Mass-Storage Systems

Fan Wu

Department of Computer Science and Engineering

Shanghai Jiao Tong University

Spring 2022

The First Commercial Disk Drive

1956
IBM RAMDAC computer included the IBM Model 350 disk storage system

5M (7 bit) characters
50 x 24 inch platters
Access time = < 1 second

Magnetic Tape

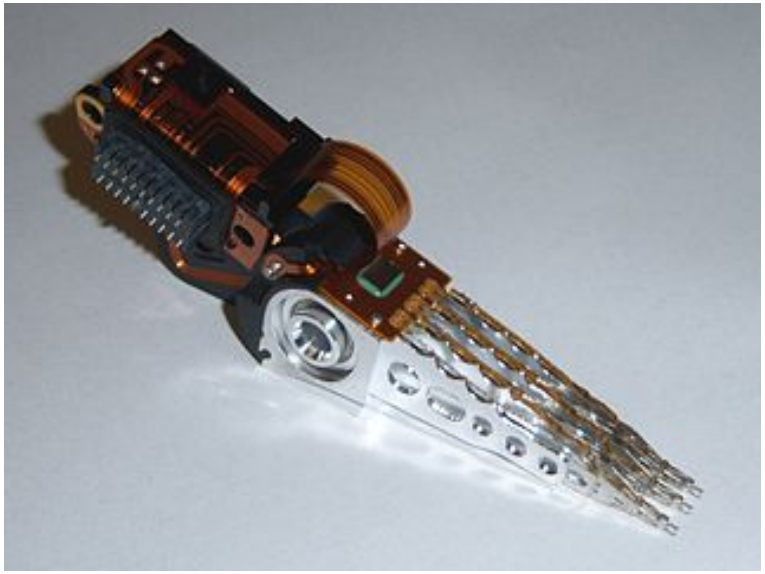
- Was early secondary-storage medium
  - Evolved from open spools to cartridges
- Relatively permanent and holds large quantities of data
- Access time slow
- Random access ~1000 times slower than disk
- Mainly used for backup, storage of infrequently-used data, transfer medium between systems
- Kept in spool and wound or rewound past read-write head
- Once data under head, transfer rates comparable to disk
  - 140MB/sec and greater
- 200GB to 185TB typical storage
- Common technologies are LTO-{3,4,5} and T10000

Hard Disk Drive

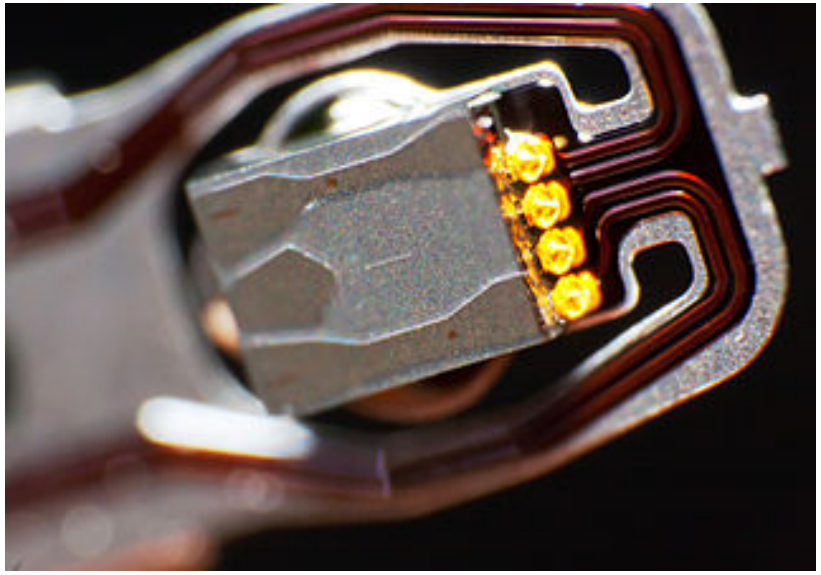
SATA HDD

Different Sized Disk Drives

Read-Write Arm and Head

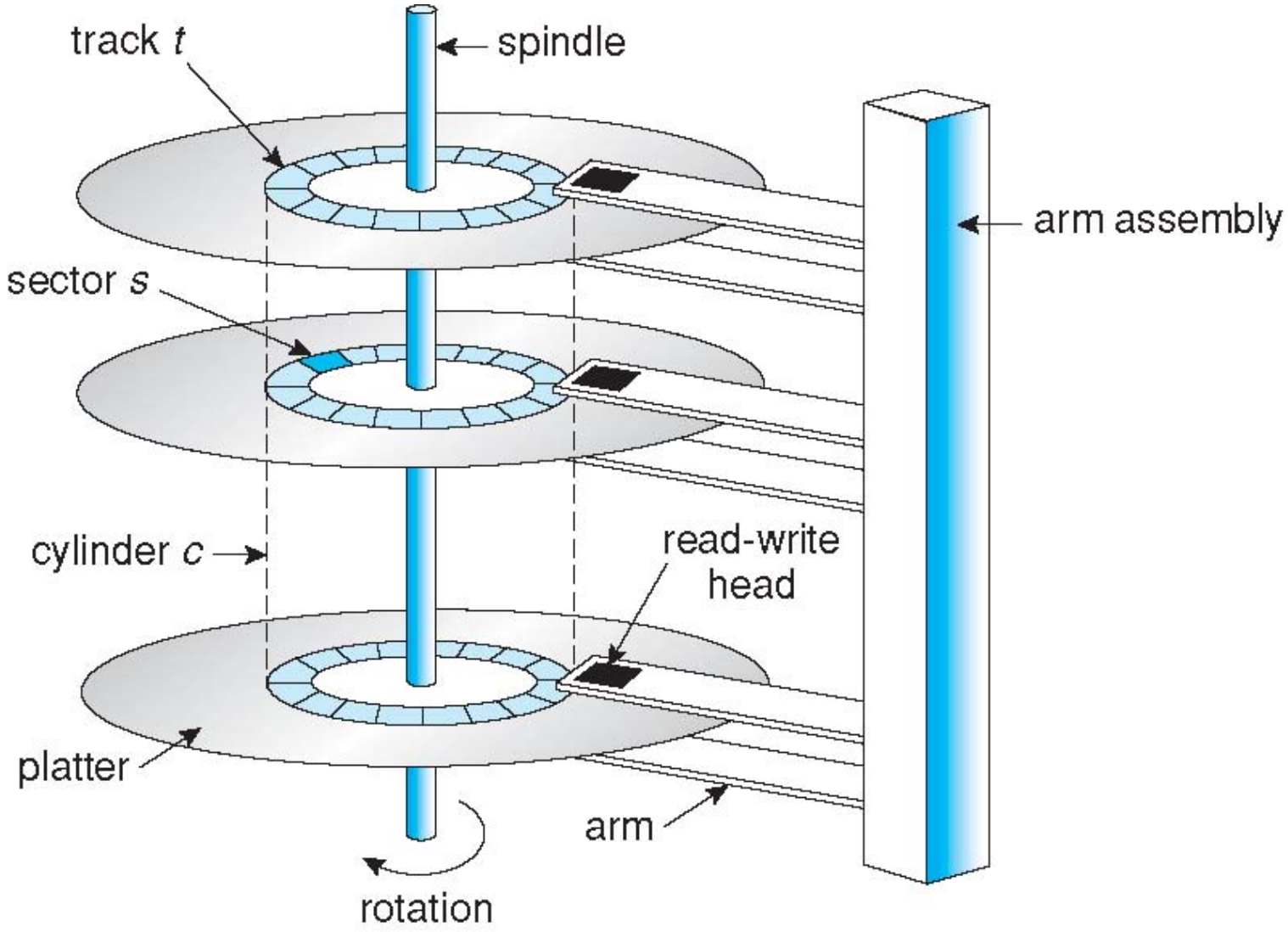


Head stack

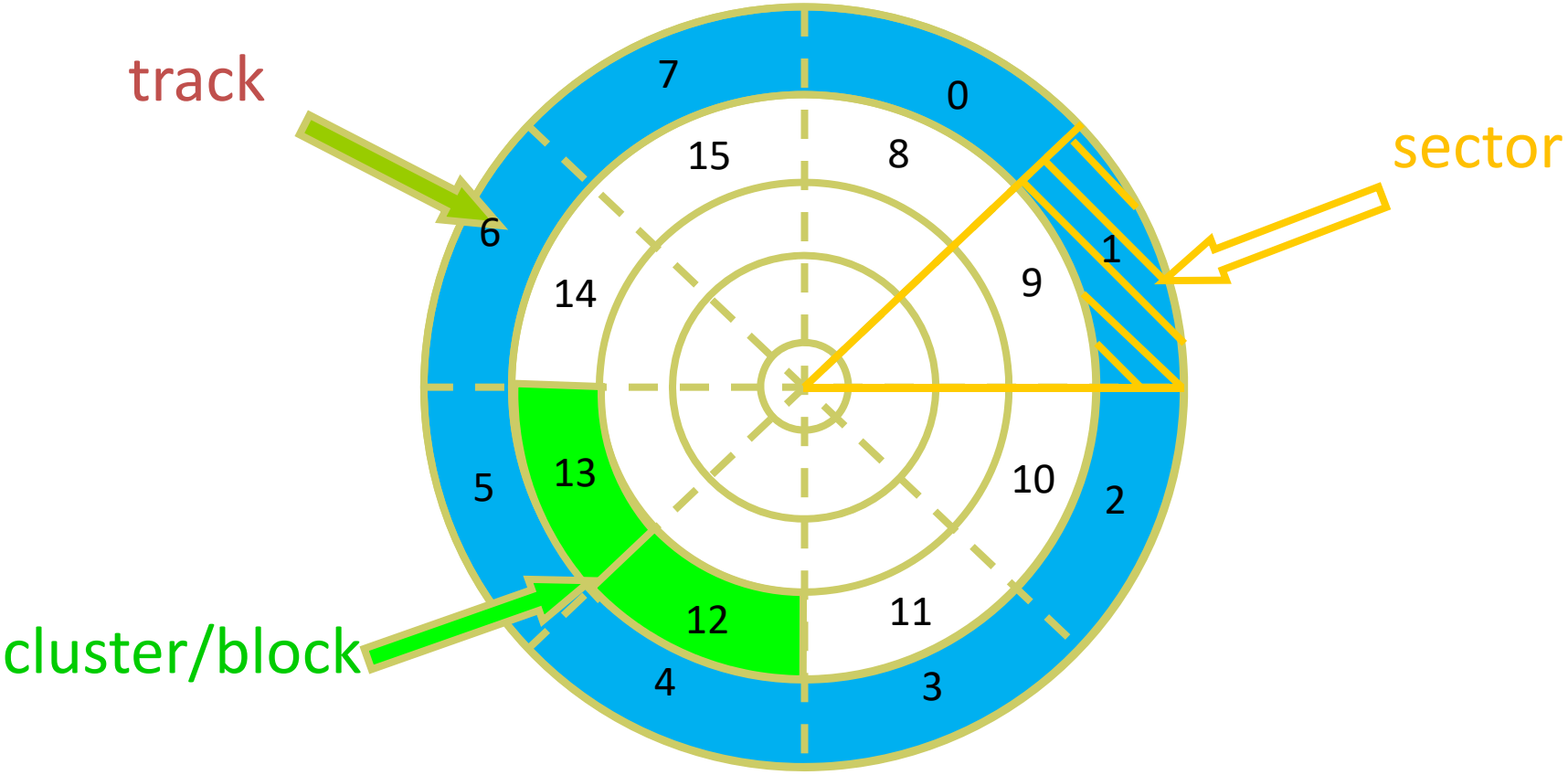


Read-write head

Moving-head Disk Mechanism

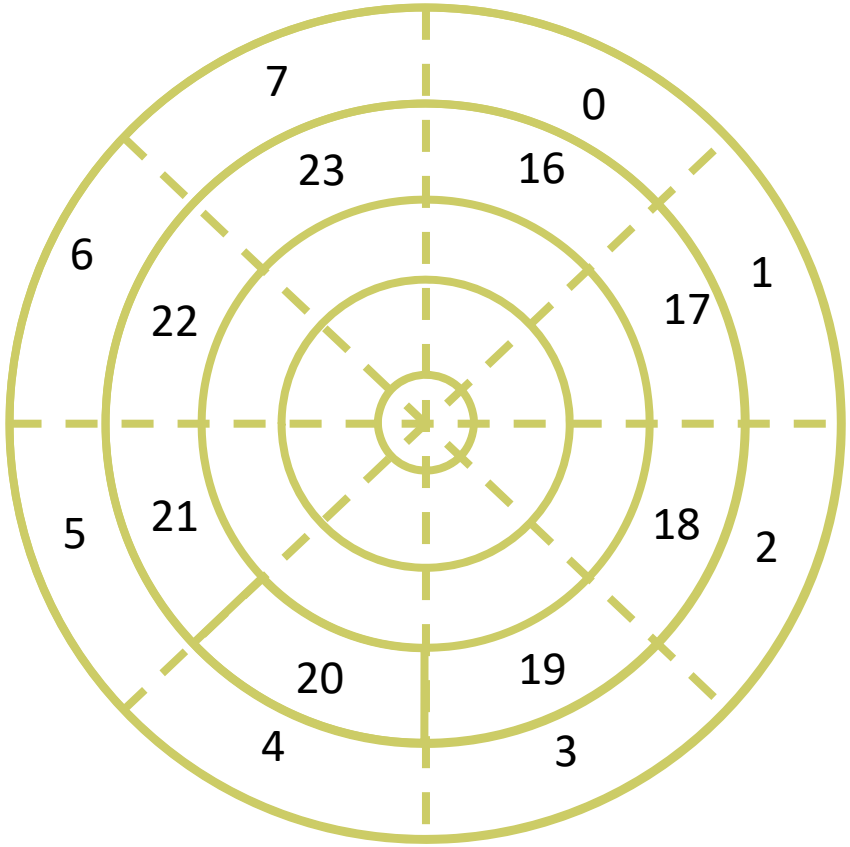


Disk Structure

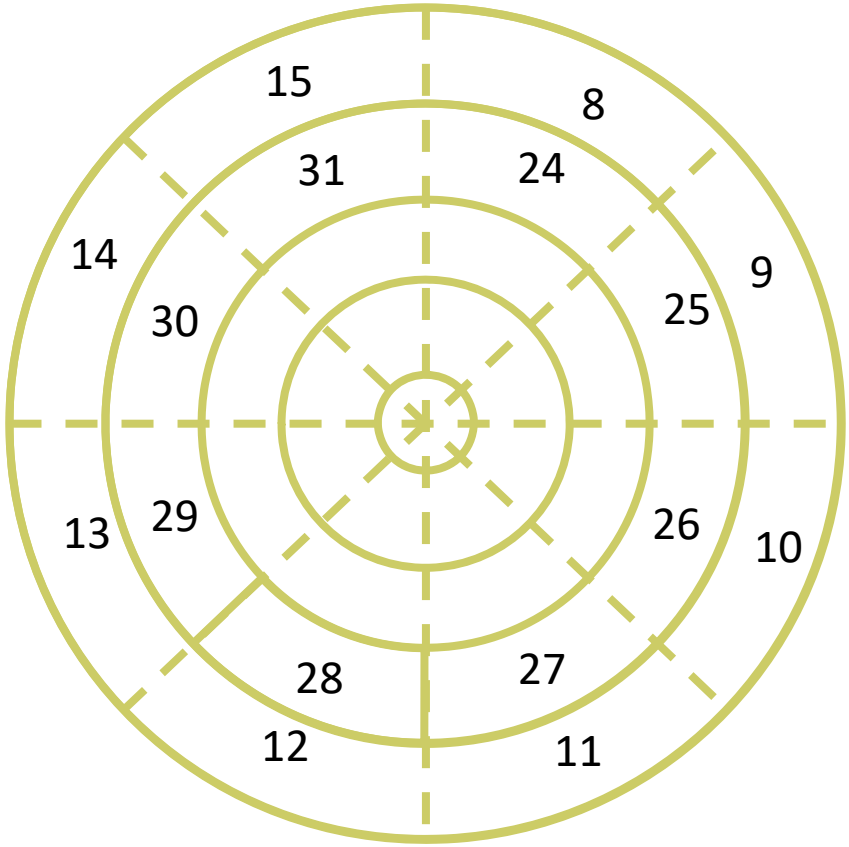


0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	.....

Disk Structure



Platter 0



Platter 1

Overview of Mass Storage Structure

- **Magnetic disks** provide bulk of secondary storage of modern computers
  - Drives rotate at 60 to 250 times per second
  - **Transfer rate** is rate at which data flows between drive and computer
  - **Positioning time** (**random-access time**) is time to move disk arm to desired cylinder (**seek time**) and time for desired sector to rotate under the disk head (**rotational latency**)

定位时间 = 寻道时间 + 旋转延迟 (seek time) (rotational latency)

1° 找到 cylinder

2° 找到 sector.

Disk Scheduling

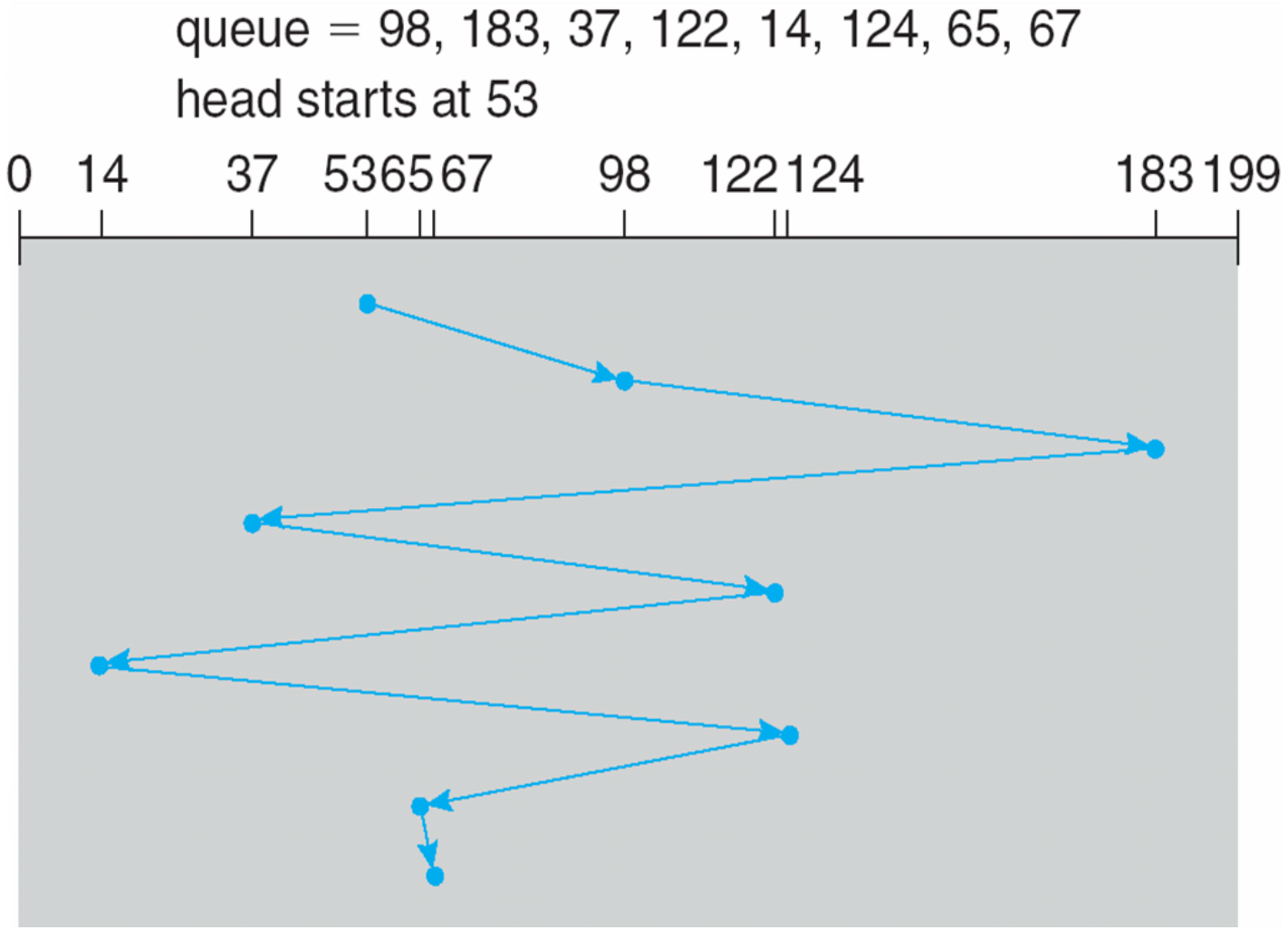
- There are many sources of disk I/O request
  - OS, System processes, Users processes
- OS maintains queue of requests, per disk or device
- Idle disk can immediately work on I/O request, busy disk means work must be queued
  - Optimization algorithms only make sense when a queue exists
- The operating system is responsible for using hardware efficiently — for the disk drives, this means having a fast access time and disk bandwidth
  - 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

FCFS Scheduling



Total head movement: 640 cylinders

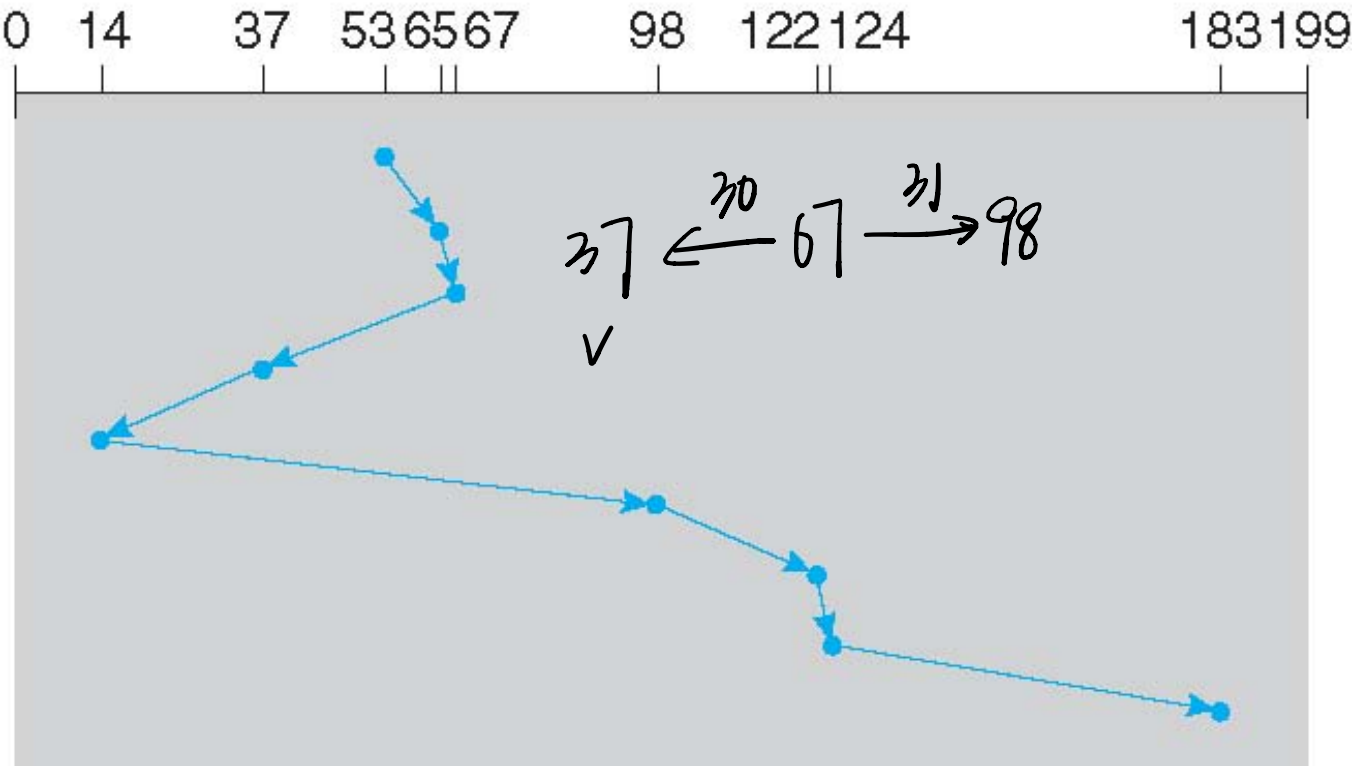
SSTF Scheduling

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

选择相对当前头指针而言的最短seek time.

queue = 98, 183, 37, 122, 14, 124, 65, 67

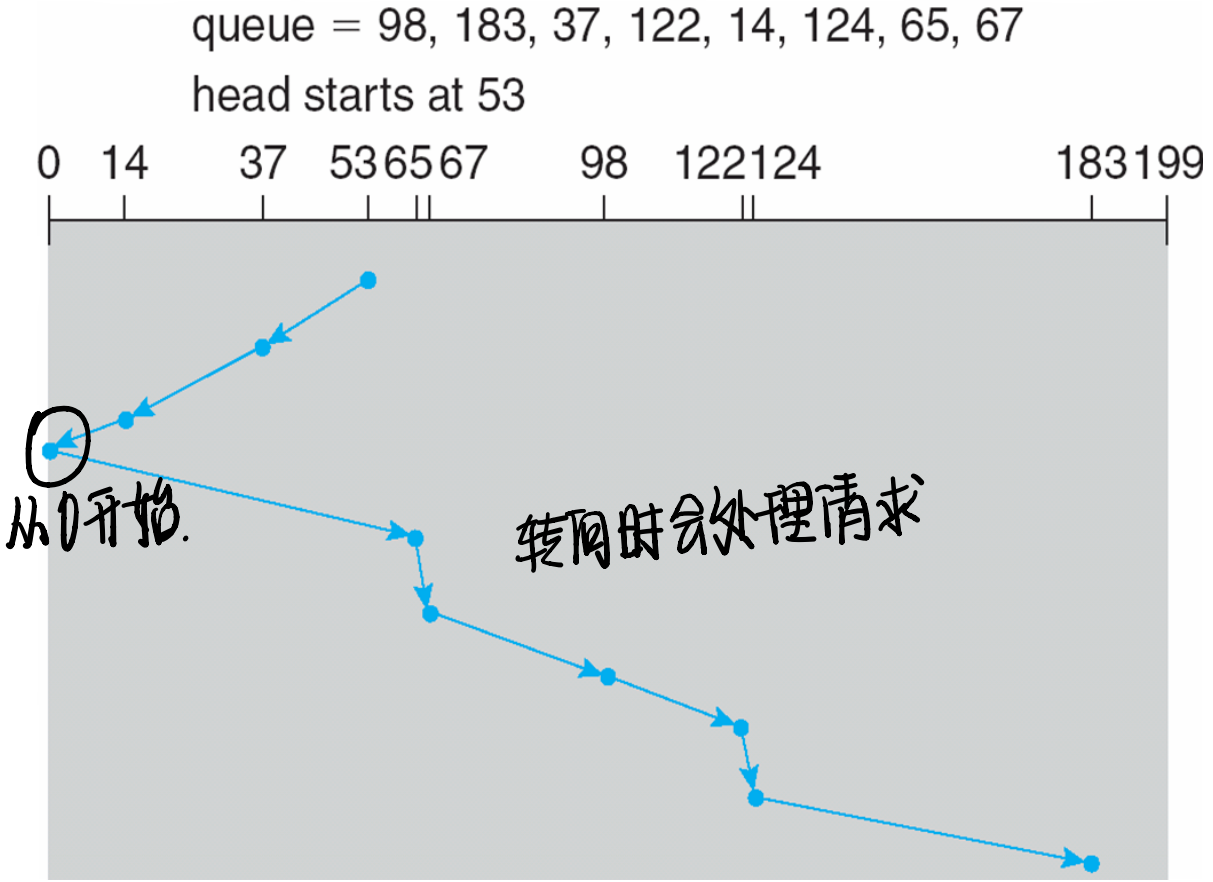
head starts at 53



Total head movement: 236 cylinders.

SCAN

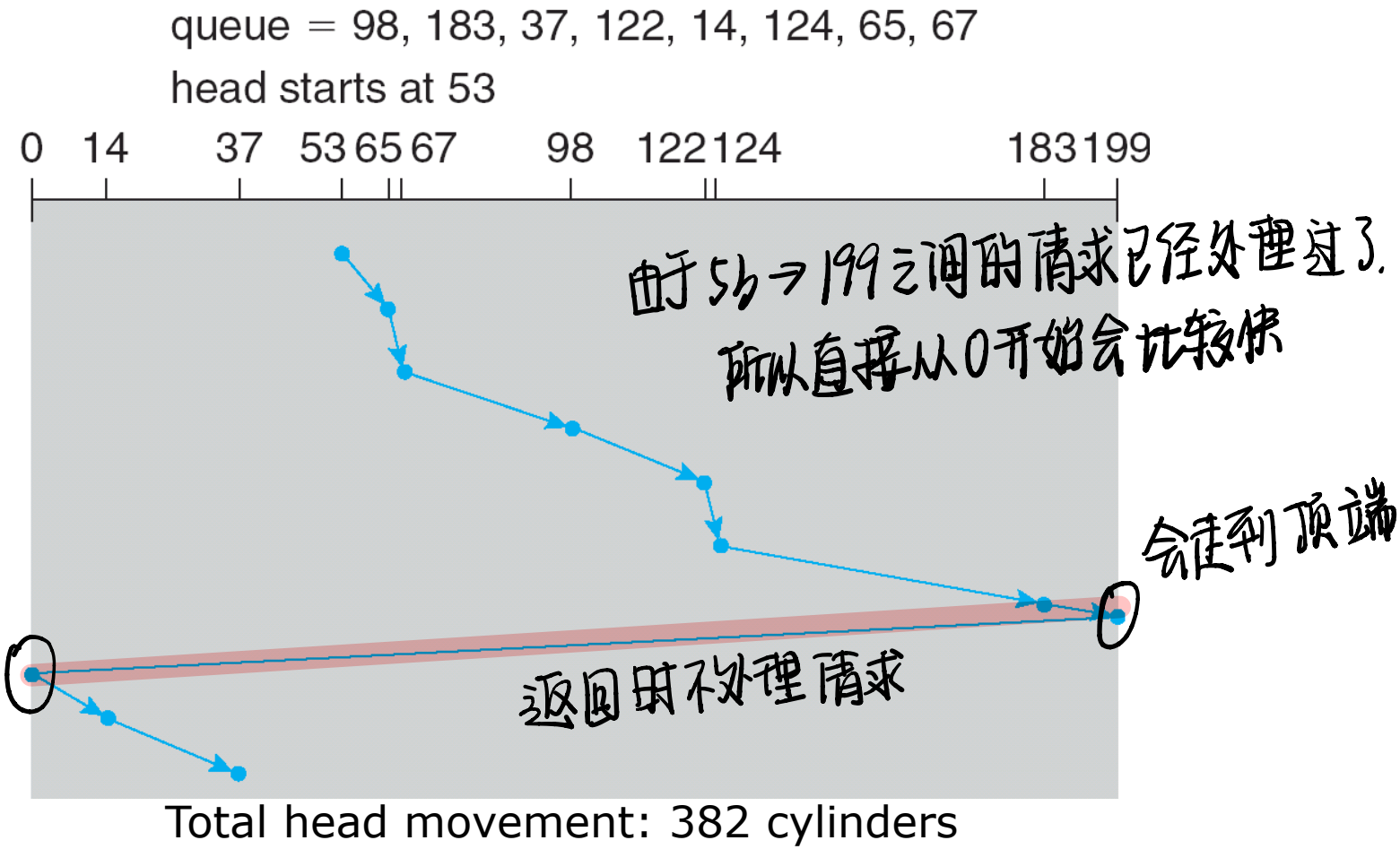
- The disk arm starts at one end of the disk, and moves toward the other end, servicing requests until it gets to the other end of the disk, where the head movement is reversed and service continues.



Total head movement: 236 cylinders

C-SCAN

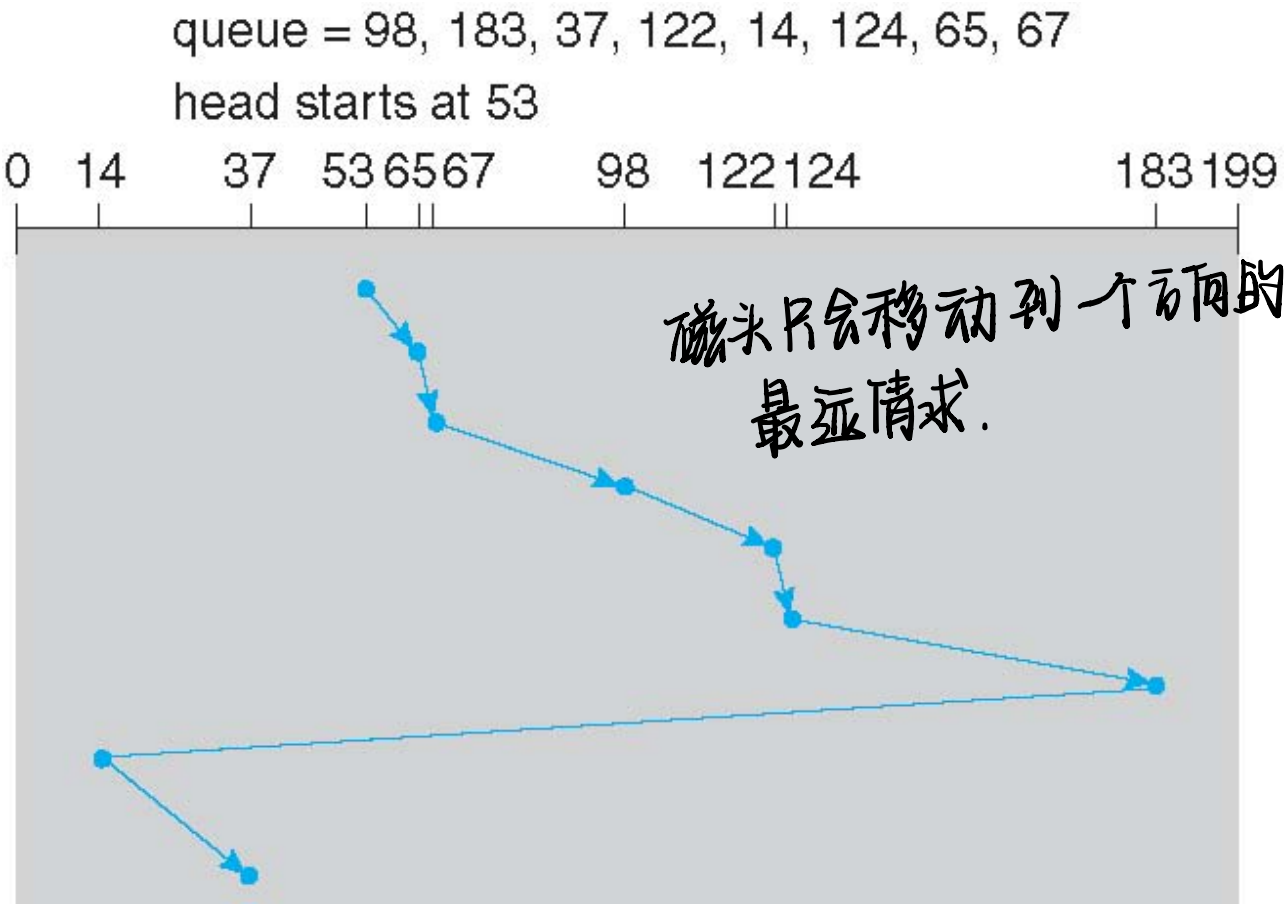
- The head moves from one end of the disk to the other, servicing requests as it goes. When it reaches the other end, it immediately returns to the beginning of the disk, without servicing any requests on the return trip.



Total head movement: 382 cylinders

C-LOOK

- Arm only goes as far as the last request in each direction, then reverses direction immediately, without first going all the way to the end of the disk



Total head movement: 322 cylinders

Pop-Quiz

- Suppose that a disk drive has 5,000 cylinders, numbered 0 to 4999. The drive is currently serving a request at cylinder 143, and the previous request was at cylinder 125. The queue of pending requests, in FIFO order, is:
  - 86,1470,913,1774,948,1509,1022,1750,130

- Starting from the current head position, what is the total distance (in cylinders) that the disk arm moves to satisfy all the pending requests for each of the following disk-scheduling algorithms?
  a) SSTF
  b) C-LOOK

a) 143 → 130 → 86 → 913 → 948 → 1022 → 1470 → 1509 → 1750 → 1774

b) 143 → 913 → 948 → 1022 → 1470 → 1509 → 1750 → 1774 → 86 → 130.

Nonvolatile Memory Devices

- If disk-drive like, then called **solid-state disks** (**SSD**s)
- Other forms include **USB drives** (thumb drive, flash drive), DRAM disk replacements, surface-mounted on motherboards, and main storage in devices like smartphones
- Can be more reliable than HDDs
- More expensive per MB
- Maybe have shorter life span – need careful management
- Less capacity
- But much faster
- Busses can be too slow -> connect directly to PCI for example
- No moving parts, so no seek time or rotational latency

Solid-State Drive (SSD)

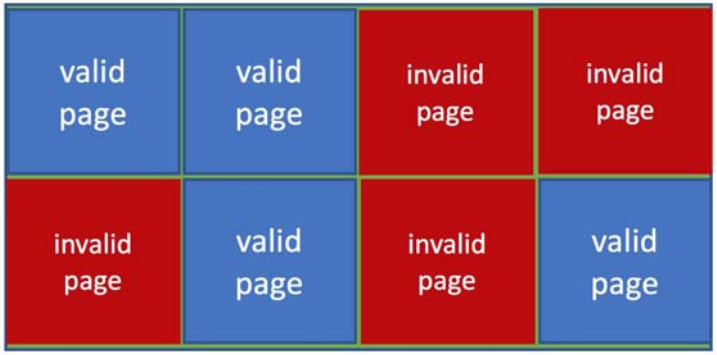
SSDs use microchips which retain data in non-volatile memory chips and contain no moving parts

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
  - 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
- Also implements **garbage collection** to free invalid page space
- Allocates **overprovisioning** to provide working space for GC
- Each cell has lifespan, so **wear leveling** needed to write equally to all cells

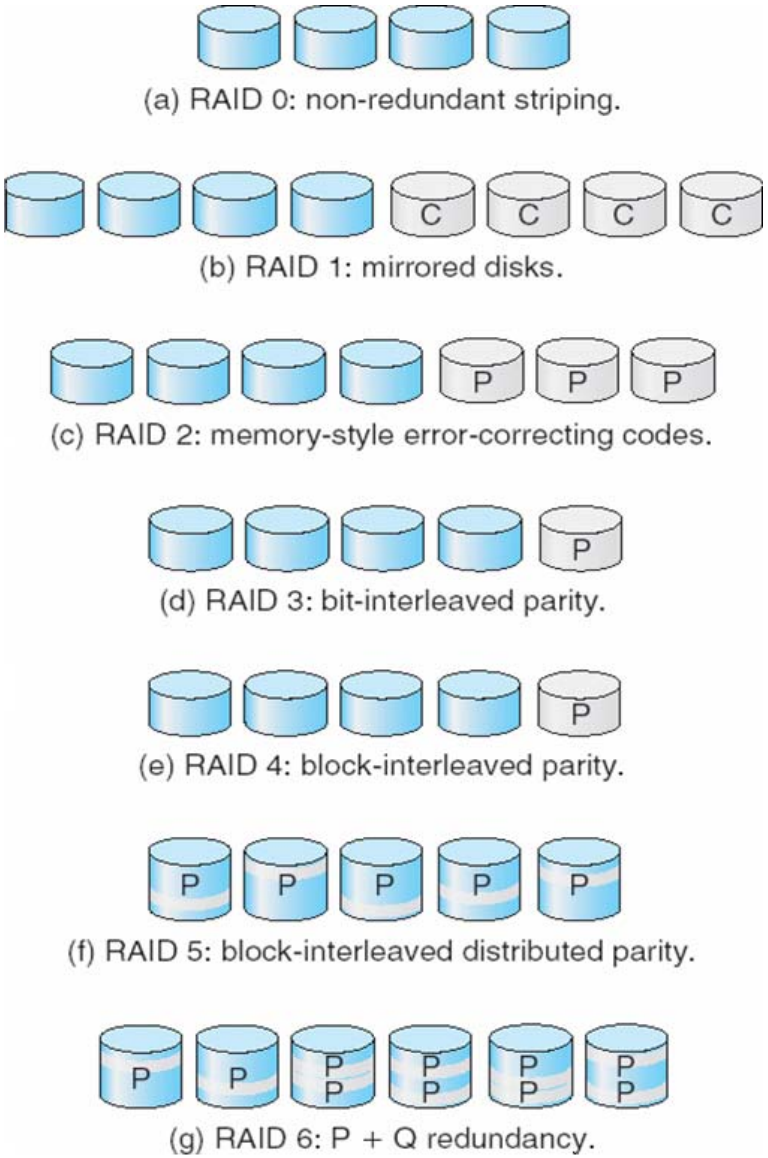


NAND block with valid and invalid pages

RAID Structure

- Redundant Arrays of Inexpensive Disks (**RAID**s)
- RAID – multiple disk drives provides reliability via **redundancy**
  - **Mirroring**
    - duplicate every disk
  - **Parity bit**
- **Parallel access** to multiple disk 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 different levels

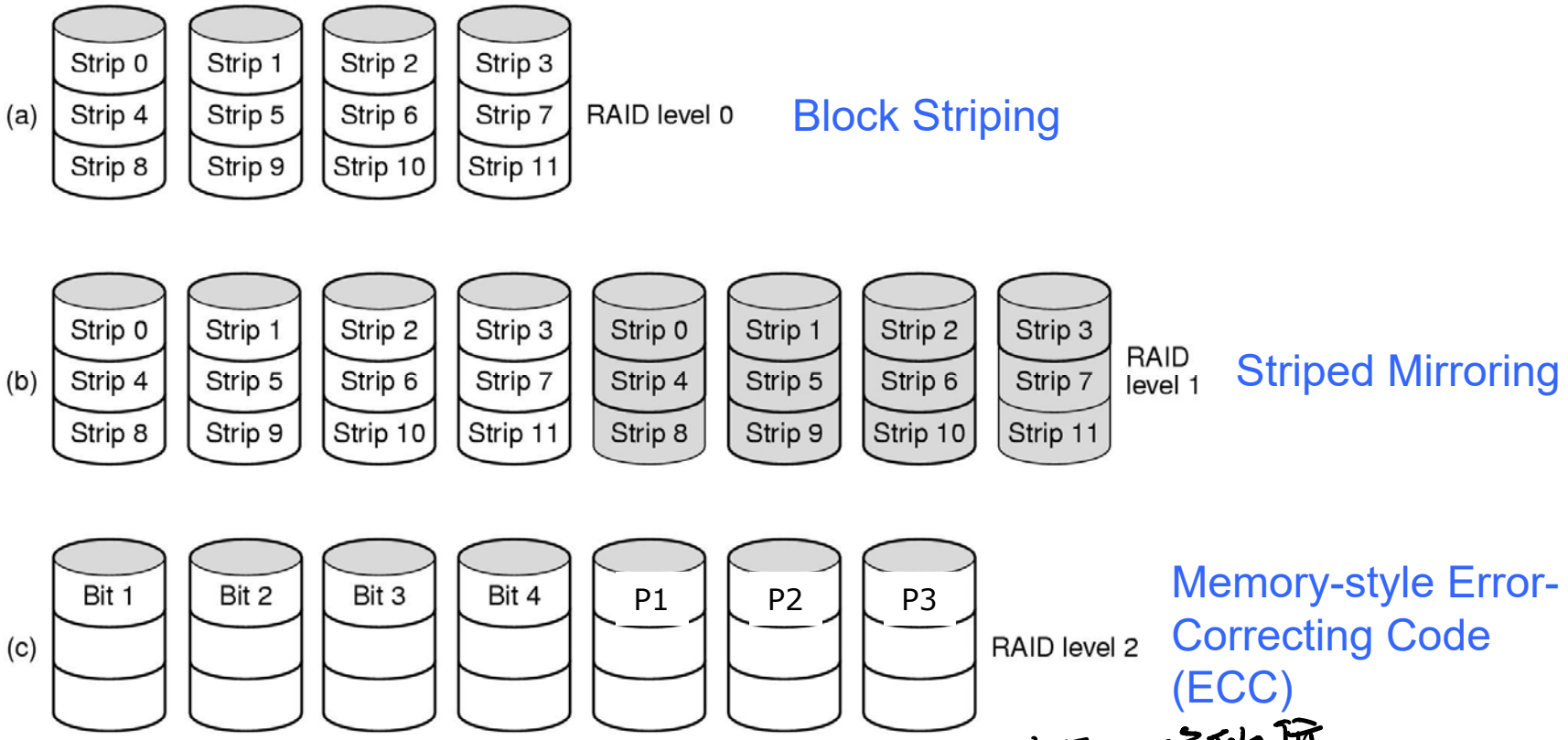
RAID Levels



C: a second copy

P: error-correcting bit

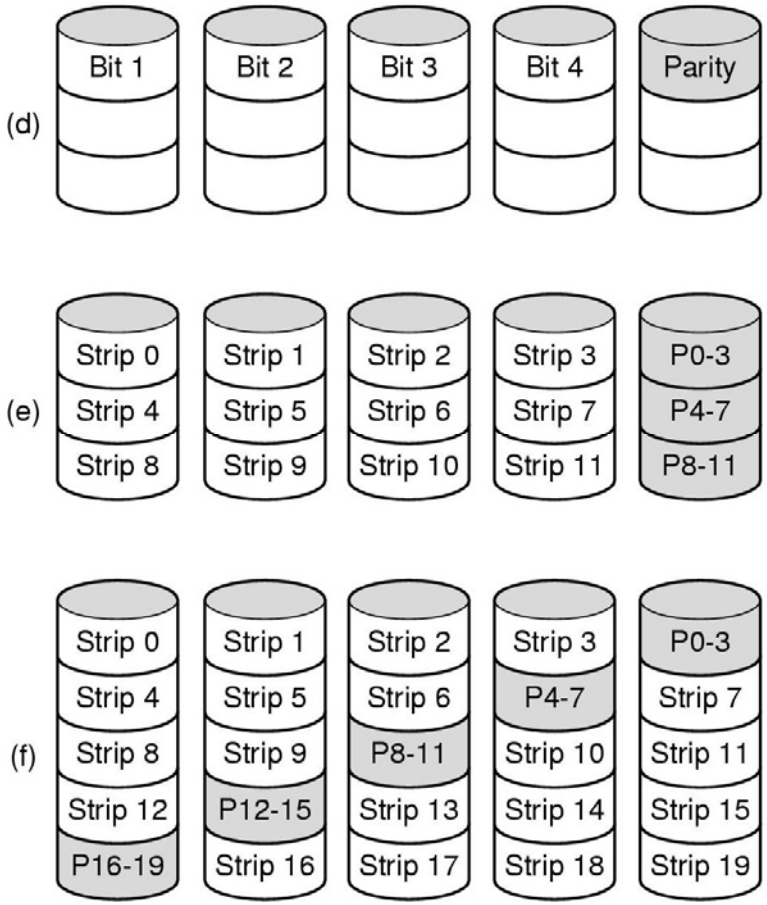
RAID Levels (Cont.)



Bit1损坏可以读取P1来查看正确的位

RAID Levels (Cont.)

Bit 2 坏.
↗ 取 Bit 1, 3, 4 和 Parity 就知道 Bit 2 是 0 还是 1



bit-interleaved parity

block-interleaved parity

block-interleaved distributed parity

RAID Levels (Cont.)

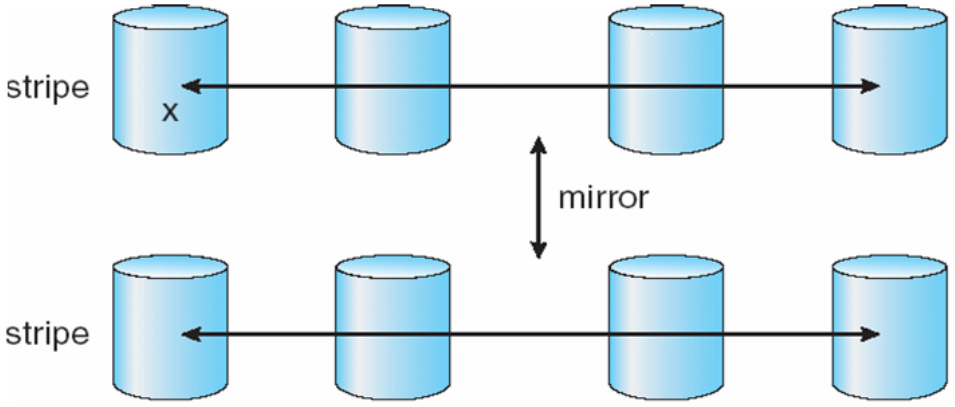
- RAID 6: P + Q redundancy
  - Reed-Solomon codes
  - 2 bits of redundant data are stored for every 4 bits of data
  - can tolerate two disk failures

每4位数据使用了2位的冗余数据.

RAID (Cont.)

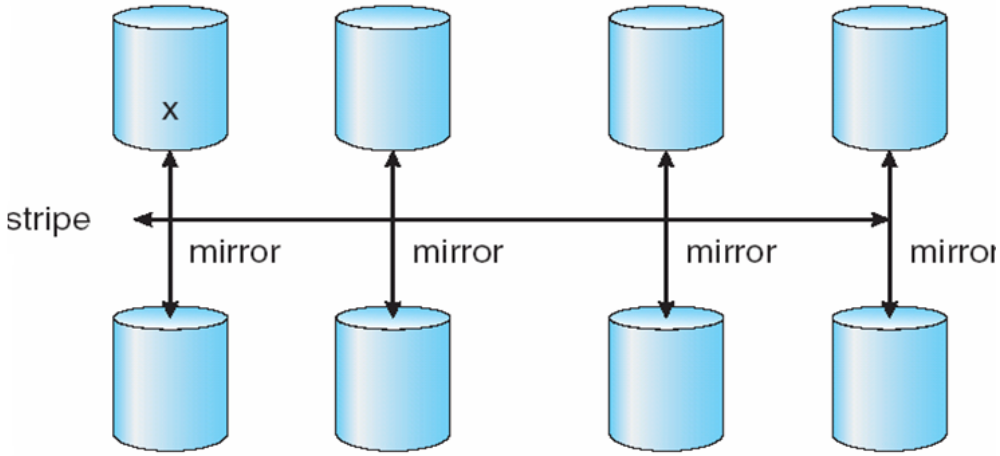
- Several improvements in disk-use techniques involve the use of multiple disks working cooperatively
- Disk **striping** uses a group of disks as one storage unit
- RAID schemes improve performance and improve the reliability of the storage system by storing redundant data
  - **Mirroring** or **shadowing** (**RAID 1**) keeps duplicate of each disk
  - Striped mirrors (**RAID 1+0**) or mirrored stripes (**RAID 0+1**) provides high performance and high reliability
  - **Block interleaved parity** (**RAID 4, 5, 6**) uses much less redundancy
- RAID within a storage array can still fail if the array fails, so automatic **replication** of the data between arrays is common
- Frequently, a small number of **hot-spare** disks are left unallocated, automatically replacing a failed disk and having data rebuilt onto them

RAID (0 + 1) and (1 + 0)



a) RAID 0 + 1 with a single disk failure.

先分条
再镜像



b) RAID 1 + 0 with a single disk failure.

先镜像
再分条

Disk Attachment

- Drive attached to computer via **I/O bus**
  - Busses vary, including **EIDE**, **ATA**, **SATA**, **USB**, **SCSI**, **Fiber Channel, SAS, Firewire**
  - **Host controller** in computer uses bus to talk to **disk controller** built into drive

- SCSI itself is a bus, up to 16 devices on one cable, **SCSI initiator** requests operation and **SCSI targets** perform tasks
  - Each target can have up to 8 **logical units** (disks attached to device controller)

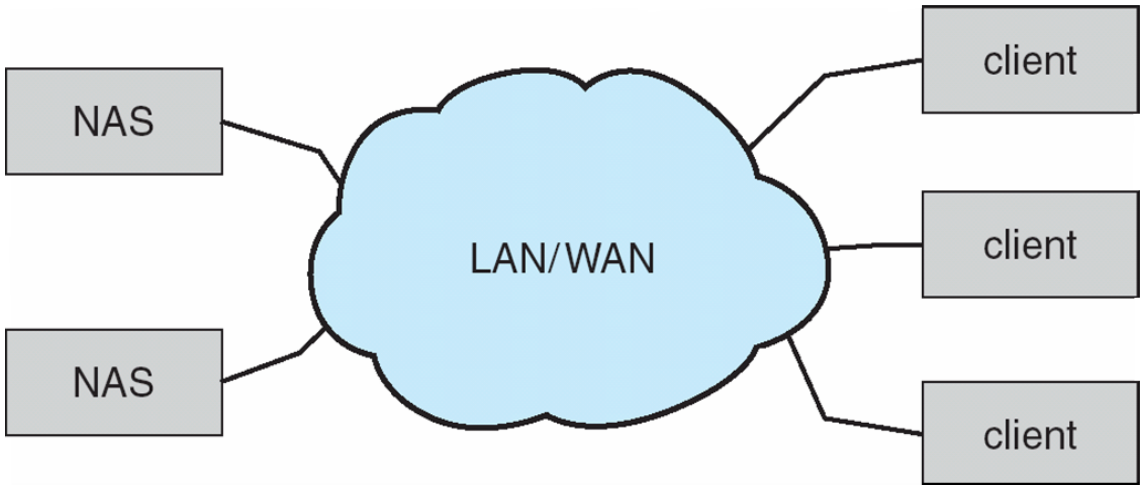
- FC is high-speed serial architecture
  - Can be switched fabric with 24-bit address space – the basis of **storage area networks (SANs)** in which many hosts attach to many storage units

Storage Array

- Can just attach disks, or arrays of disks
- Storage Array has controller(s), provides features to attached host(s)
  - Ports to connect hosts to array
  - Memory, controlling software (sometimes NVRAM, etc)
  - A few to thousands of disks
  - RAID, hot spares, hot swap (discussed later)
  - Shared storage -> more efficiency
  - Features found in some file systems
    - Snaphots, clones, thin provisioning, replication, deduplication, etc

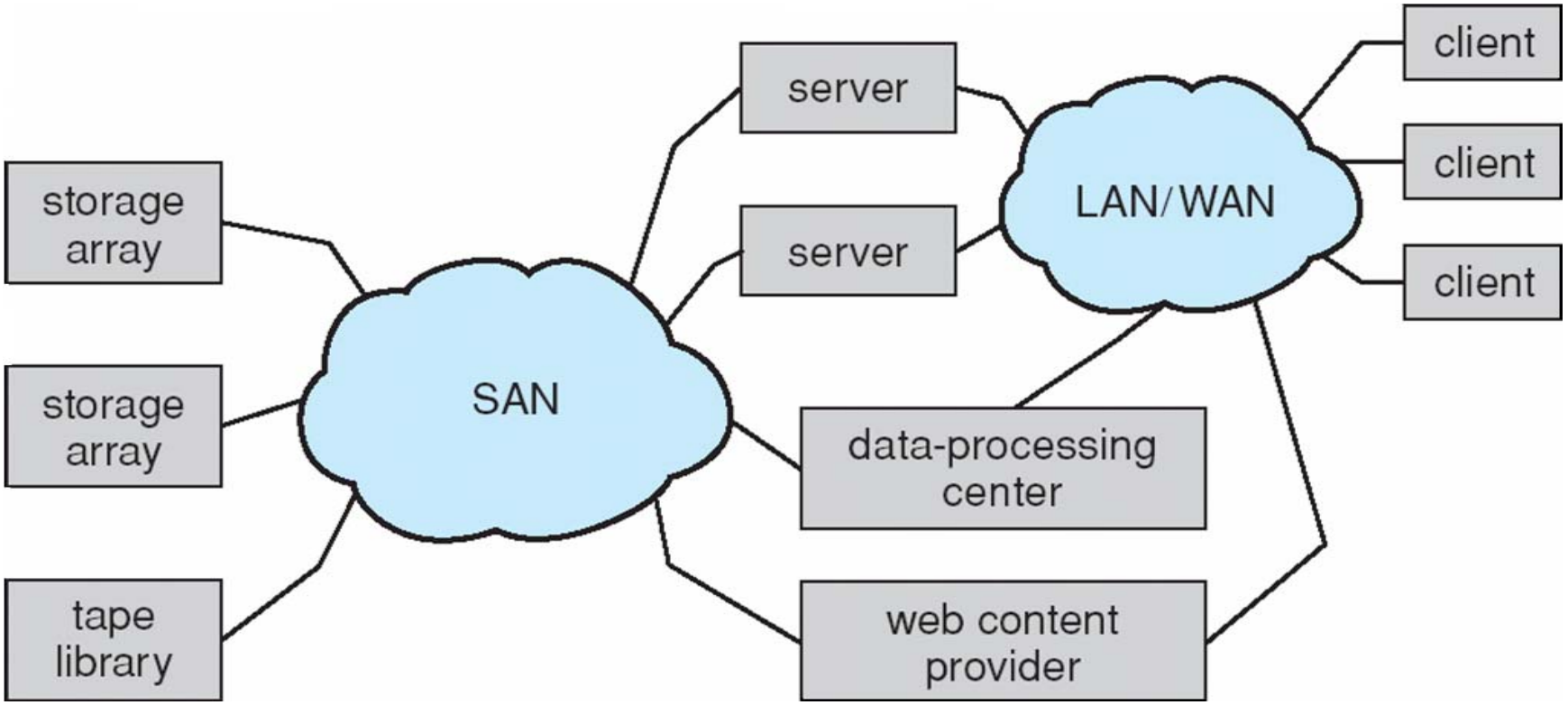
Network-Attached Storage

- 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
- 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
  - Remotely attaching to devices (blocks)



Storage Area Network

- Common in large storage environments
- Multiple hosts attached to multiple storage arrays - flexible



Homework

- Reading
  - Chapter 11
- Exercise
  - See course website