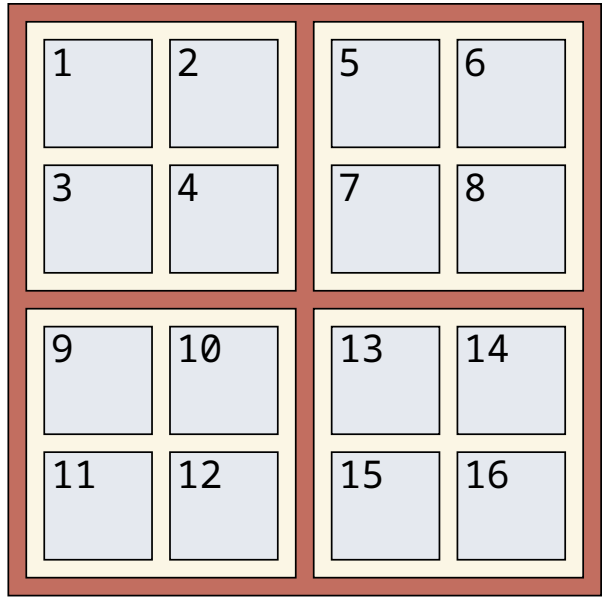
Page writing

- How long does it take to write to a single page?
- Example flash drive specifications
  - 4 KiB page
  - 3 ms block erasure time
  - 512 KiB block (128 pages)
  - 50 µs read/write page

Smarter approach

- *Flash translation layer*
  - Map logical pages to physical pages
- Make free erased block(s)
- Cost of erasure is amortized
- $Total = (3ms/128) + 50\mu s = 73.4\mu s$

Logic	Phys
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12



Flash storage

Durability

Wear out

Flash memory stops reliably storing a bit

- After many erasures (on the order of 10^3 to 10^6)
- After nearby cells are read many times (read disturb)

Solutions

- Error correcting codes
- Wear leveling
  - Using write remapping
- Bad pages/erasure blocks
- Spare pages and erasure blocks

Flash storage

Example: Intel 710 series SSD

Specifications	
Capacity	300 GiB
Page size	4 KB
Bandwidth (seq reads)	270 MiB/s
Bandwidth (seq writes)	210 MiB/s
R/W latency	75 µs
Random reads/s	38,500 (ie 26 µs/read)
Random writes/s	2,000

Description

- Workload
  - 500 read requests
  - Randomly chosen sectors
- How long to service them?

Result

- 500 * 26 µs = 13 ms
  - (compared to 7.3 s for magnetic disk...!)