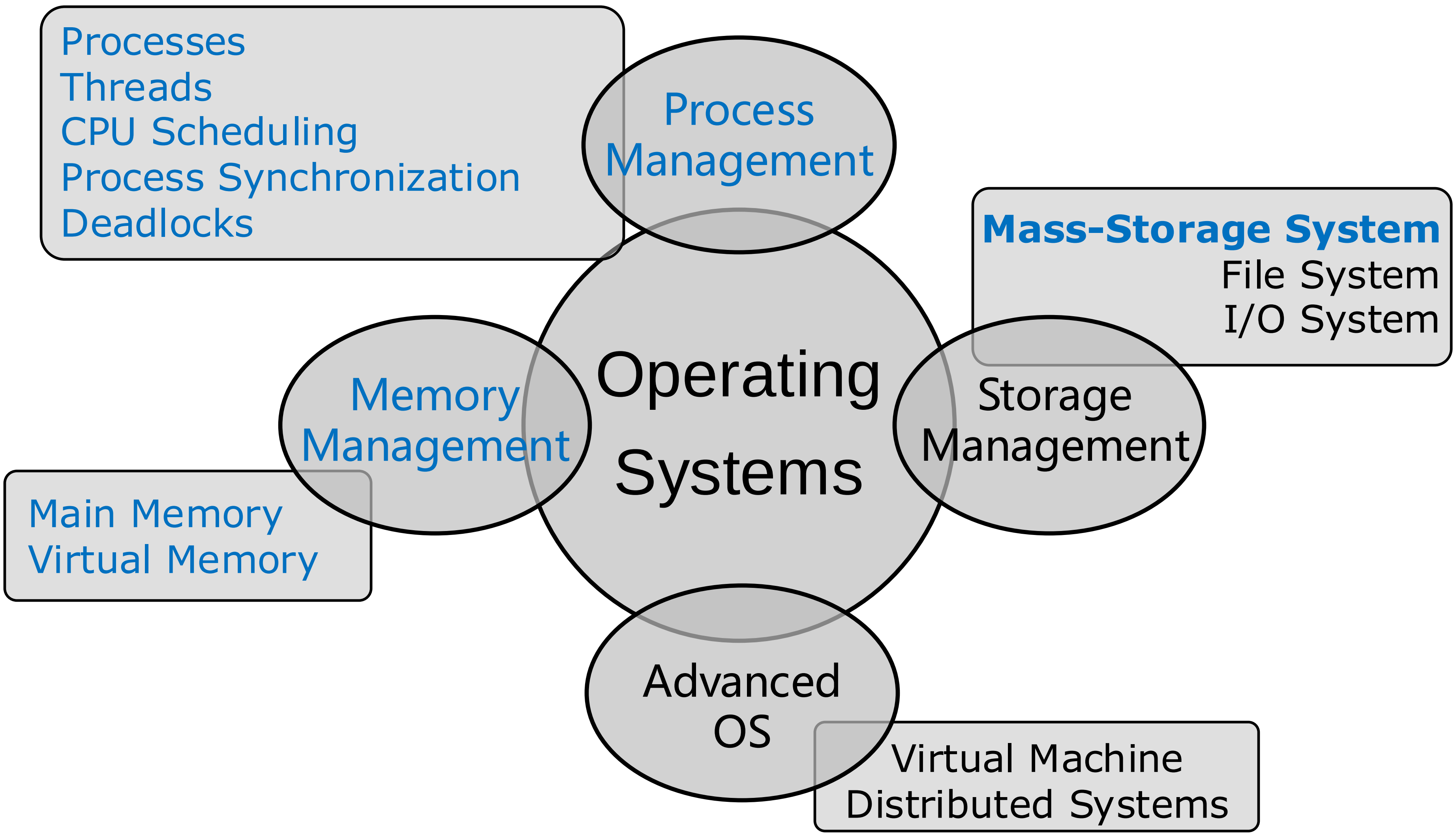
Mass-Storage Systems

Shengzhong Liu
Department of Computer Science and Engineering
Shanghai Jiao Tong University

Operating System Topics



Outline

- Overview of Mass Storage Structure
- HDD Scheduling
- NVM Scheduling
- Storage Device and Swap Space Management
- Storage Attachment
- RAID Structure

Mass Storage Overview

Overview of Mass Storage Structure

- Most secondary storage for modern computers are **hard disk drives (HDDs)** and **nonvolatile memory (NVM, 非易失性内存)** devices

- **HDDs** spin platters of magnetically-coated material under moving read-write heads
  - **Transfer rate** is rate at which data flow between drive and computer
    - 10s to 100s MB per second
  - **Positioning time (random-access time)** = seek time + rotational latency
    - **Seek time**: time to move disk arm to desired cylinder
    - **rotational latency**: time for desired sector to rotate under the disk head
    - Several miliseconds

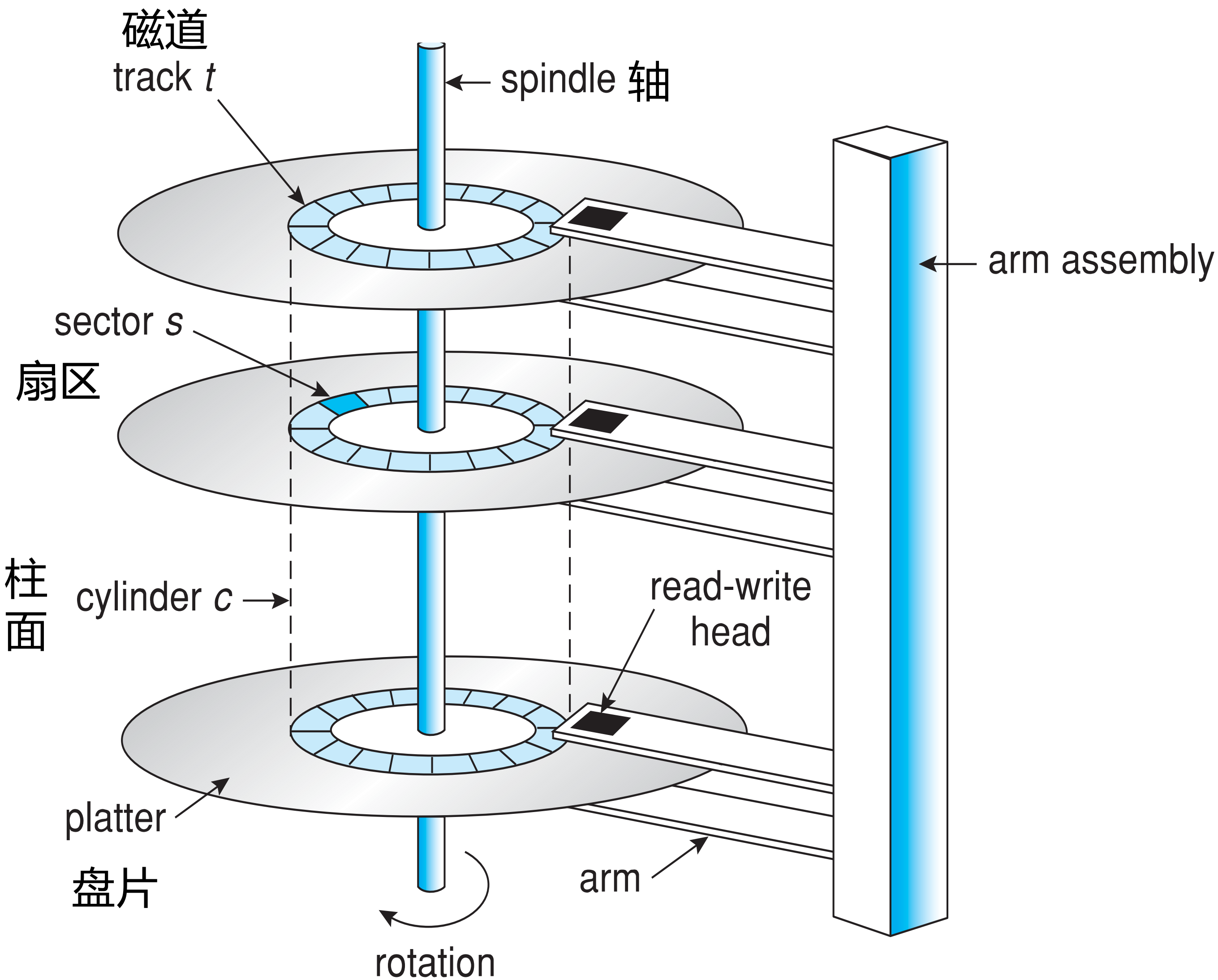
- Disks can be removable

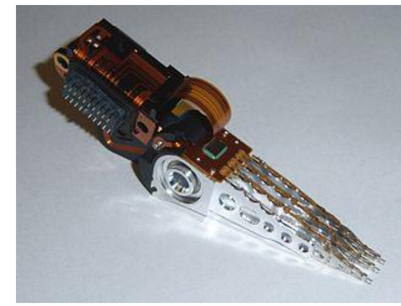
Mass Storage: (1) Hard Disk Drives

- Platters range from **.85" to 14"** (historically)
  - Commonly 3.5", 2.5", and 1.8"
- Range from **30GB to 20TB** per drive
- Performance
  - **Transfer Rate** – theoretical – 6 Gb/sec
    - Effective Transfer Rate – real – 1Gb/sec
  - **Seek time** from 3ms to 12ms – 9ms common for desktop drives
    - Average seek time measured or calculated based on 1/3 of tracks
  - **Latency** based on spindle speed
    - 1 / (RPM / 60) = 60 / RPM
    - Average latency = ½ latency

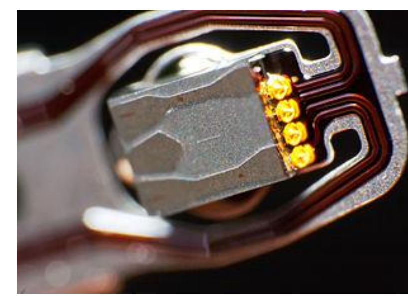
Mass Storage: (1) HDD - Seek

Mass Storage: (1) HDD - Mechanisms





Head stack

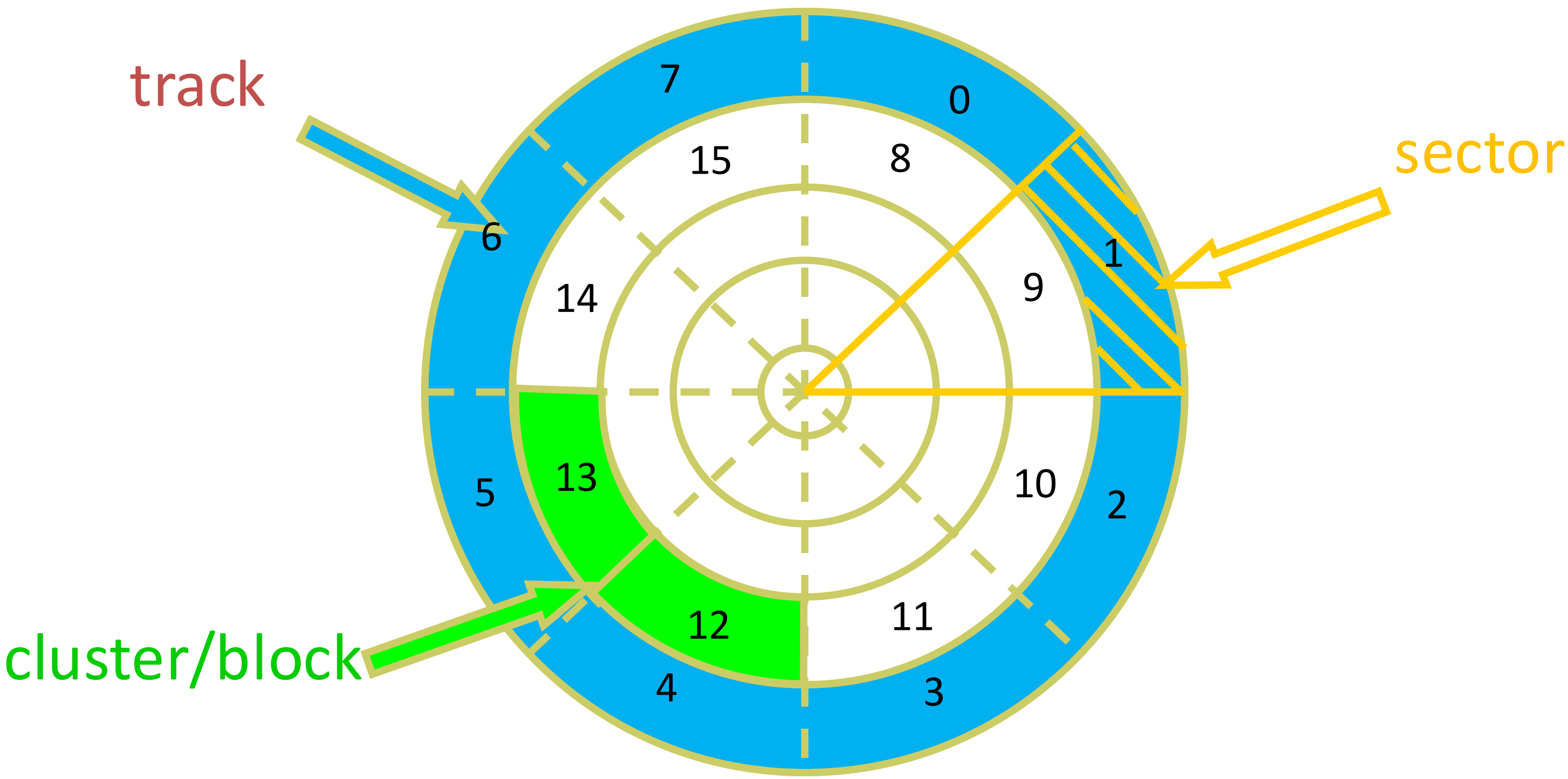


Read-write head

Mass Storage: (1) HDD – Address Mapping

- Disk drives are addressed as large 1-dimensional arrays of **logical blocks**, where the logical block is the smallest unit of transfer
  - Low-level formatting creates **logical blocks** on physical media
- The 1-dimensional array of logical blocks is mapped into the sectors of the disk sequentially
  - Sector 0 is the first sector of the first track on the outermost cylinder
  - Mapping proceeds:
    - in order through that track,
    - then the rest of the tracks in that cylinder
    - then through the rest of the cylinders from outermost to innermost
  - Logical to physical address should be easy
    - Except for bad sectors

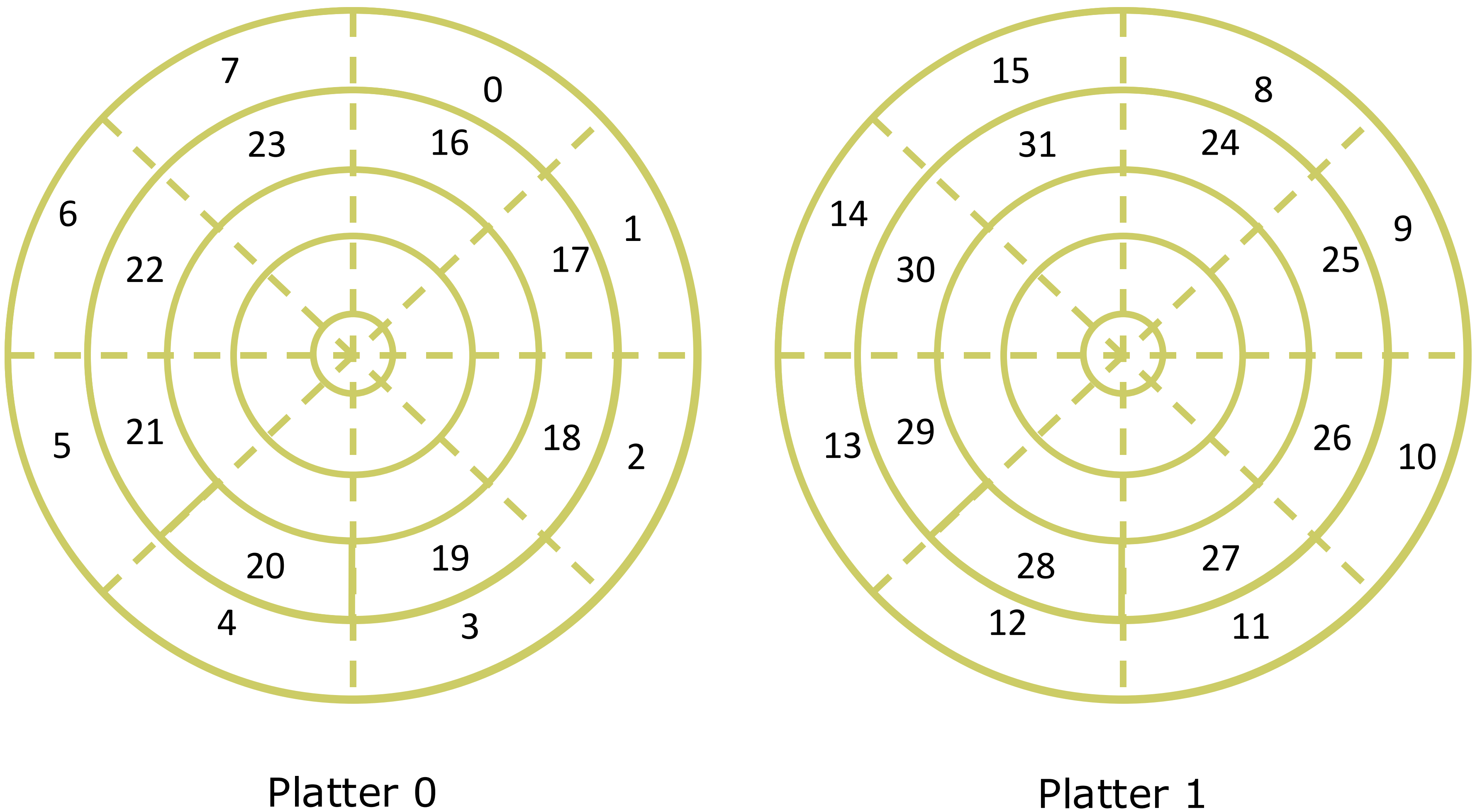
Mass Storage: (1) HDD - Disk Structure



Logical Blocks

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	……

Mass Storage: (1) HDD - Disk Structure



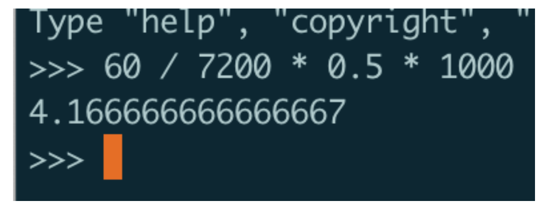
Mass Storage: (1) HDD - Performance

- **Average access time** = average seek time + average latency
  - For fastest disk 3ms + 2ms = 5ms
  - For slow disk 9ms + 5.56ms = 14.56ms
- **Average I/O time** =

  average access time +

  (amount to transfer / transfer rate) +

  controller overhead



- Example:
  - To transfer a 4KB block on a 7200 RPM disk with a 5ms average seek time, 1Gb/sec transfer rate with a .1ms controller overhead =

    5ms (**seek time**) + 4.17ms (**rotation latency**) + 0.1ms + transfer time
  - Transfer time = 4KB / 1Gb/s = 32 / (1024^2) = 0.031 ms
  - **Average I/O time for 4KB block = 9.27ms + 0.031ms = 9.301ms**

Mass Storage: (2) Nonvolatile Memory Devices

- **Nonvolatile memory devices** are types of computer memory that retain stored information even when power is removed.
  - Flash-memory-based NVM used in disk-drive-like container is called **solid-state disks (SSDs)**
  - Other forms include **USB drives** (thumb drive, flash drive), DRAM disk replacements, surface-mounted on motherboards, and main storage in devices like smartphones
- NVM vs. HDD:
  - Can be more reliable than HDDs
  - More expensive per MB
  - Maybe have a shorter life span – need careful management
  - Less capacity
  - Much faster
- Standard buses can be too slow -> connect directly to system bus (e.g. PCIe)
- No moving parts, so **no seek time or rotational latency**

SSD vs. HDD in Price

SSD Price

HDD Price

Mass Storage: (2) Nonvolatile Memory Devices

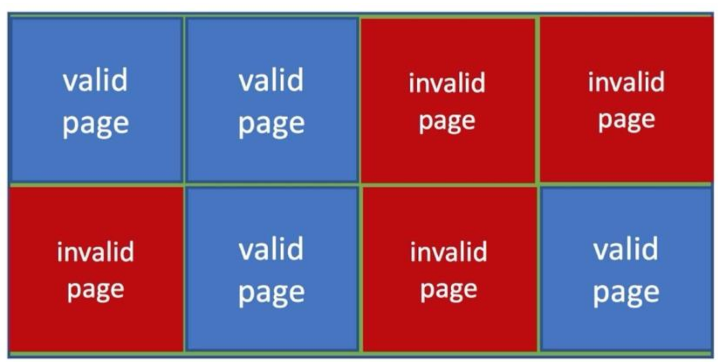
- Have characteristics that present challenges
- Read and written in “page” increments (think sector) but can’t overwrite in place
  - Must first be erased, and erases happen in larger ”block” increments
  - Can only be erased a limited number of times before worn out – ~100,000 times
  - Life span measured in **drive writes per day** (**DWPD**)
    - A 1TB NAND drive with rating of 5DWPD is expected to have 5TB per day written within warrantee period without failing

NAND Flash Controller Algorithms

- With no overwrite, pages end up with mix of valid and invalid data
- To track which logical blocks are valid, controller maintains **flash translation layer (FTL)** table

- **Garbage collection**:
  - Copy good data to other locations, freeing up blocks that could be erased
  - Allocates overprovisioning to provide working space for GC
  - For example, preserve 20% of pages for writing data to during GC

- **Wear leveling**:
  - Each cell has lifespan, need to write equally to all cells
  - Avoid frequently erased blocks making the device lifespan shorter.



Mass Storage: (3) Volatile Memory

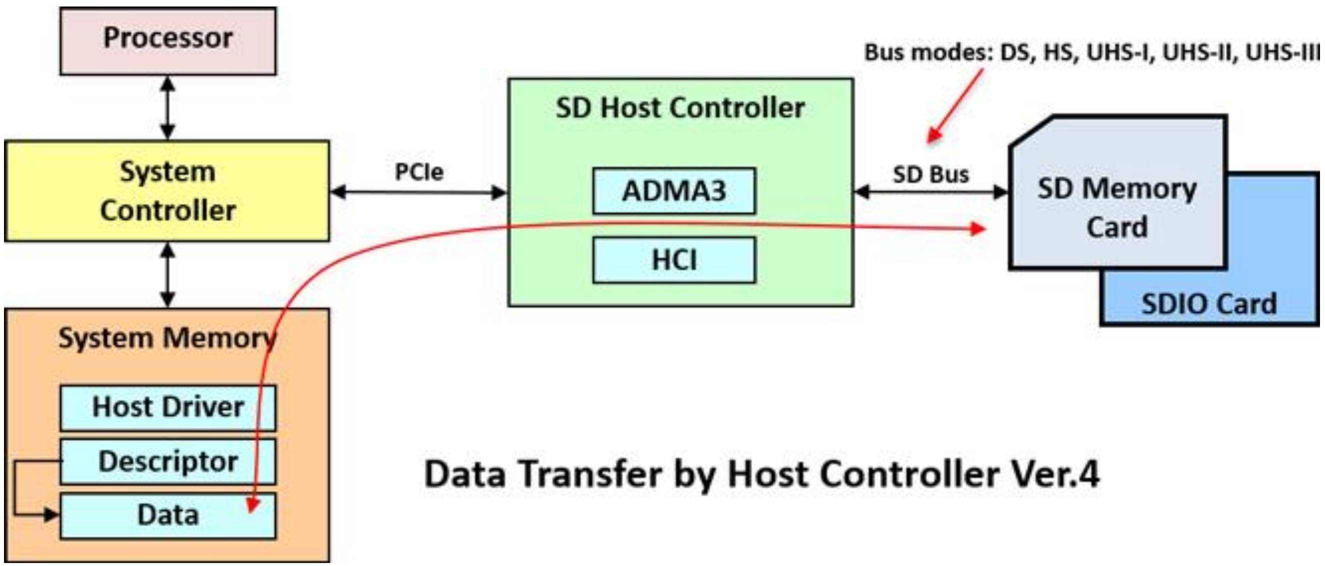
- DRAM frequently used as mass-storage device
  - Not technically secondary storage because volatile, but can have file systems, be used like very fast secondary storage
- **RAM drives** present as raw block devices, commonly file system formatted
- RAM drives allow the user to place data in memory for temporary safekeeping using standard file operations
  - Found in all major operating systems
    - Linux `/dev/ram`
    - macOS `diskutil` to create them
    - Linux `/tmp` of file system type `tmpfs`
- Useful as high-speed temporary storage
  - I/O operations to RAM drives are the fastest way to create, read, write, and delete files and their contents

Disk Attachment

- Host-attached storage accessed through I/O ports talking to **I/O busses**
- Several busses available, including
  - advanced technology attachment (ATA)
  - **serial ATA (SATA), most common**
  - eSATA
  - serial attached SCSI (SAS)
  - universal serial bus (USB)
  - fiber channel (FC)
- Because NVM much faster than HDD, new fast interface for NVM called **NVM express (NVMe)**, connecting directly to PCI bus

Disk Attachment

- Data transfers on a bus carried out by special electronic processors called **controllers** (or **host-bus adapters**, **HBAs**)
  - **Host controller** on the computer end, **device controller** on the device end
    - Computer places command on host controller, using memory-mapped I/O ports
    - Host controller sends messages to device controller
  - Data transferred via DMA between device and computer DRAM
    - DMA: Direct memory access



Data Transfer by Host Controller Ver.4

Disk Scheduling

Disk Scheduling

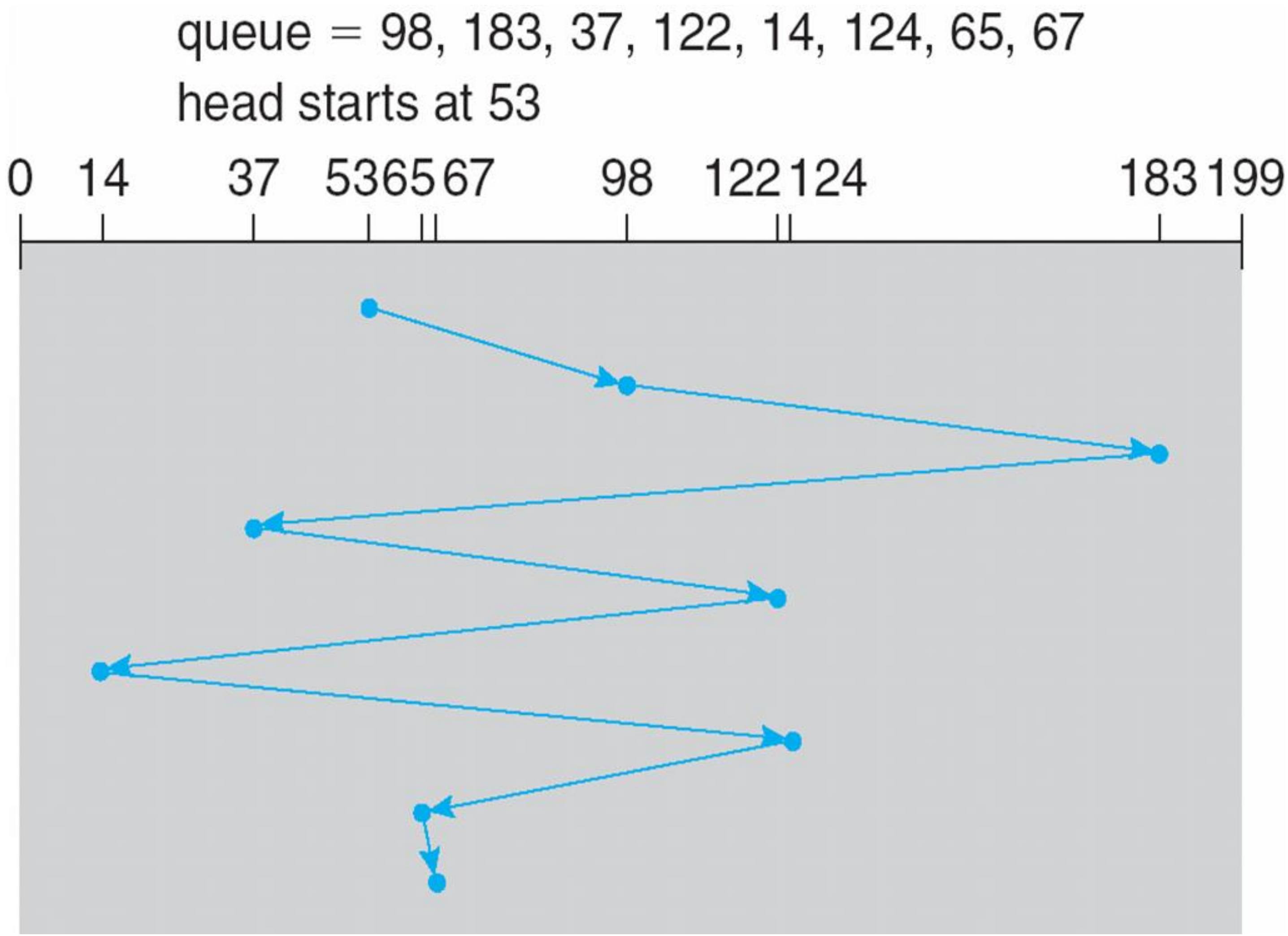
- There are many sources of disk I/O request
  - OS, System processes, Users processes
  - OS maintains queue of requests, per disk or device
- Idle disk can work on I/O request, busy disk means work must be queued
  - Optimization algorithms only make sense when a queue exists
- The operating system is responsible for using hardware efficiently —
  - Having a fast access time and disk bandwidth
  - **Objective: Minimize seek time ≈ seek distance**
  - What about rotational latency?
    - Difficult for OS to calculate
- Several algorithms exist to schedule the servicing of disk I/O requests
- We illustrate scheduling algorithms with a request queue (0-199)

98, 183, 37, 122, 14, 124, 65, 67 Head pointer 53

Disk-Scheduling Algorithms

- First-Come, First-Served (FCFS) Scheduling
- Shortest Seek Time First (SSTF) Scheduling
- SCAN Scheduling
- C-SCAN Scheduling
- LOOK/C-LOOK Scheduling

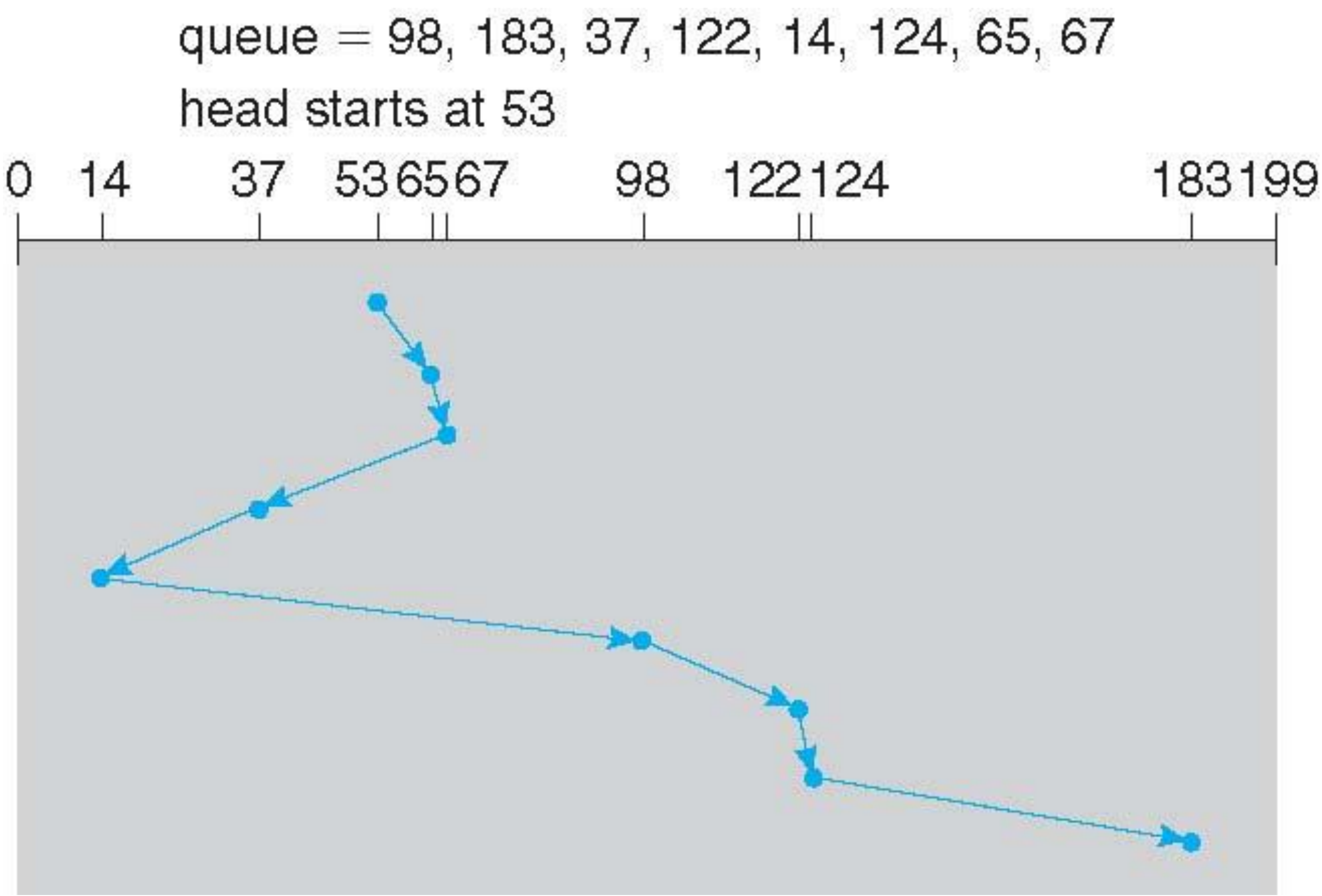
Disk Scheduling: (1) FCFS



Total head movement: **640 cylinders**

Disk Scheduling: (2) SSTF

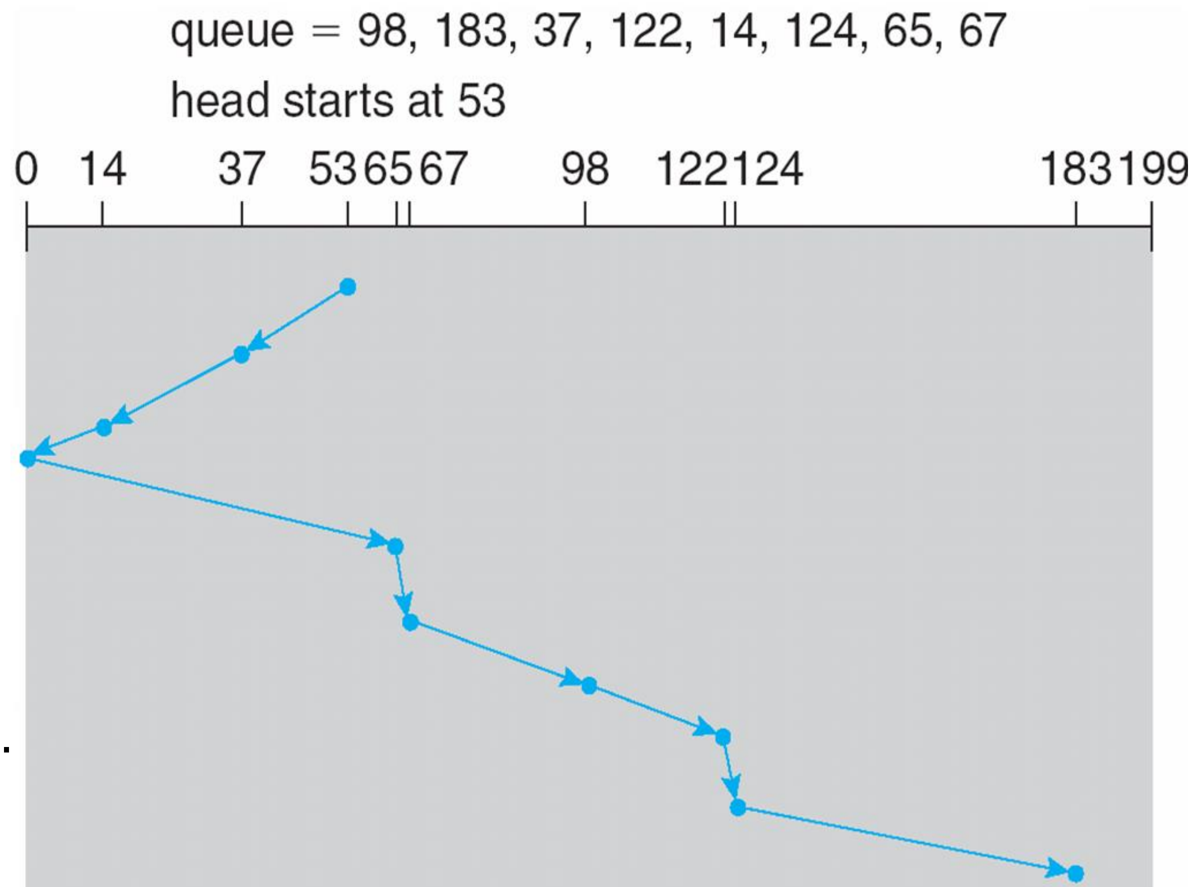
- **Shortest Seek Time First (SSTF)** selects the request with the minimum seek time from the current head position



Total head movement: **236 cylinders**

Disk Scheduling: (3) SCAN

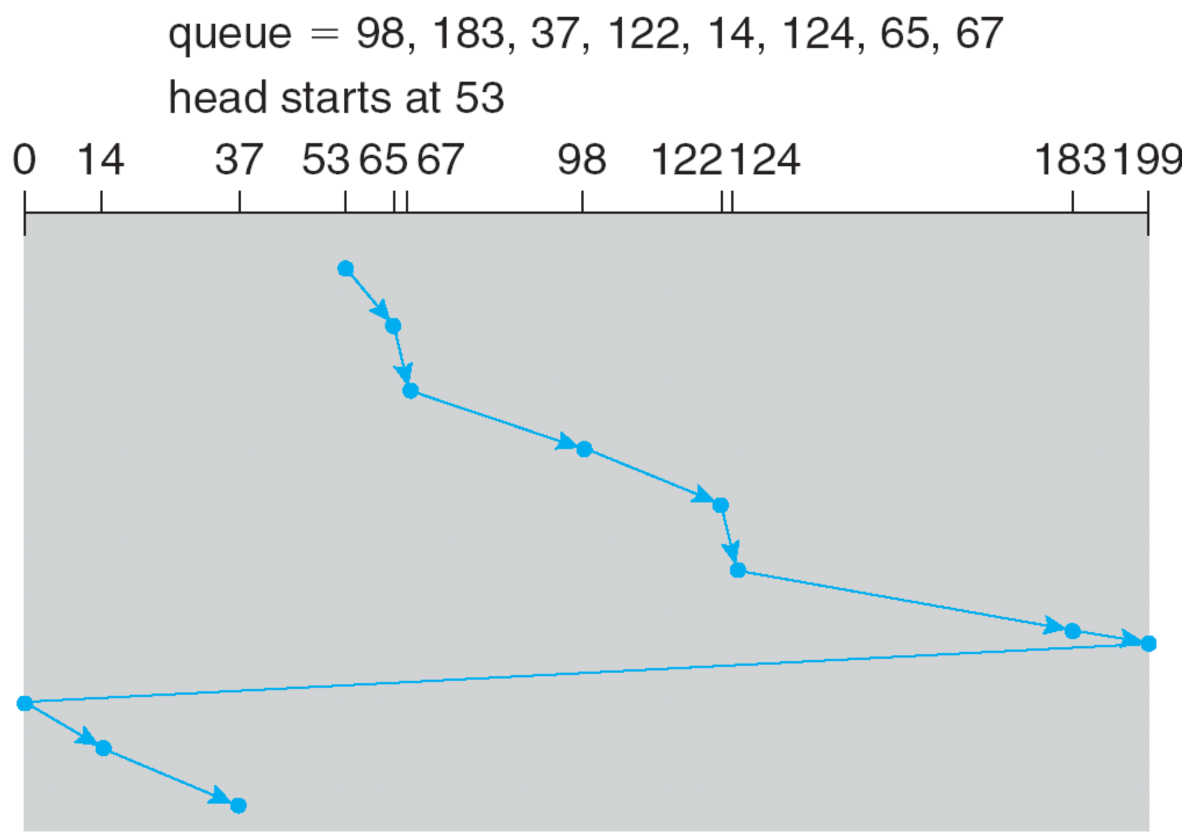
- **SCAN algorithm** Sometimes called the **elevator algorithm**
  - The disk arm starts at one end of the disk
  - Moves toward the other end, servicing requests until it gets to the other end of the disk
  - Then the head movement is reversed and servicing continues.



Total head movement: **236 cylinders**

Disk Scheduling: (4) C-SCAN

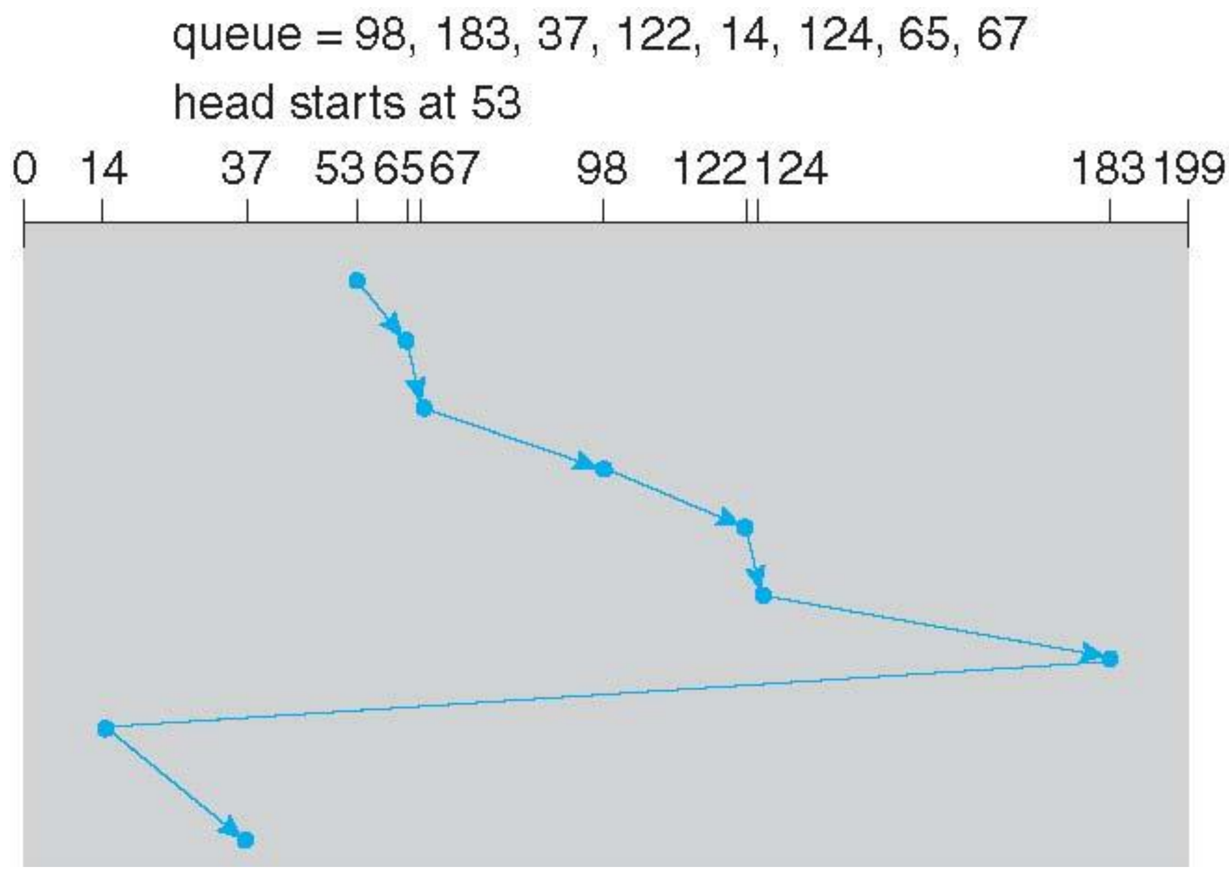
- Provides a more uniform wait time than SCAN
- The head moves from one end of the disk to the other, servicing requests as it goes
  - When it reaches the other end, it immediately returns to the beginning of the disk, without servicing any requests on the return trip
- Treats the cylinders as a circular list that wraps around from the last cylinder to the first one



Total head movement: **382 cylinders**

Disk Scheduling: (5) LOOK/C-LOOK

- LOOK a version of SCAN, C-LOOK a version of C-SCAN
- Arm only goes as far as the last request in each direction, then reverses direction immediately
- Do not go all the way to the end of the disk



Total head movement: **322 cylinders**

Selecting a Disk-Scheduling Algorithm

- SSTF is common and has a natural appeal
- SCAN and C-SCAN perform better for systems with a heavy disk load
  - Less starvation, but still possible
- LOOK and C-LOOK have a little improvement over SCAN and C-SCAN

NVM Scheduling

- No disk heads or rotational latency but still room for optimization
- In RHEL 7 **NOOP** (no scheduling) is used but adjacent LBA requests are combined
  - NVM best at random I/O, HDD better at sequential
  - Throughput can be similar
  - **Input/Output operations per second** (**IOPS**) much higher with NVM (hundreds of thousands vs hundreds)
  - But **write amplification** (once write, causing garbage collection and many read/writes) can reduce the performance advantage

Storage and Swap Management

Storage Device Management

- **Low-level formatting**, or **physical formatting** — Dividing a disk into sectors that the disk controller can read and write
  - Each sector can hold header information, plus data, plus error correction code (ECC)
  - Usually **512 bytes** of data but can be selectable
- To use a disk to hold files, the operating system still needs to record its own data structures on the disk
  - **Partition** the disk into one or more groups of cylinders, each treated as a logical disk
  - **Logical formatting (分区)** or "making a file system"
  - To increase efficiency most file systems group blocks into **clusters**
    - Disk I/O done in blocks
    - File I/O done in clusters

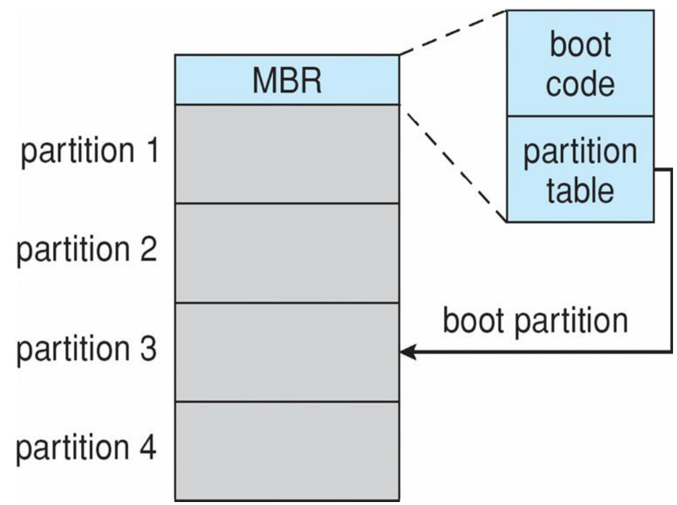
Storage Device Management (cont.)

- **Root partition** contains the OS, other partitions can hold other OS types, other file systems, or be raw
  - **Mounted** (**挂载**) at boot time
  - Other partitions can mount automatically or manually
- At mount time, file system consistency checked
  - Is all metadata correct?
    - If not, fix it, try again
    - If yes, add to mount table, allow access
- Boot block can point to boot volume or boot loader set of blocks that contain enough code to know how to load the kernel from the file system
  - Or a boot management program for multi-OS booting

Device Storage Management (Cont.)

- **Raw I/O**: Raw disk access for apps that want to do their own block management, keep OS out of the way (databases for example)

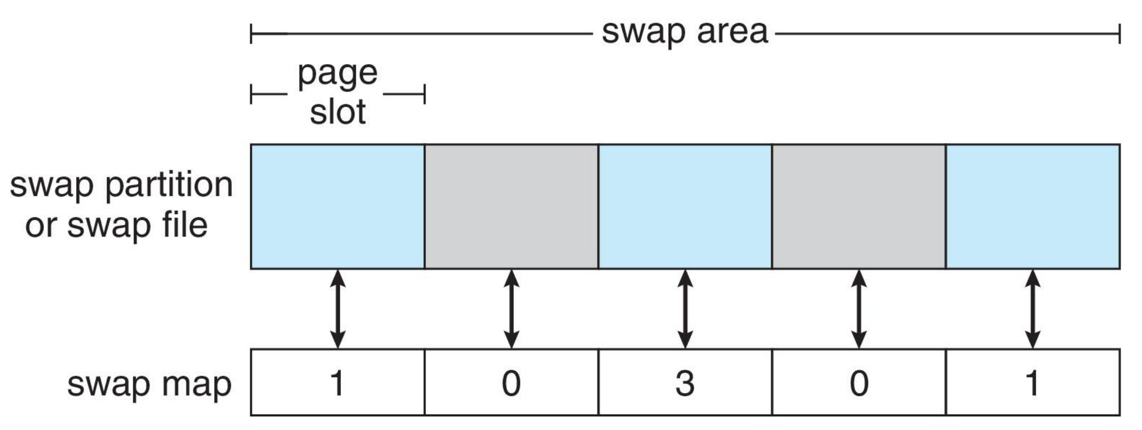
- Boot block initializes system
  - The bootstrap is stored in ROM, firmware
  - **Bootstrap loader** program stored in boot blocks of boot partition



Booting from secondary storage in Windows

Swap-Space Management

- **Swap**: Used for moving entire processes (swapping), or pages (paging), from DRAM to secondary storage when DRAM not large enough for all processes
- Operating system provides **swap space management**
  - Secondary storage slower than DRAM, so important to optimize performance
  - Usually multiple swap spaces possible – decreasing I/O load on any given device
    - Best to have dedicated devices
  - Can be in raw partition or a file within a file system (for convenience of adding)
- Example: Data structures for swapping on Linux systems:



Storage Attachment

Storage Attachment

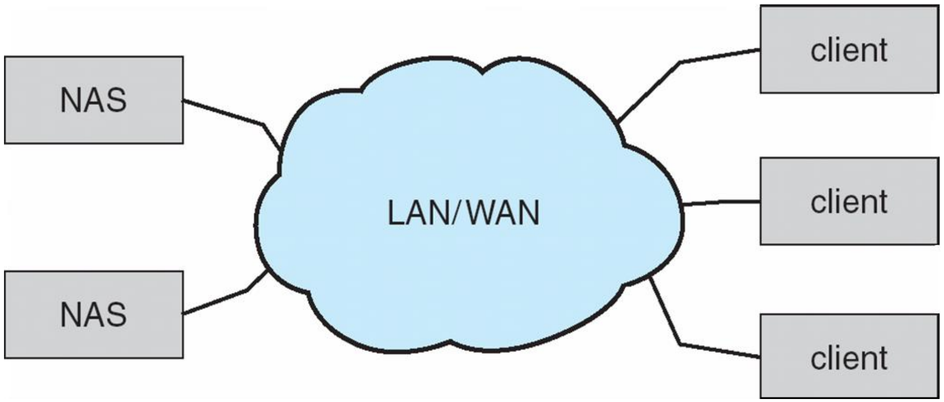
- Computers access storage in three ways
  - host-attached
  - network-attached
  - cloud

Storage Attachment: (1) Host-Attached

- Host attached access through local I/O ports, using one of several technologies
  - The most common is SATA, and a typical system has one or a few SATA ports.
  - To attach many devices, use storage busses such as USB, firewire, thunderbolt
  - High-end systems use **fiber channel (FC)**
    - High-speed serial architecture using fiber or copper cables (铜缆)
    - Multiple hosts and storage devices can connect to the FC fabric

Storage Attachment: (2) Network-Attached

- Network-attached storage (**NAS**) is storage made available over a network rather than over a local connection (such as a bus)
  - Remotely attaching to file systems
  - NFS and CIFS are common protocols
- Implemented via remote procedure calls (RPCs) between host and storage over typically TCP or UDP on IP network
- **iSCSI** protocol uses IP network to carry the SCSI protocol
  - Latest network-attached storage protocol
  - Networks, rather than SCSI cables used as the interconnects between hosts and their storage



Storage Attachment: (3) Cloud Storage

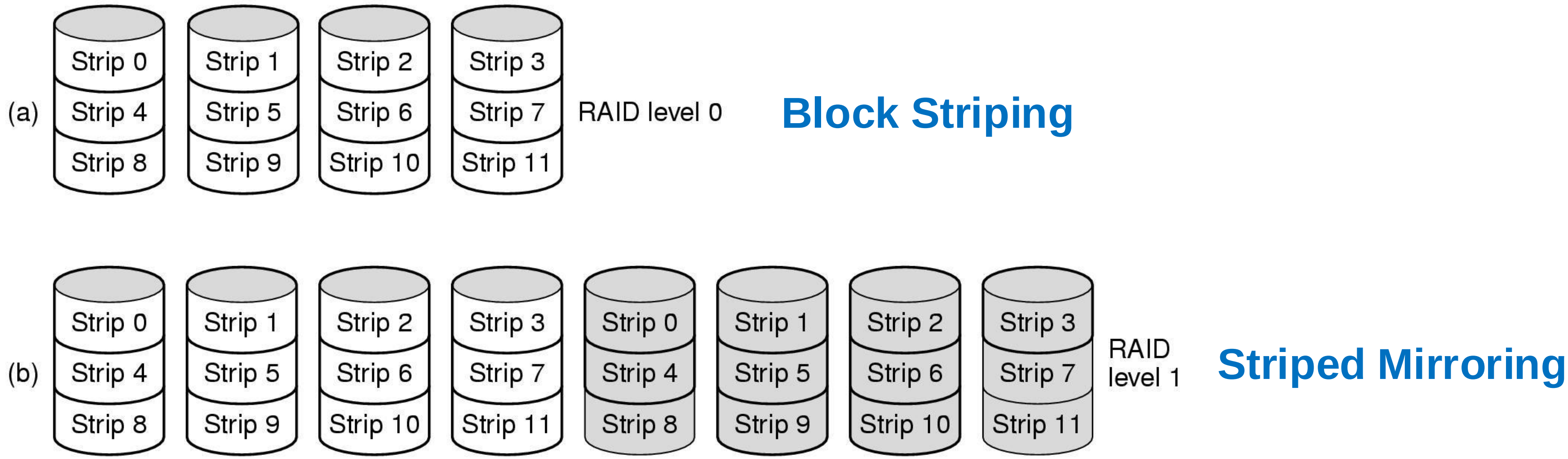
- Similar to NAS, provides access to storage across a network
  - Unlike NAS, accessed **over the Internet or a WAN** to remote data center
- NAS presented as just another file system, while cloud storage is API based, with programs using the APIs to provide access
  - Examples include Dropbox, Microsoft OneDrive, Apple iCloud, and SJTU Jbox
  - Use APIs because of latency and failure scenarios (NAS protocols wouldn't work well)

RAID Structure

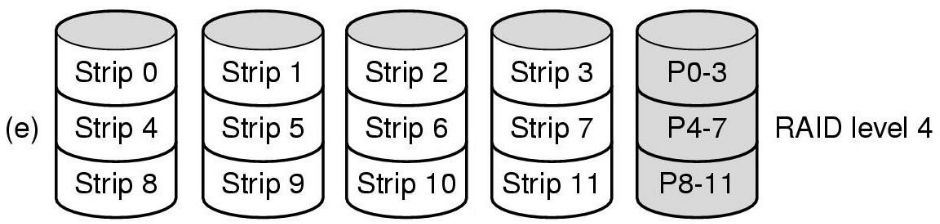
RAID Structure

- Redundant Arrays of Independent Disks (RAIDs)
- **Reliability**: multiple disk drives provide reliability via **redundancy**
  - **Mirroring**
    - duplicate every disk
  - **Parity bit**
- **Parallel access** to multiple disks improves performance
  - **Bit-level striping**
    - split the bits of each byte across multiple disks
  - **Block-level striping**
    - blocks of a file are striped across multiple disks
- RAID is arranged into seven or more different levels

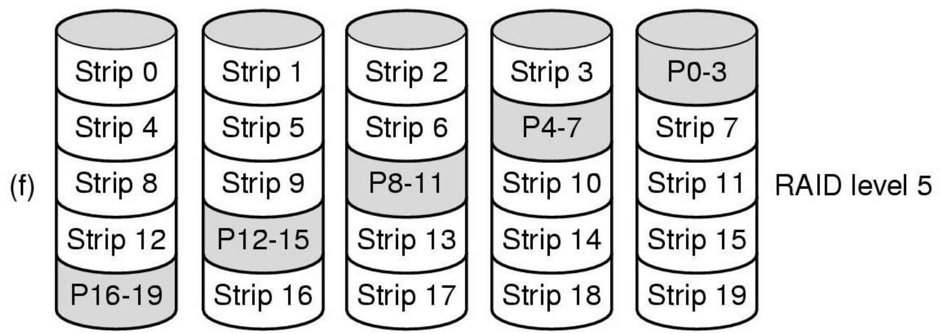
RAID Levels (Cont.)



RAID Levels (Cont.)



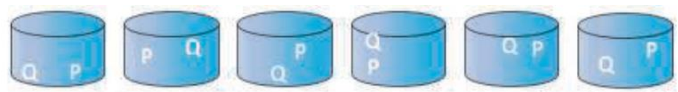
block-interleaved parity



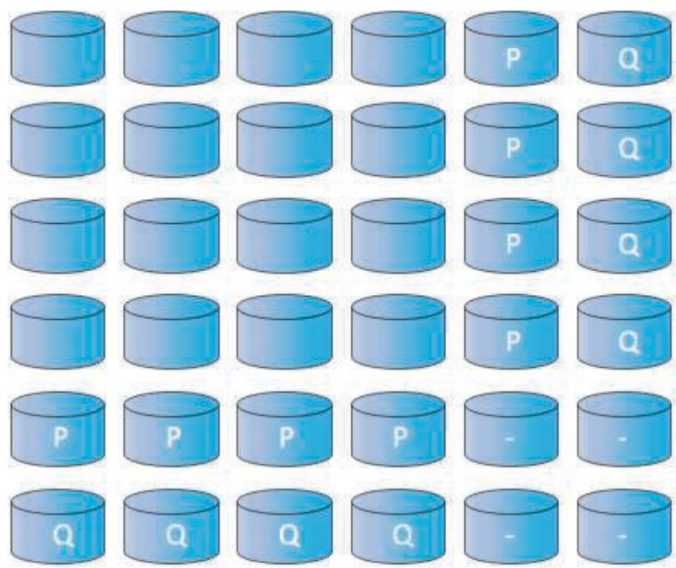
block-interleaved distributed parity

RAID Levels (Cont.)

- RAID 6: **P + Q redundancy**
  - Reed-Solomon codes
  - 2 bits of redundant data are stored for every 4 bits of data
  - can tolerate two disk failures



(e) RAID 6: P + Q redundancy.



(f) Multidimensional RAID 6.

Summary

- **Hard disk drives (HDD)** and **nonvolatile memory (NVM)** devices are the major secondary storage I/O units on most Computers
- Attached to a computer system through (1) local IO ports on the host computer, (2) directly connected to the motherboards, (3) communication network or storage network connection
- Disk scheduling algorithms improve the effective bandwidth, average response time, and variance in response time.
- Secondary storage devices are frequently made redundant via RAID algorithms for reliability and efficiency

Homework

- Reading
  - Chapter 11

- Exercise
  - Check Canvas for HW3 release, due on **Mar. 28, 23:59**!