CS330 Operating Systems and Lab.

Mass-Storage Systems

Spring 2024

Instructor : Insik Shin

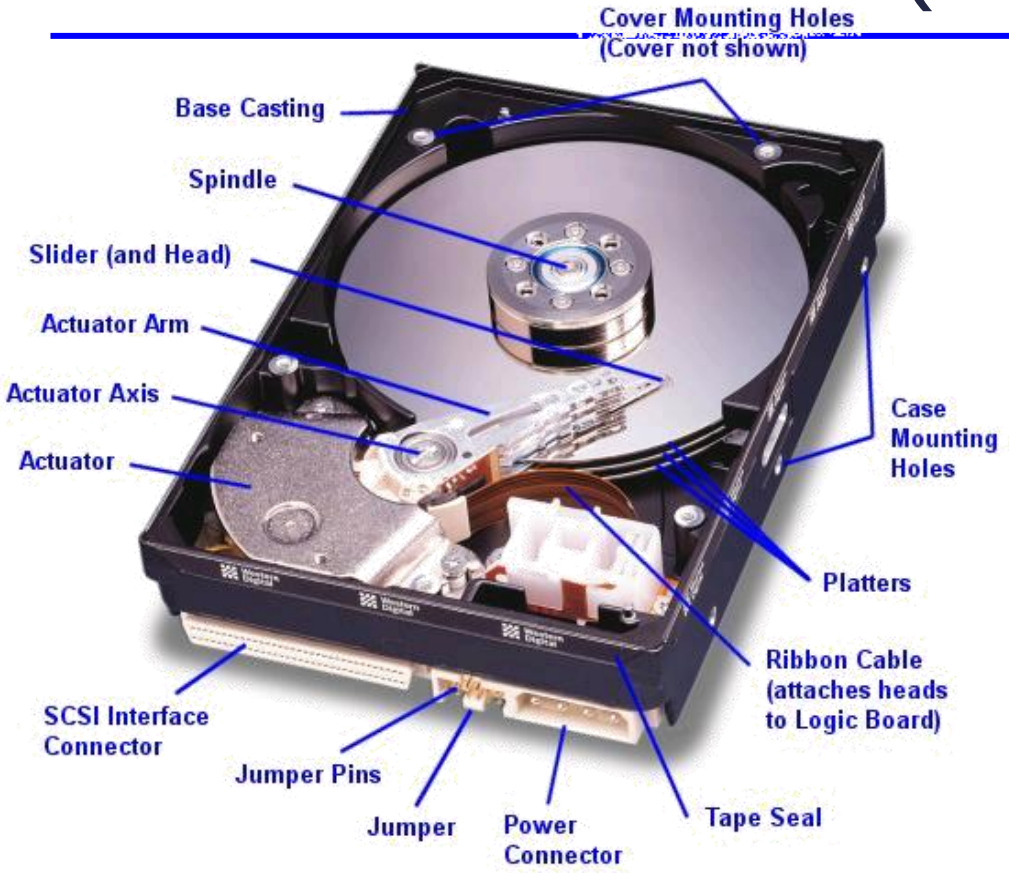
Logistics

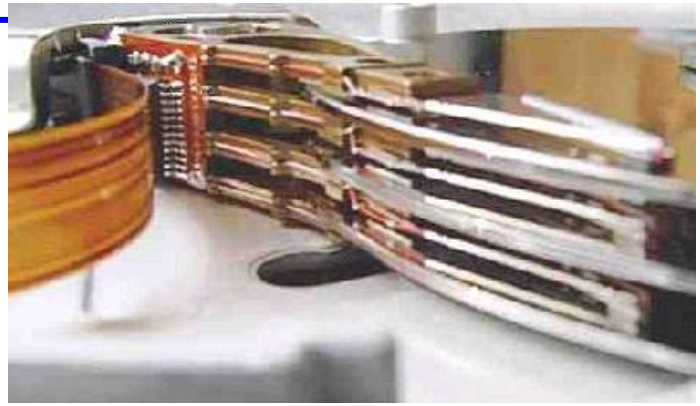
- No Class
  - May 6 (Mon) : Holiday
  - May 13 (Mon) : Conference Trip
  - May 15 (Wed) : Holiday

Mass-Storage Systems

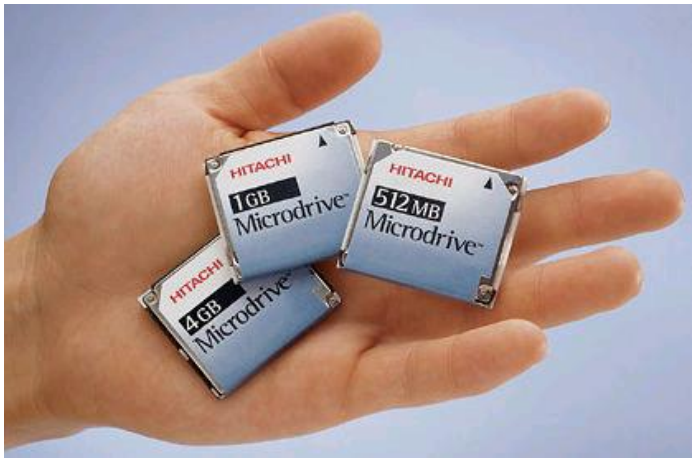
- Disks and SSDs

Hard Disk Drives (HDDs)





Read/Write Head
Side View



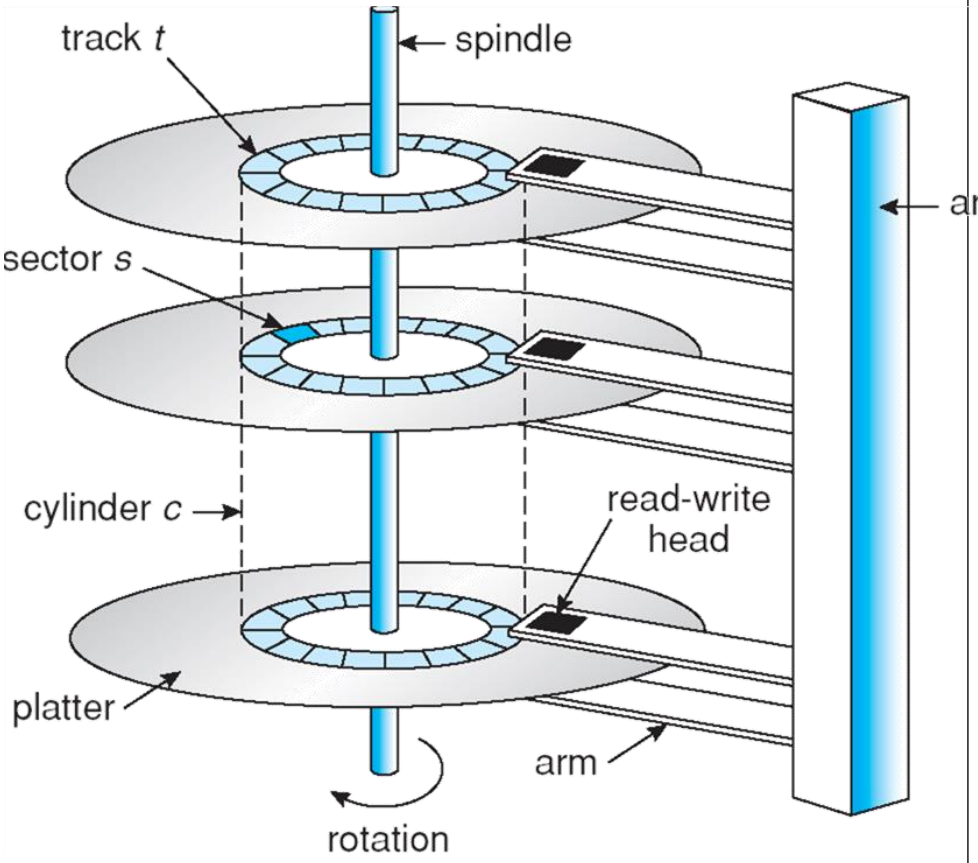
IBM/Hitachi Microdrive

IBM Personal Computer/AT (1986)

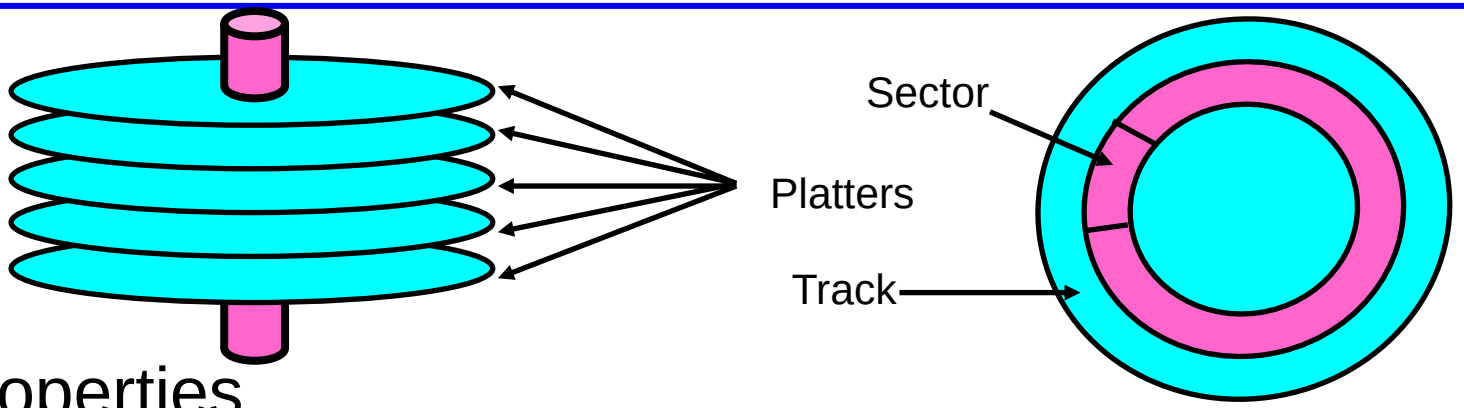
- 30 MB hard disk - $500
- 30-40ms seek time
- 0.7-1 MB/s (est.)

Disk

- Stack of magnetic platters
  - Rotate together on a central spindle at 3,600-15,000 RPM
- Disk arm assembly
  - Arms rotate around pivot, all move together
  - Arms contain disk heads – one for each recording surface
  - Heads read and write data to platters

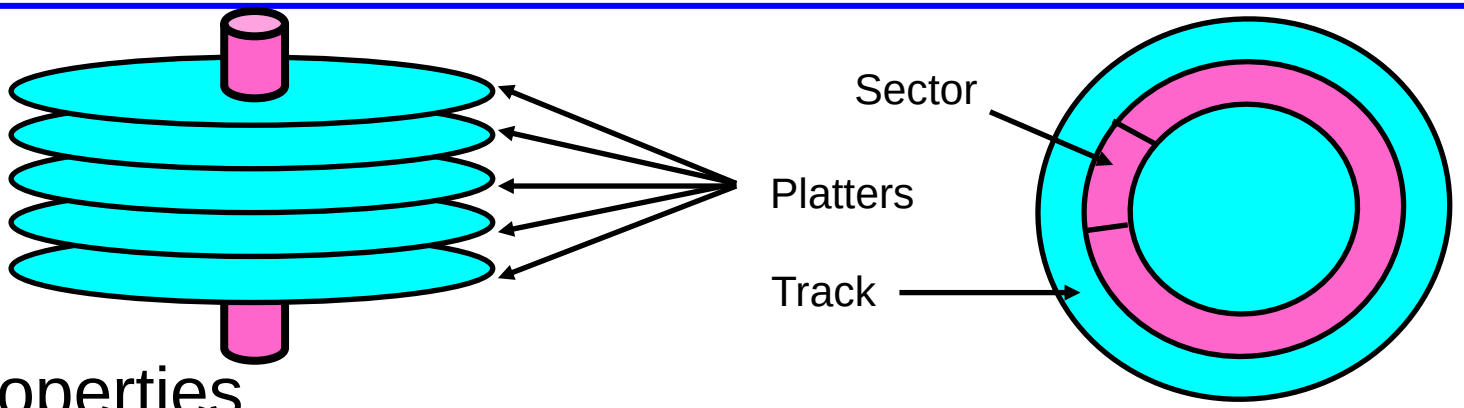


Properties of a Magnetic Hard Disk

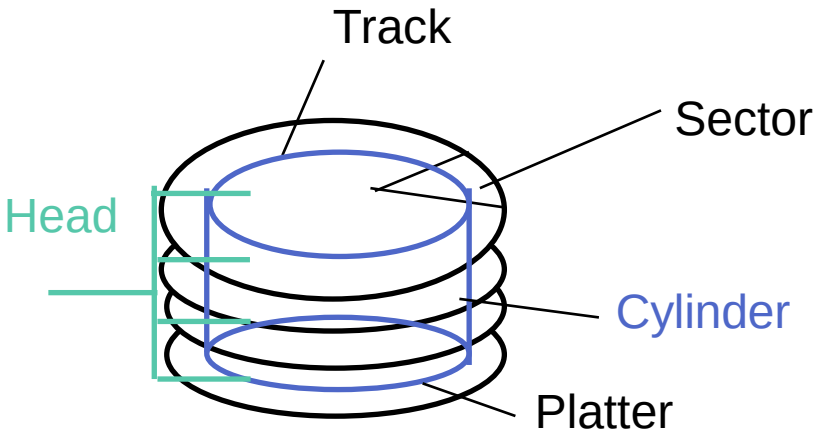


- Properties
  - Independently addressable element: sector
    - OS always transfers groups of sectors together—"blocks"
  - A disk can access directly any given block either sequentially or randomly.
- Typical numbers (depending on the disk size):
  - 500 to more than 20,000 tracks per surface
  - 32 to 800 sectors per track
- Zoned bit recording
  - Constant bit density: more bits (sectors) on outer tracks

Properties of a Magnetic Hard Disk



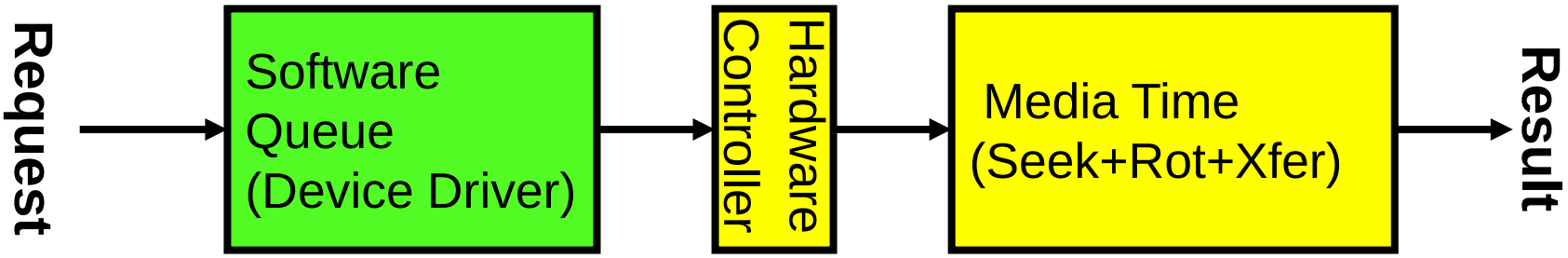
- Properties
  - Independently addressable element: sector
    - OS always transfers groups of sectors together—"blocks"
  - Cylinder: all the tracks under the head at a given point on all surfaces



Magnetic Disk Characteristic

- Read/write: three-stage process:
  - **Seek time**: position the head/arm over the proper track (into proper cylinder)
  - **Rotational latency**: wait for the desired sector to rotate under the read/write head
  - **Transfer time**: transfer a block of bits (sector) under the read-write head

- Disk Latency = Queuing Time + Controller time + Seek Time + Rotation Time + Xfer Time



- Highest Bandwidth:
  - Transfer large group of blocks sequentially from one track

Typical Numbers of a Magnetic Disk

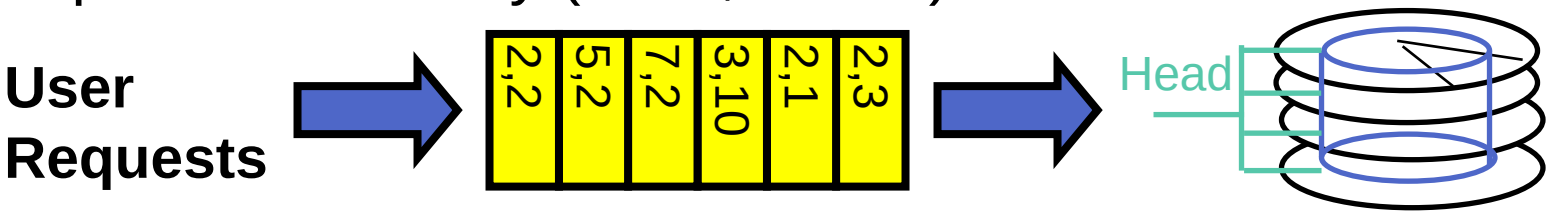
Parameter	Info / Range
Average seek time	**Typically 5-10 milliseconds**. Depending on reference locality, actual cost may be 25-33% of this number.
Average rotational latency	Most laptop/desktop disks rotate at 3600-7200 RPM (8-16 ms/rotation). Server disks up to 15,000 RPM. Average latency is halfway around disk yielding corresponding times of **4-8 milliseconds**
Controller time	Depends on controller hardware
Transfer time	**Typically 50 to 100 MB/s**. Depends on: • Transfer size (usually a sector): 512B – 1KB per sector • Rotation speed: 3600 RPM to 15000 RPM • Recording density: bits per inch on a track • Diameter: ranges from 1 in to 5.25 in
Cost	Drops by a factor of two every 1.5 years (or even faster). **$0.01/GB in 2021** ($0.025/GB in 2019)

Disk Performance Examples

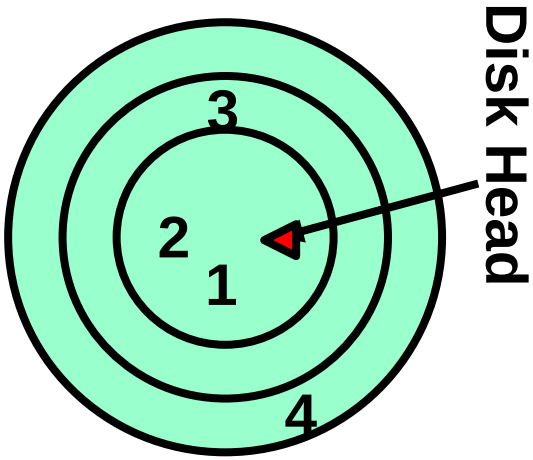
- Assumptions:
  - Ignoring queuing and controller times for now
  - Avg seek time of 5ms,
  - 7200RPM $\Rightarrow$ Time for one rotation: 60000ms/7200 ~= 8ms
  - Transfer rate of 4MByte/s, sector size of 1 KByte
- Read sector from random place on disk:
  - Seek (5ms) + Rot. Delay (4ms) + Transfer (0.25ms)
  - Approx 10ms to fetch/put data: **100 KByte/sec**
- Read sector from random place in same cylinder:
  - Rot. Delay (4ms) + Transfer (0.25ms)
  - Approx 5ms to fetch/put data: **200 KByte/sec**
- Read next sector on same track:
  - Transfer (0.25ms): **4 MByte/sec**
- Key to using disk effectively (especially for file systems) is to *minimize seek and rotational delays*

Disk Scheduling

- Disk can do only one request at a time; What order do you choose when handling queued requests?
  - Request denoted by (track, sector)



- Scheduling algorithms:
  - First In First Out (FIFO)
  - Shortest Seek Time First
  - SCAN
  - C-SCAN



- In our examples we will ignore the sector
  - Consider only track #

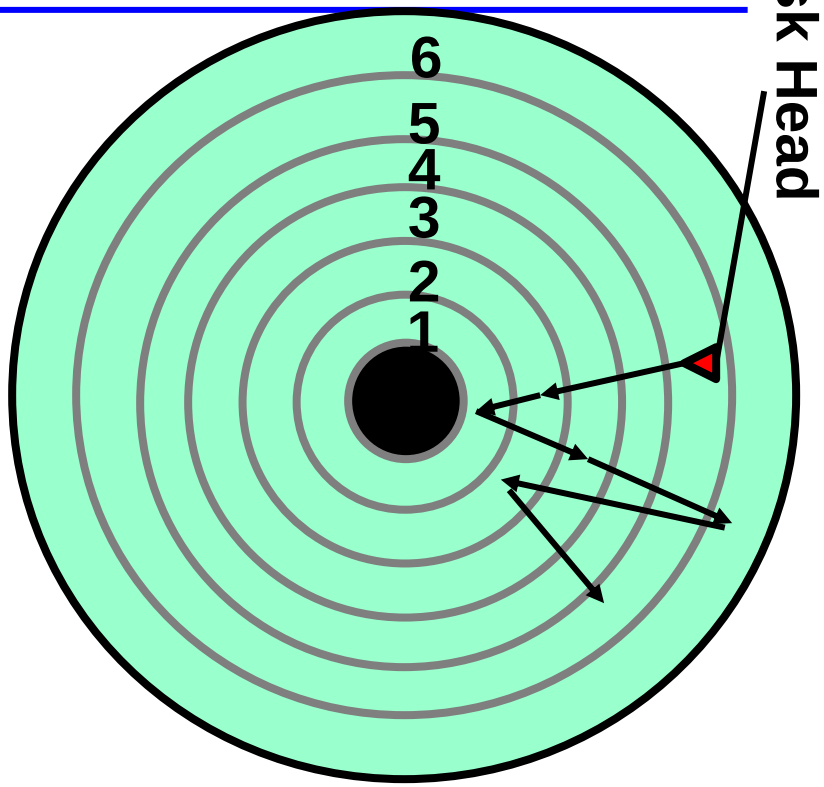
FIFO: First In First Out

- Schedule request in the order they arrive in the queue

- Example:
  - Request queue: 2, 1, 3, 6, 2, 5
  - Scheduling order: 2, 1, 3, 6, 2, 5

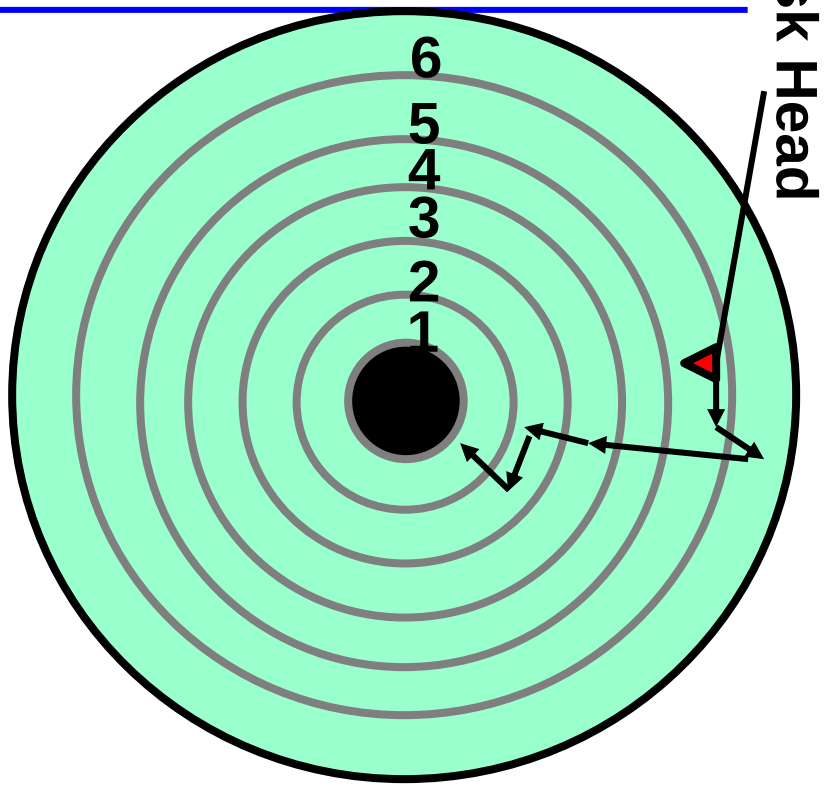
  - Pros: Fair among requesters

  - Cons: Order of arrival may be to random spots on the disk ⇒ Very long seeks



SSTF: Shortest Seek Time First

- Pick the request that's closest to the head on the disk
  - Although called SSTF, include rotational delay in calculation, as rotation can be as long as seek
- Example:
  - Request queue: 2, 1, 3, 6, 2, 5
  - Scheduling order: 5, 6, 3, 2, 2, 1
- Pros: reduce seeks
- Cons: may lead to starvation



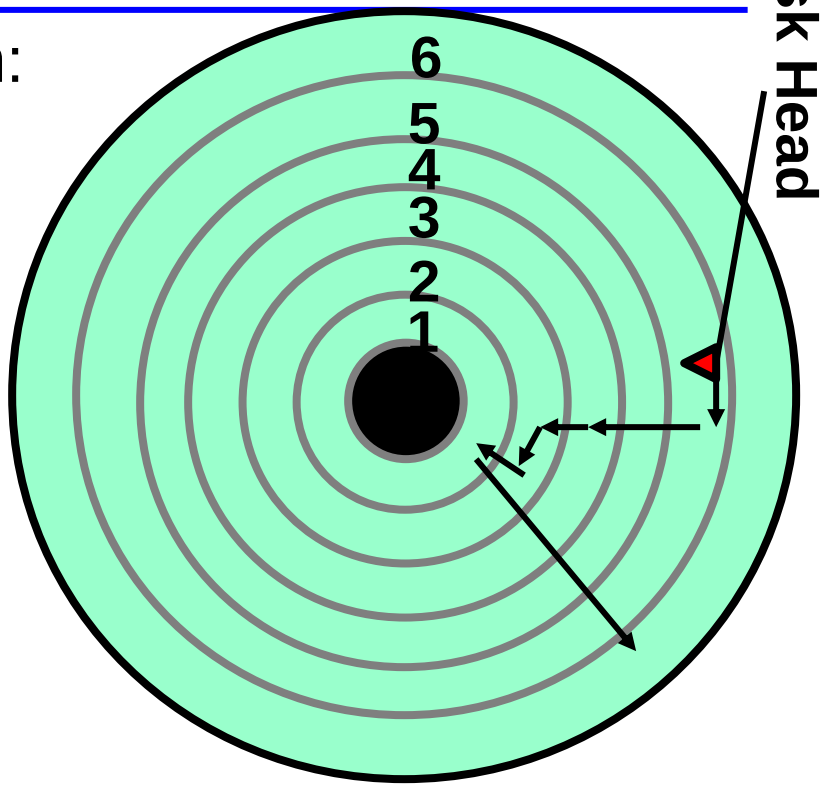
SCAN

- Implements an Elevator Algorithm: take the closest request in the direction of travel

- Example:
  - Request queue: 2, 1, 3, 6, 2, 5
  - Head is moving towards center
  - Scheduling order: 5, 3, 2, 2, 1, 6

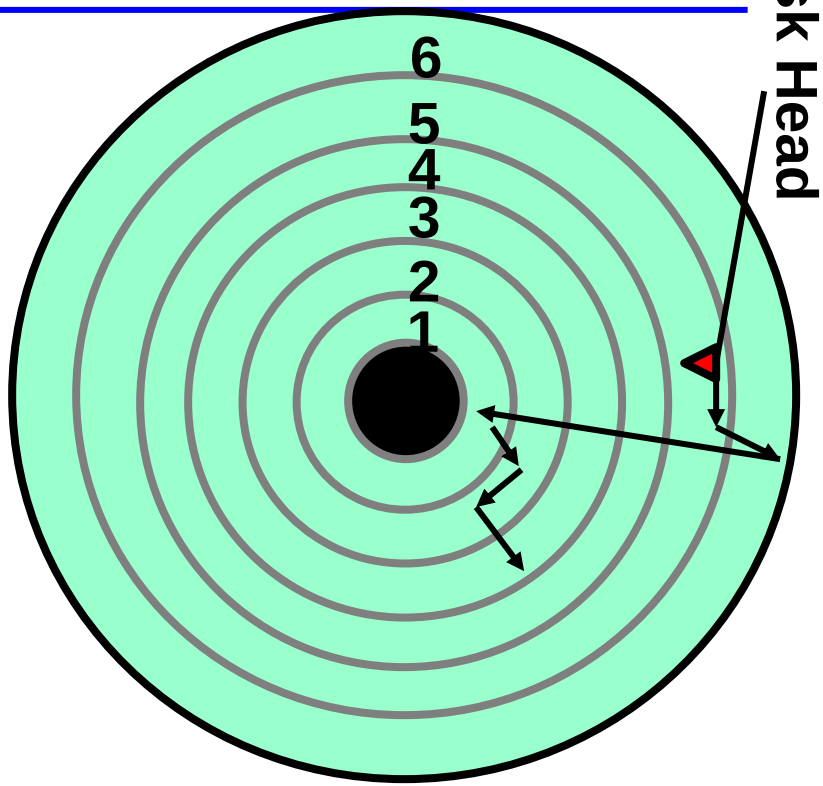
- Pros:
  - No starvation
  - Low seek

- Cons: favor middle tracks

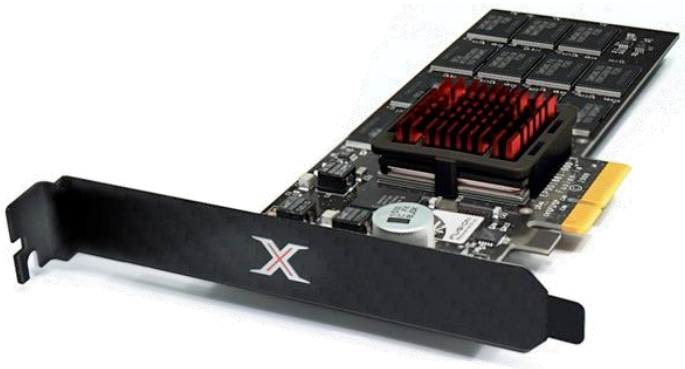


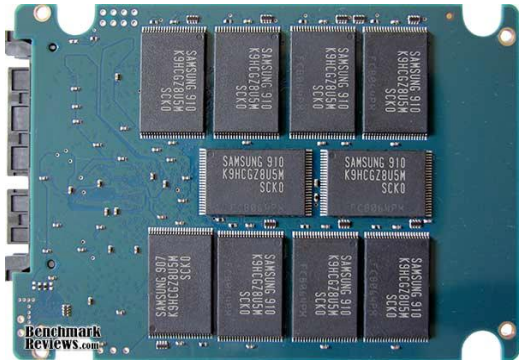
C-SCAN

- Like SCAN but only serves request in only one direction
- Example:
  - Request queue: 2, 1, 3, 6, 2, 5
  - Head only servers request on its way from center towards edge
  - Scheduling order: 5, 6, 1, 2, 2, 3
- Pros:
  - Fairer than SCAN
- Cons: longer seeks on the way back



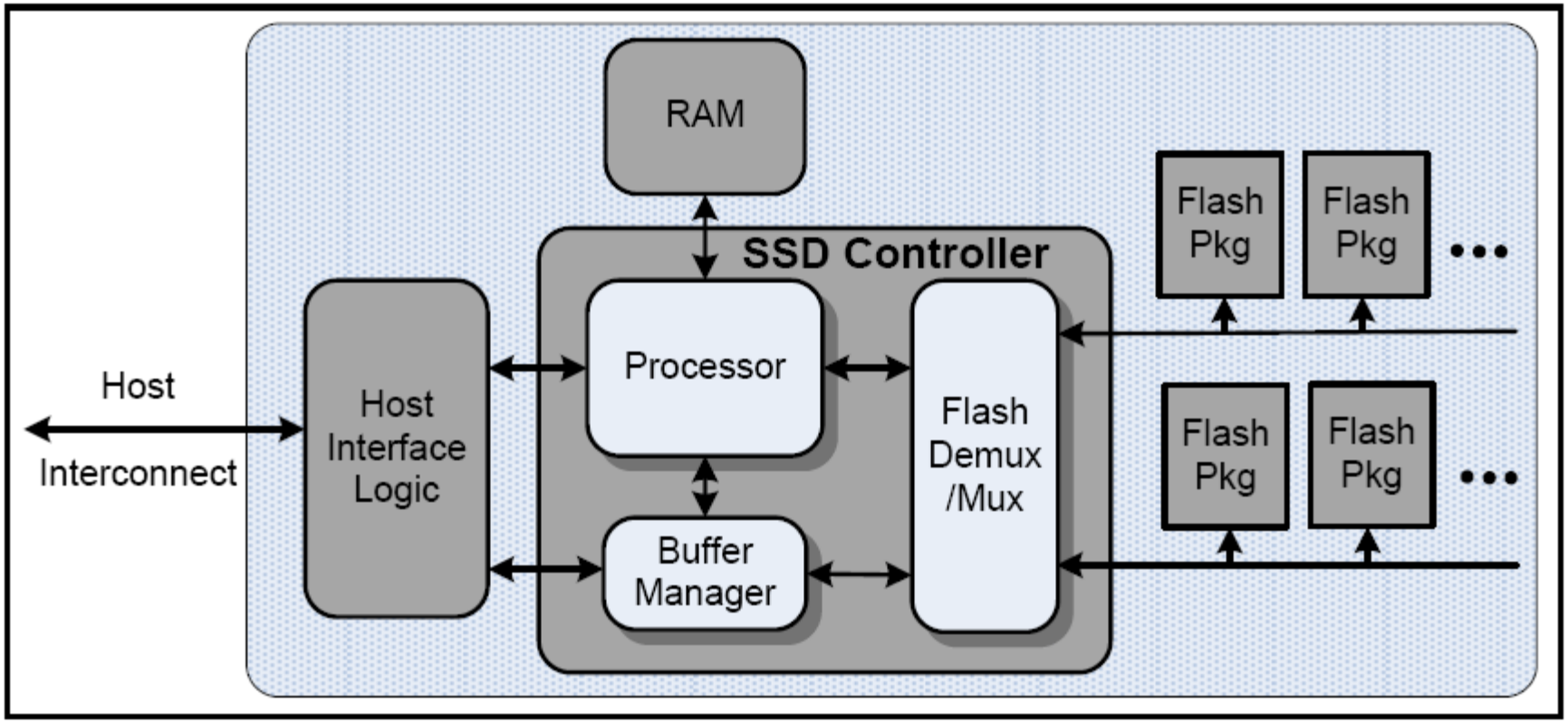
Solid State Disks (SSDs)



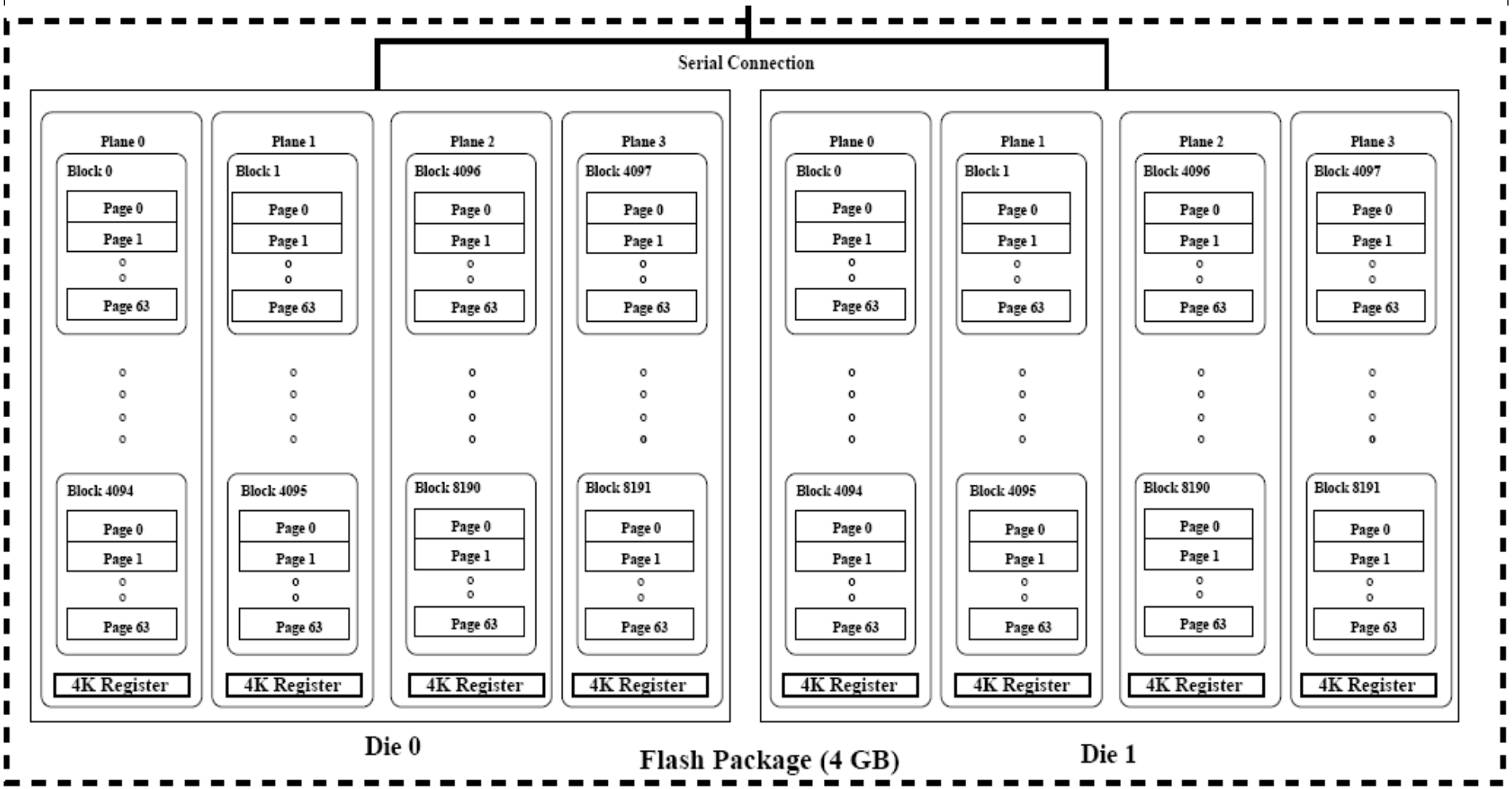


- 1995 – Replace rotating magnetic media with non-volatile memory (battery backed DRAM)
  - Since 2009, use NAND Flash: Single Level Cell (1-bit/cell), Multi-Level Cell (2-bit/cell)
- Sector addressable, but stores 4-64 “sectors” per memory page
- No moving parts (no rotate/seek motors)
  - Eliminates seek and rotational delay (0.1-0.2ms access time)
  - Very low power and lightweight

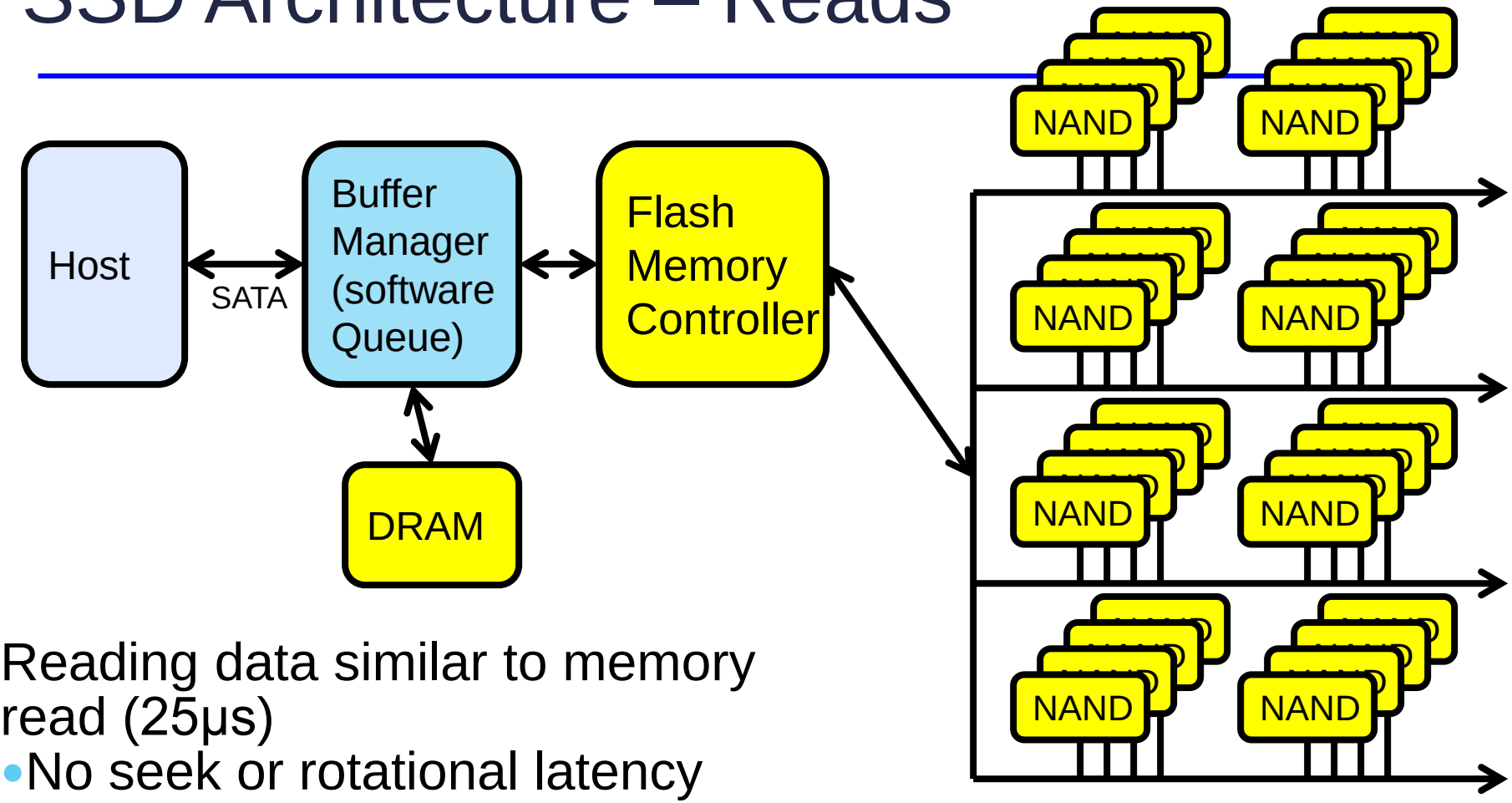
SSD Logic components



Flash Internals



SSD Architecture – Reads



Reading data similar to memory read (25µs)

- No seek or rotational latency
- Transfer time: transfer a block of bits (sector)
  - Limited by controller and disk interface (SATA: 300-600MB/s)
- Latency = Queuing Time + Controller time + Xfer Time
- Highest Bandwidth: Sequential OR Random reads

SSD Architecture – Writes

- Writing data is complex! (~200µs – 1.7ms)
- Can only write empty pages (erase takes ~1.5ms)
- Controller maintains pool of empty pages by coalescing used sectors (read, erase, write), also reserve some % of capacity

- Write and erase cycles require “high” voltage
- Damages memory cells, limits SSD lifespan
- Controller uses ECC, performs wear leveling

- Result is very workload dependent performance
- Latency = Queuing Time + Controller time (Find Free Block) + Xfer Time
- Highest BW: Seq. OR Random writes (limited by empty pages)
  - Sequential easier to implement since can write all data to same pg

Rule of thumb: writes 10x more expensive than reads, and erases 10x more expensive than writes

Storage Performance & Price

	Bandwidth (sequential R/W)	Cost/GB	Size
HDD	50-100 MB/s	$0.01-0.05/GB	2-10 TB
SSD	200-600 MB/s (SATA) 6 GB/s (PCI)	$0.1-0.5/GB	512GB-4TB
DRAM	10-16 GB/s	$0.5-1/GB	4GB-64GB

BW: SSD up to x10 than HDD, DRAM > x10 than SSD

Price: HDD x10 less than SSD, SSD x5 less than DRAM

Quiz 12.3: HDDs and SSDs

- Q1: True _ False _ The block is the smallest addressable unit on a disk
- Q2: True _ False _ An SSD has zero seek time
- Q3: True _ False _ For an HDD, the read and write latencies are similar
- Q4: True _ False _ For an SSD, the read and write latencies are similar
- Q5: Consider the following sequence of requests (2, 4, 1, 8), and assume the head position is on track 9. Then, the order in which SSTF services the requests is

  _____

SSD Summary

- Pros (vs. hard disk drives):
  - Low latency, high throughput (eliminate seek/rotational delay)
  - No moving parts:
    - Very light weight, low power, silent, very shock insensitive
  - Read at memory speeds (limited by controller and I/O bus)
- Cons
  - Expensive (3-20x disk)
    - Hybrid alternative: combine small SSD with large HDD
  - Asymmetric block write performance: read pg/erase/write pg
    - Controller garbage collection (GC) algorithms have major effect on performance
  - Limited drive lifetime
    - 1-10K writes/page for MLC NAND
    - Avg failure rate is 6 years, life expectancy is 9–11 years
- These are changing rapidly!