Chapter 12: Mass Storage Structure

肖 卿 俊

办公室：九龙湖校区计算机楼212室

电邮：csqjxiao@seu.edu.cn

主页： https://csqjxiao.github.io/PersonalPage

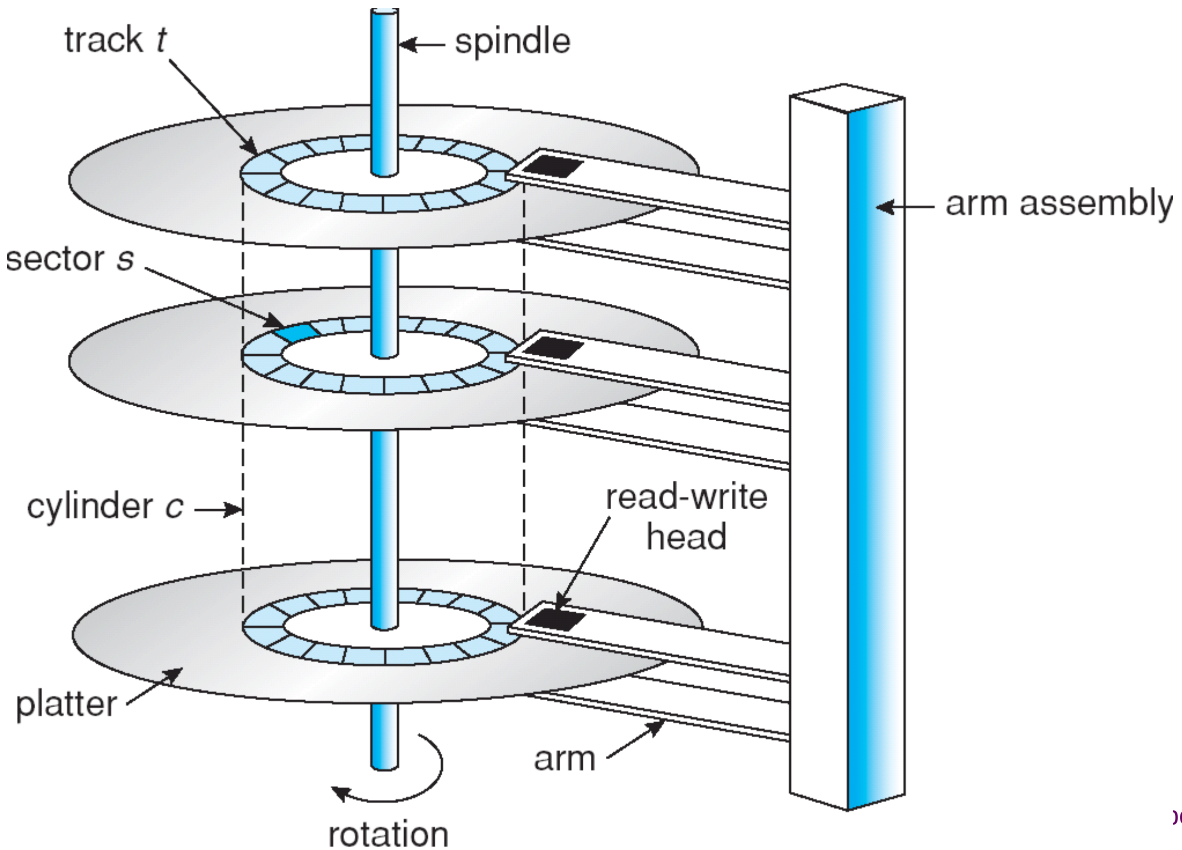
电话：025-52091022

Chapter 12: Mass-Storage Systems

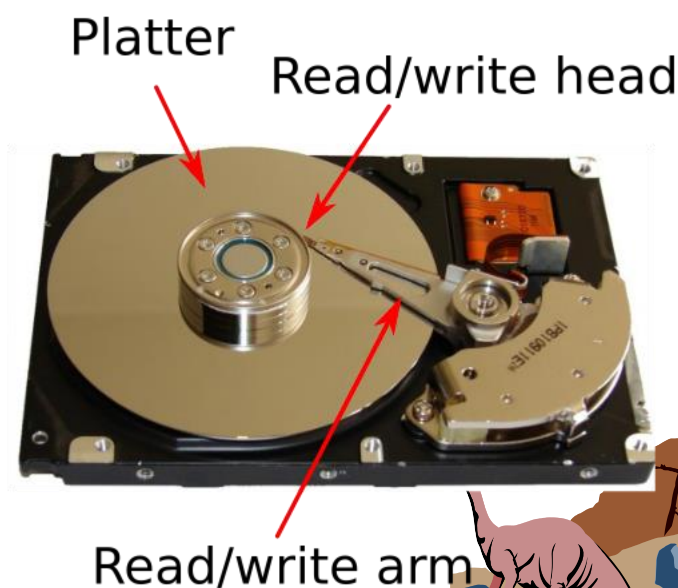
- Overview of Mass Storage Structure
- Disk Structure
- Disk Attachment
- Disk Scheduling
- Disk Management
- Swap-Space Management
- RAID Structure
- Stable-Storage Implementation

Magnetic Disk Hardware

- Disk drives are addressed as large 1-dimensional arrays of logical blocks
  - logical block: smallest unit of transfer (sector)

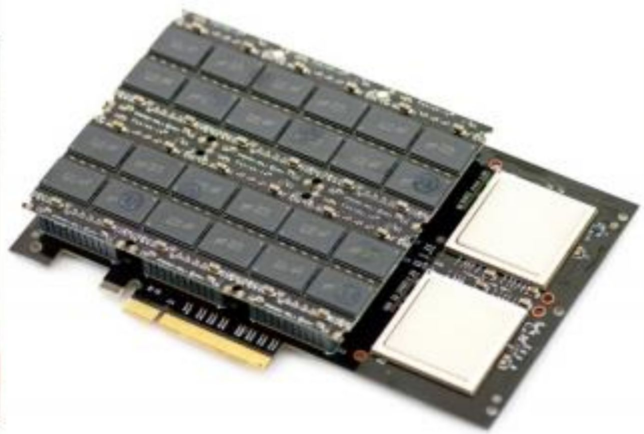


A Real HDD



Disk Attachment

- Host-attached storage accessed through I/O ports talking to **I/O bus**
  - Buses vary, including **EIDE**, **ATA**, **SATA**, **USB**, **Fibre Channel**, **SCSI, SAS, Firewire**
  - I/O bus is controlled by controller
    - Host controller (computer end)
    - Disk controller (built into disk drive)



Overview of Mass Storage Structure

- Platters of magnetic disks range from .85” to 14” (historically)
  - Commonly 3.5”, 2.5”, and 1.8”
- Range from 30GB to 3TB per drive
- Disks can be removable

3.5”

2.5”

Solid-State Disks

- Nonvolatile memory used like a magnetic disk
- More expensive per MB
- Less capacity
- But much faster

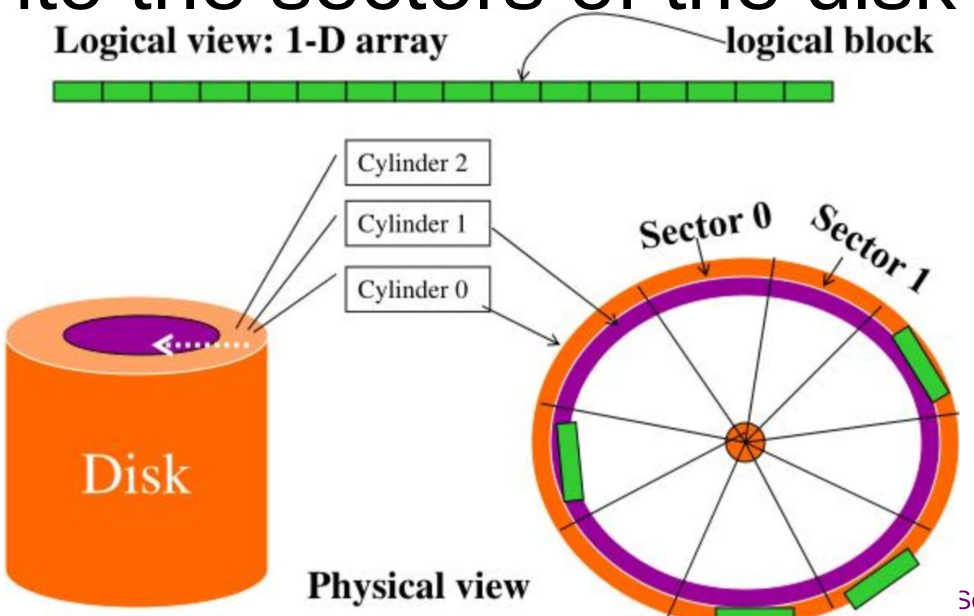
- Can be more reliable than HDDs
- Maybe have shorter life span than magnetic
- No moving parts, so no cylinder seek time or rotational latency

Chapter 12: Mass-Storage Systems

- Overview of Mass Storage Structure
- Disk Structure
- Disk Attachment
- Disk Scheduling
- Disk Management
- Swap-Space Management
- RAID Structure
- Stable-Storage Implementation

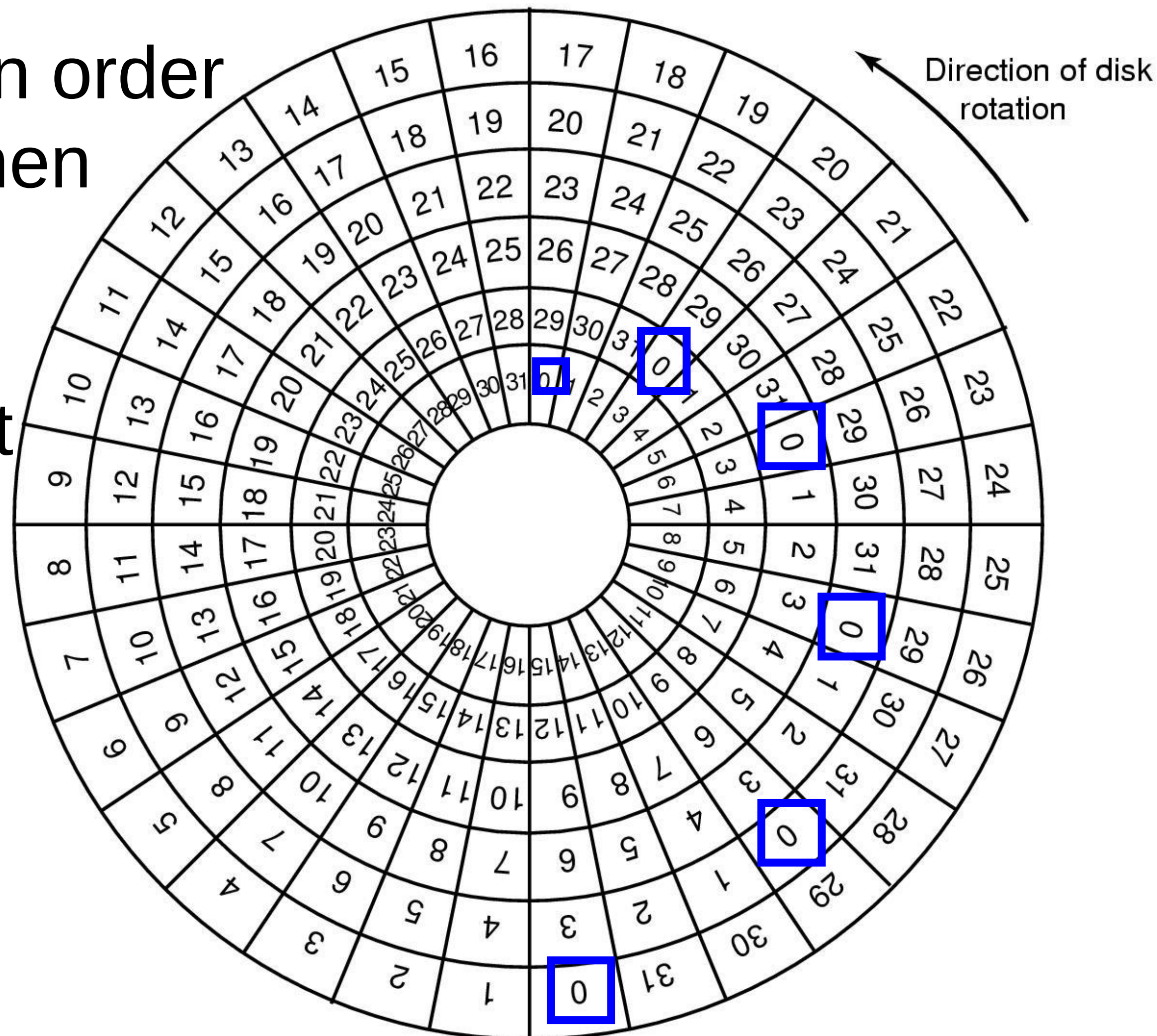
Disk Structure

- Disk drives are addressed as 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



Disk Structure

- 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, and then through the rest of the cylinders from outermost to innermost.



Chapter 12: Mass-Storage Systems

- Overview of Mass Storage Structure
- Disk Structure
- Disk Attachment
- Disk Scheduling
- Disk Management
- Swap-Space Management
- RAID Structure
- Stable-Storage Implementation

Disk Attachment

- Host-attached storage accessed through I/O ports talking to I/O buses
- 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**, which are addressed by logical unit number (LUN)
- 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
- I/O directed to bus ID, device ID, logical unit

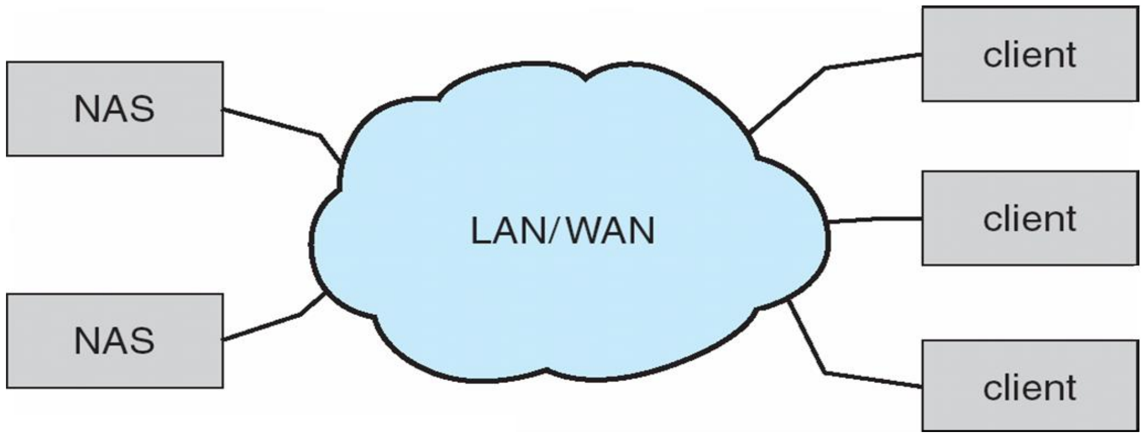
Storage Array (存储阵列)

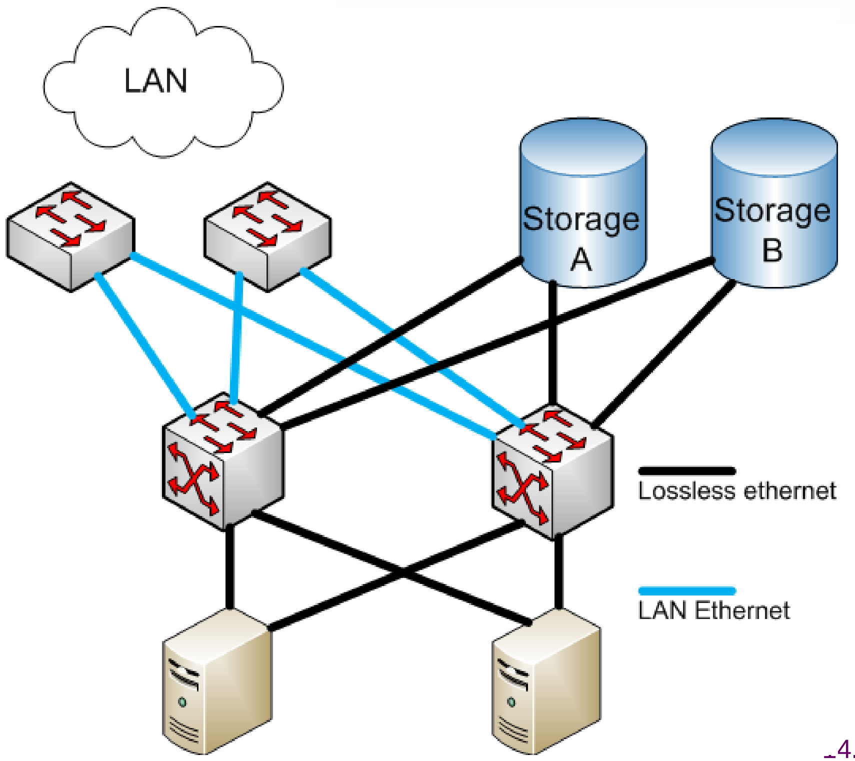
- Can just attach disks, or arrays of disks
- Storage array has controller(s), provides features to attached host(s)
  - Ports to connect hosts to array
  - Memory, controlling software (sometimes NVRAM, etc)
  - A few to thousands of disks
  - RAID, hot spares, hot swap (discussed later)
  - Shared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

Storage Area Network (Cont.)

- SAN is one or more storage arrays
  - Connected to one or more Fibre Channel switches
- Hosts also attach to the switches
- Storage made available via **LUN Masking** from specific arrays to specific servers
- Easy to add or remove storage, add new host and allocate it storage
  - Over low-latency Fibre Channel fabric
- Why have separate storage networks and communications networks?
  - Consider FCOE, iSCSI

Network-Attached Storage





Combined storage and local area network, based on **Fibre Channel over Ethernet (FCoE)**

Network-Attached Storage

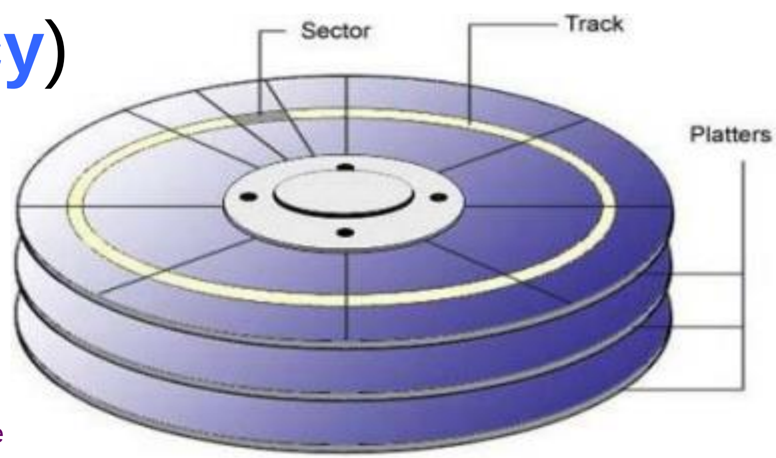
- Network-attached storage (**NAS**) is storage made available over a network rather than over a local connection (such as a bus)
  - Remotely attaching to file systems
- NFS and CIFS are common protocols
- Implemented via remote procedure calls (RPCs) between host and storage over typically TCP or UDP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

Chapter 12: Mass-Storage Systems

- Overview of Mass Storage Structure
- Disk Structure
- Disk Attachment
- Disk Scheduling
- Disk Management
- Swap-Space Management
- RAID Structure
- Stable-Storage Implementation

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 flow 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**)



https://en.wikipedia.org/wiki/Hard_disk_drive_performance_characteristics#Rotational_latency

Magnetic Disk Performance

- 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
    - ✓60s / RPM => time per revolution
  - Avg rotational latency=½ latency

revolutions per minute (RPM)

Typical HDD figures	
HDD Spindle [rpm]	Average rotational latency [ms]
4,200	7.14
5,400	5.56
7,200	4.17
10,000	3.00
15,000	2.00

(From Wikipedia)

Magnetic Disk Performance

- **Access Latency** = **Average access time** = average seek time + average rotational latency
  - For fastest disk 3ms + 2ms = 5ms
  - For slow disk 9ms + 5.56ms = 14.56ms
- Average I/O time = average access time + (bytes to transfer / transfer rate) + controller overhead
  - For example to transfer a 4KB block on a 7200 RPM disk with a 5ms average seek time, 1Gb/sec transfer rate, with a 0.1ms controller overhead,
  - Average I/O time = 5ms + 4.17ms + 4KB / 1Gb/sec + 0.1ms = 9.27ms + 0.12ms = 9.39ms

Disk Scheduling

- The operating system is responsible for using hardware efficiently — for the disk drives, this means having a fast *access time* and a large *disk bandwidth*.

- Disk bandwidth is the total number of bytes transferred, divided by the total time between the first request for service and the completion of the last transfer.

Disk Scheduling (cont.)

- Access time has three major components
  - *Seek time* is the time for the disk to move the heads to the cylinder containing the desired sector.
  - *Rotational latency* is the additional time waiting for the disk to rotate the desired sector to the disk head.
  - *Transfer time* is the time to transfer a block of data from the disk to the host computer.

http://www.csc.villanova.edu/~achang/diskintro.html
https://en.wikipedia.org/wiki/Hard_disk_drive_performance_characteristics

Disk Scheduling (cont.)

- Here, we consider seek time as the most dominate parameter, and attempt to minimize the seek time
  - Seek time ≈ seek distance
  - As hard disk seek time improves over time, its role as a bottleneck in hard disk performance diminishes and rotational delay will become a major bottleneck in the not too distant future. Algorithms base not only on seek time reduction, but also on rotational distance reduction already exist in theory.

http://www.csc.villanova.edu/~achang/diskintro.html

Disk Scheduling (cont.)

- Several algorithms exist to schedule the servicing of disk I/O requests.

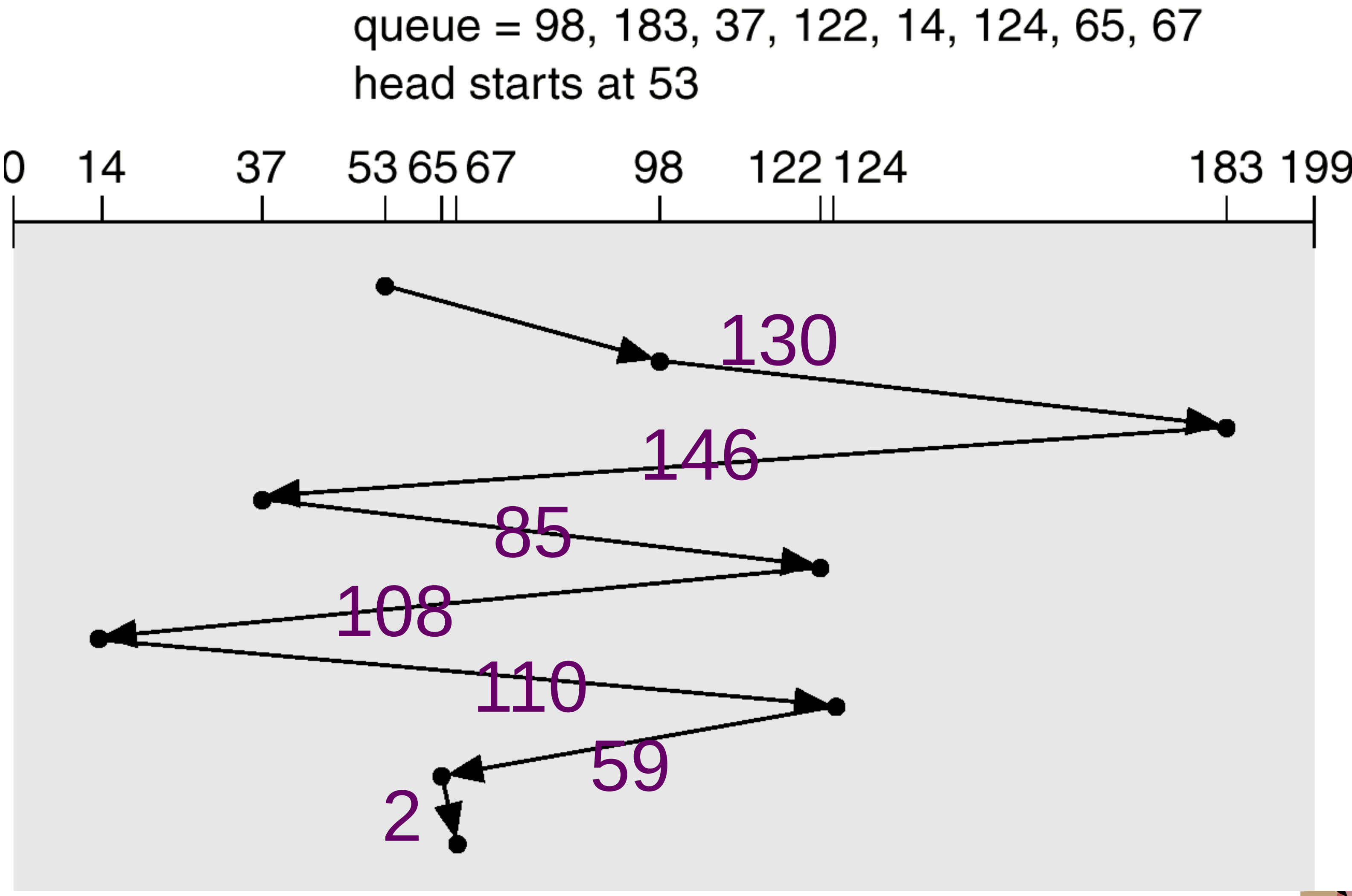
- We illustrate them with a request queue (falling into the range of 0 ~ 199).

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

Head pointer 53

FCFS

Illustration shows total head movement of 640 cylinders.

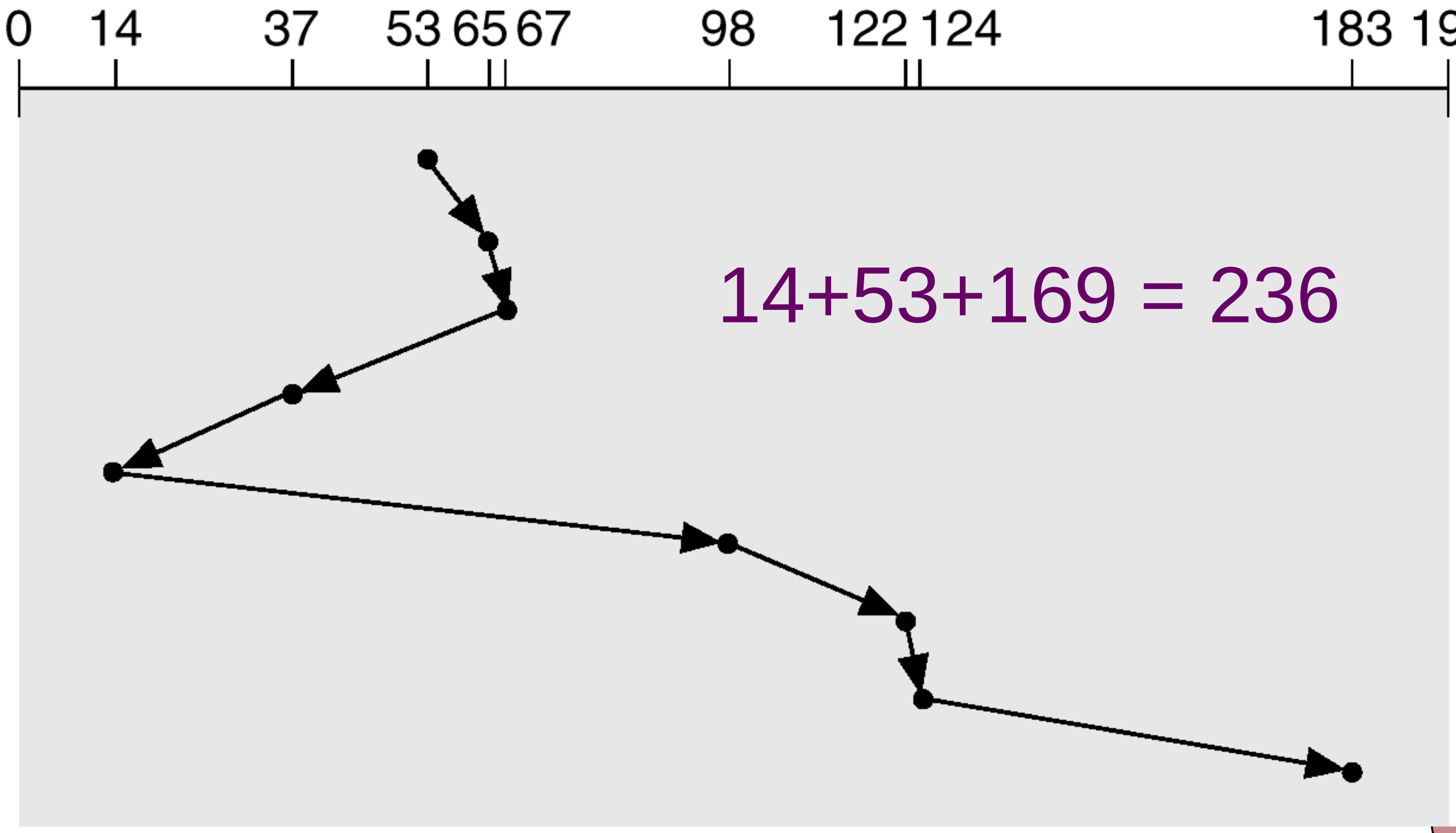


SSTF (Shortest Seek Time First)

- Selects the request with the minimum seek time from the current head position.
- SSTF scheduling is a form of SJF scheduling; may cause starvation of some requests.
- Illustration shows total head movement of 236 cylinders.

SSTF (cont.)

queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53

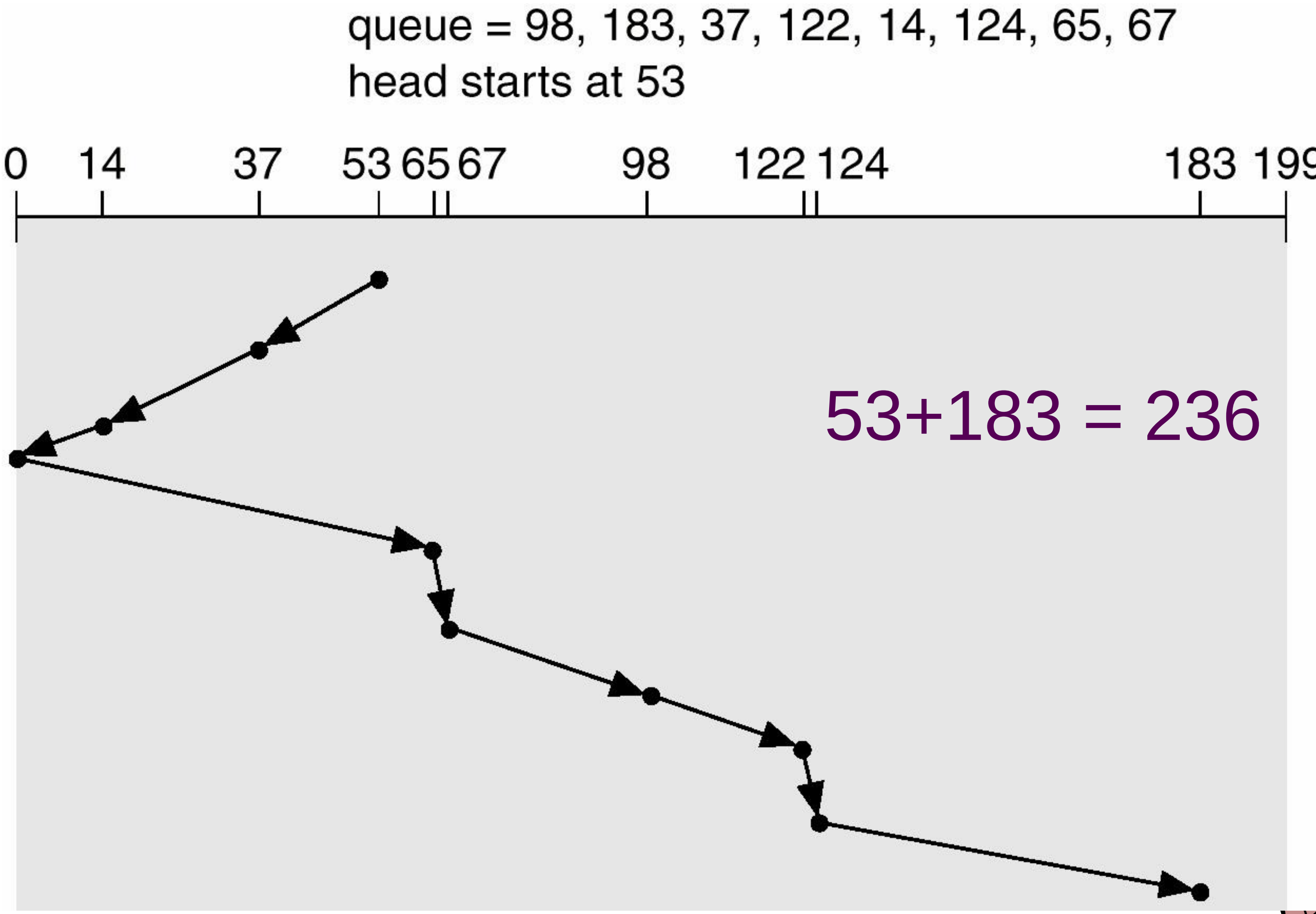


SCAN

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 servicing continues.

- Sometimes called the *elevator algorithm*.
- Illustration shows total head movement of 236 cylinders.

SCAN (cont.)



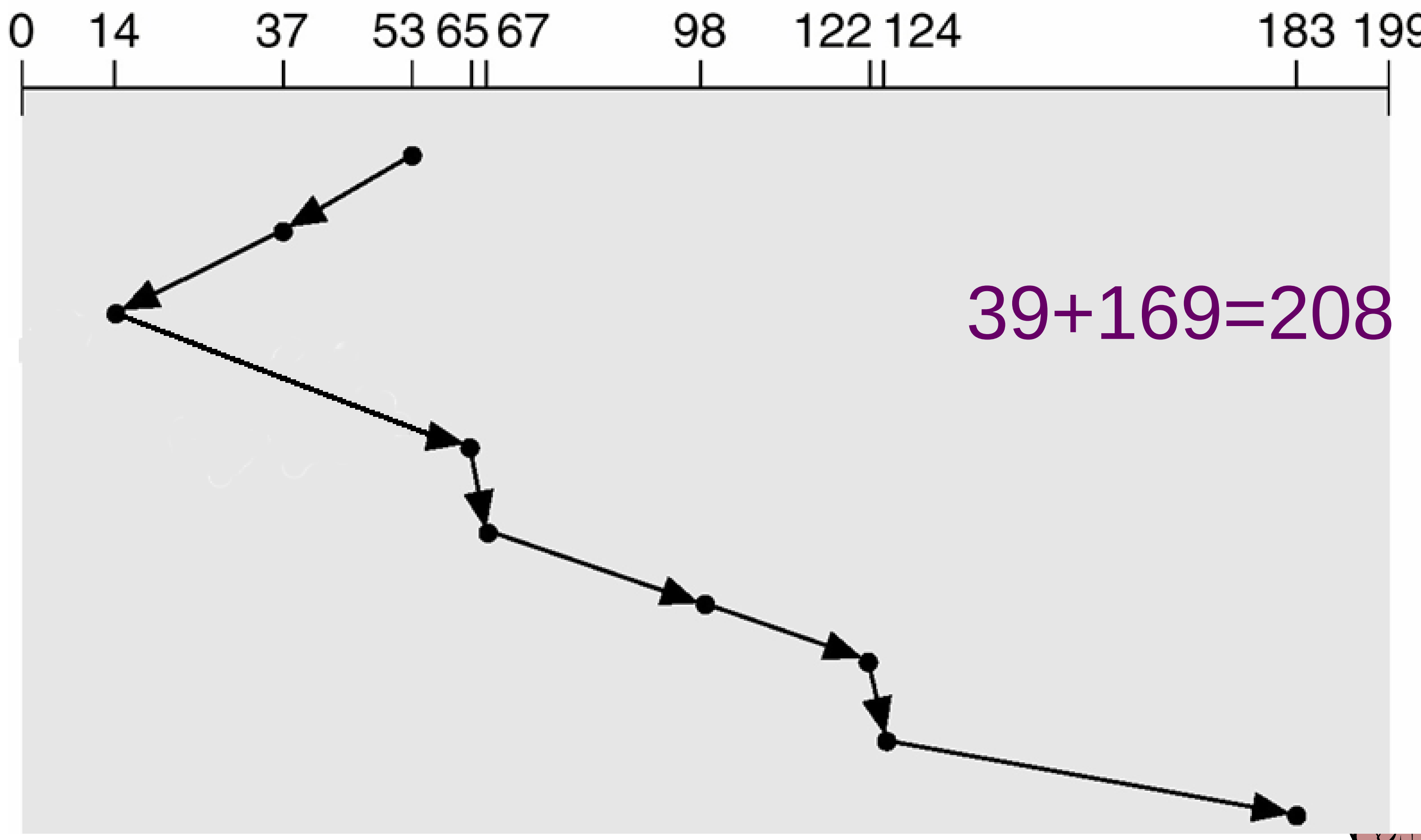
53+183 = 236

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.
- Illustration shows total head movement of 208 cylinders.

LOOK (cont.)

queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53

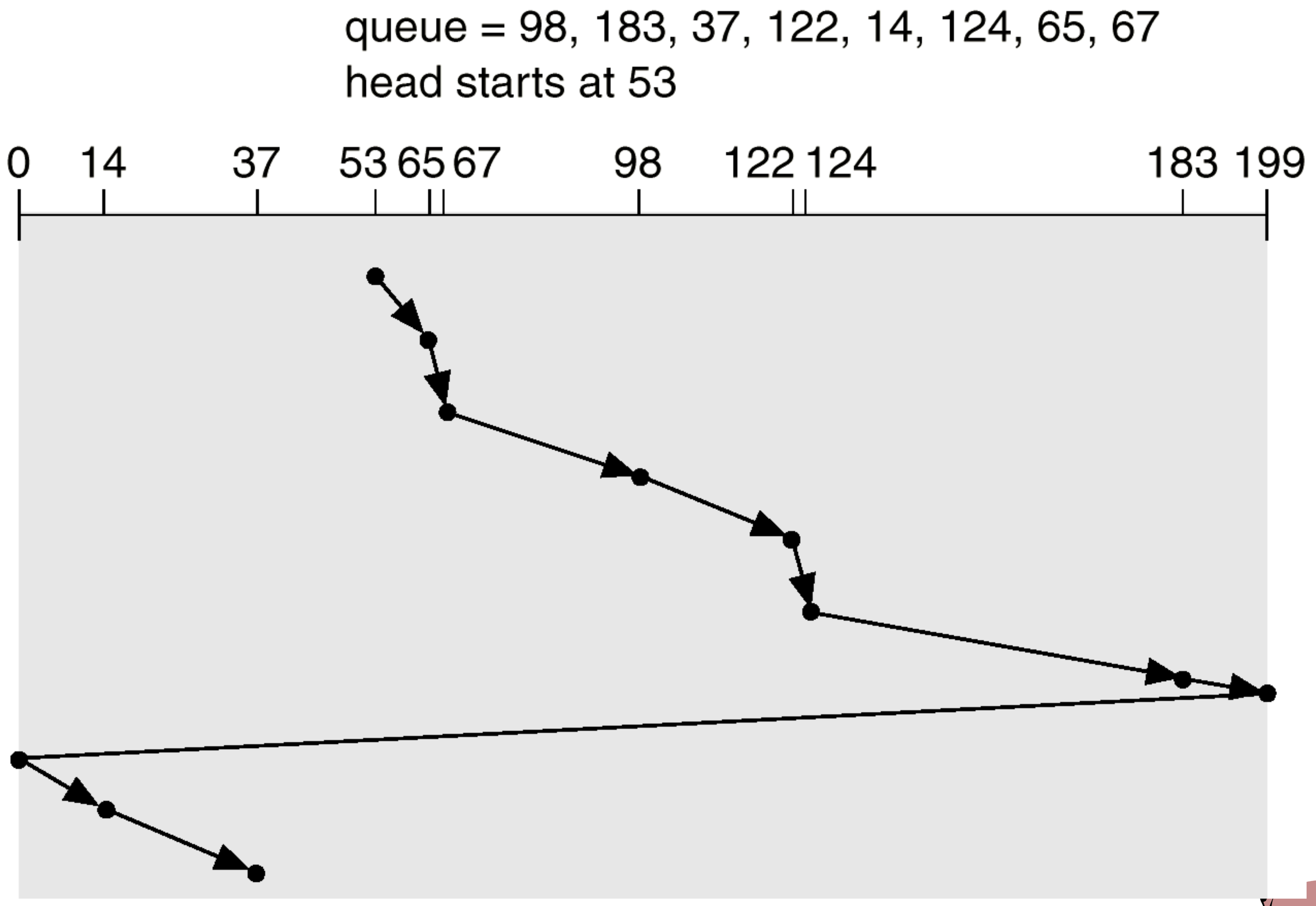


39+169=208

C-SCAN (Circular-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, however, 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.

C-SCAN (cont.)

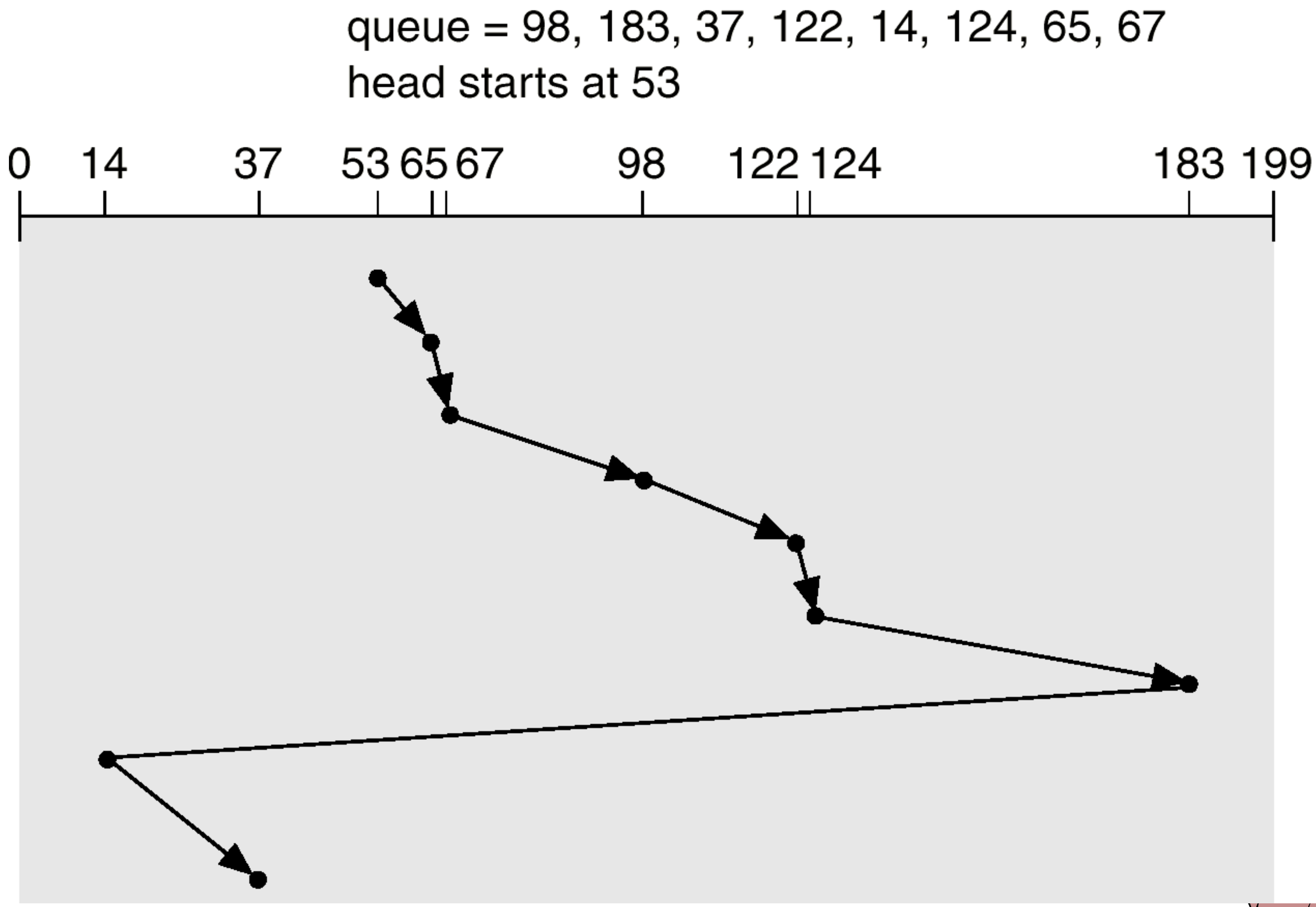


queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0 14 37 53 65 67 98 122 124 183 199

C-LOOK

- Version of C-SCAN
- 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.

C-LOOK (cont.)


queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0 14 37 53 65 67 98 122 124 183 199

Select a Disk-Scheduling Algorithm

- SSTF is common and has a natural appeal
- SCAN and C-SCAN perform better for systems that place a heavy load on the disk
  - Less starvation
- Performance depends on the number and types of requests
- Requests for disk service can be influenced by the file-allocation method
  - And metadata layout
  - 可以参考考研真题

Select a Disk-Scheduling Algorithm (cont.)

- The disk-scheduling algorithm should be written as a separate module of the operating system, allowing it to be replaced with a different algorithm if necessary
- Either SSTF or LOOK is a reasonable choice for the default algorithm
- What about rotational latency?
  - Difficult for OS to calculate
- How does disk-based queuing effect OS queue ordering efforts?

Chapter 12: Mass-Storage Systems

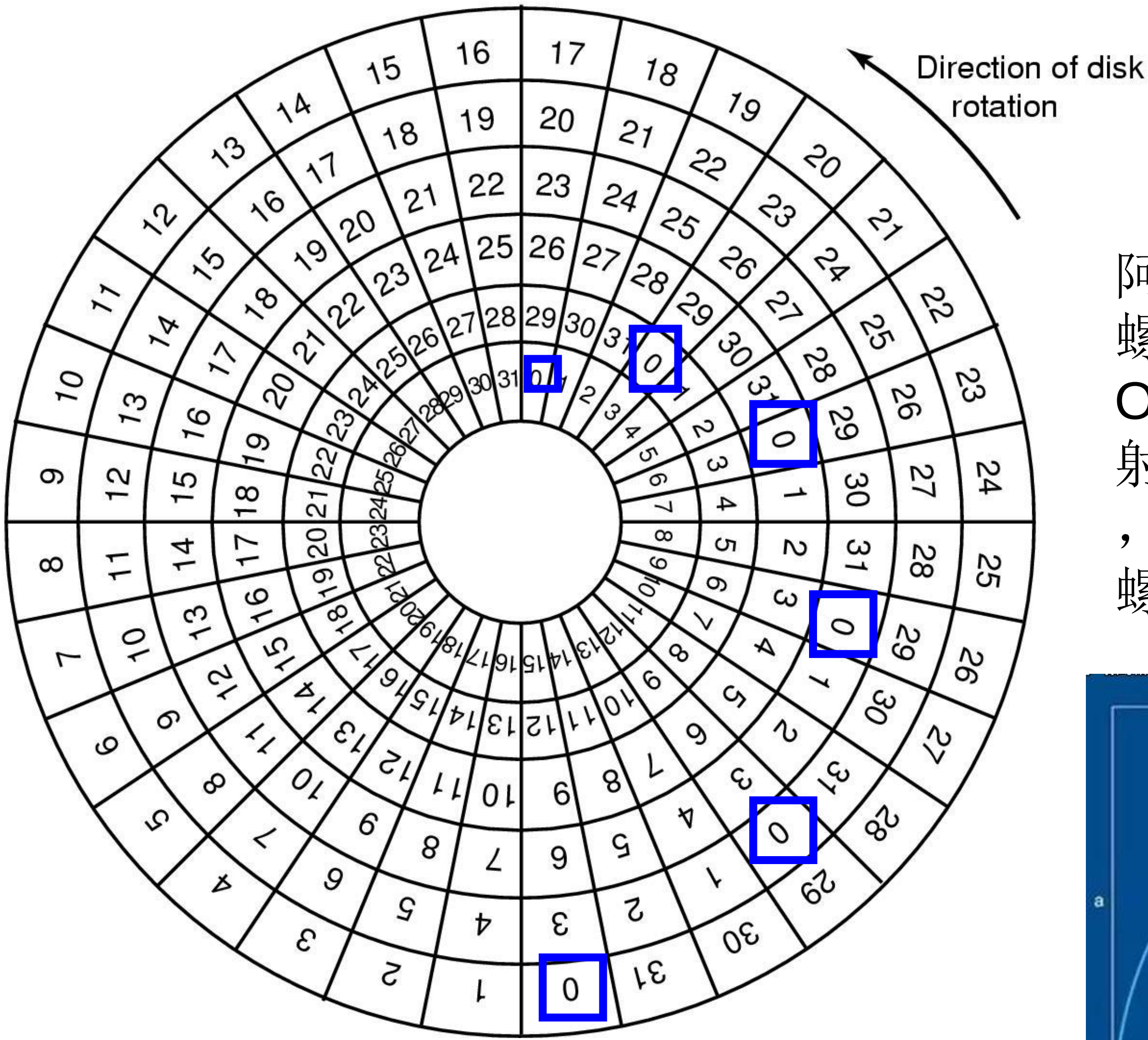
- Overview of Mass Storage Structure
- Disk Structure
- Disk Attachment
- Disk Scheduling
- Disk Management
- Swap-Space Management
- RAID Structure
- Stable-Storage Implementation

Disk Formatting

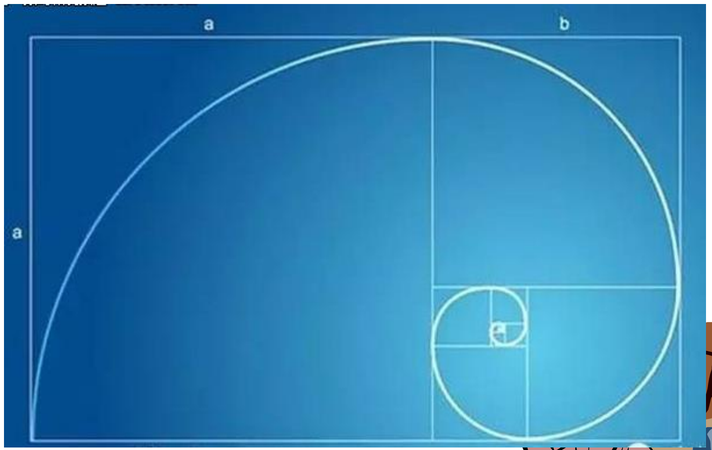
- *Low-level formatting*, or *physical formatting* — Dividing a disk into sectors that the disk controller can read and write.

- 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.
  - *Logical formatting* or “making a file system”.

Disk Formatting (2)

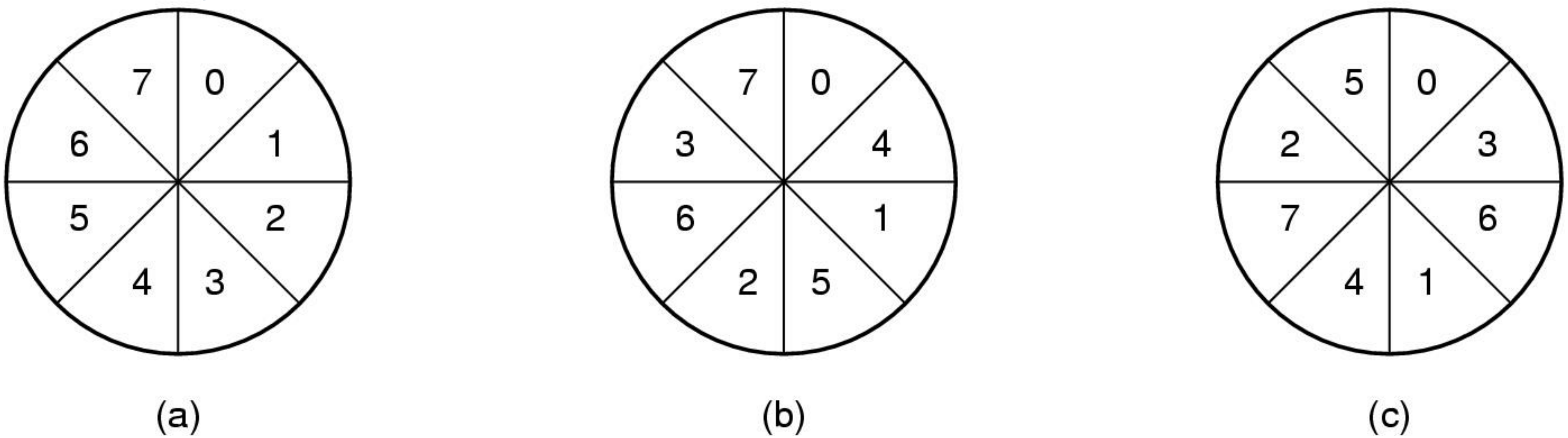


阿基米德螺线，亦称“等速螺线”。当一点P沿动射线OP一等速率运动的同时，这射线有以等角速度绕点O旋转，点P的轨迹称为“阿基米德螺线”。



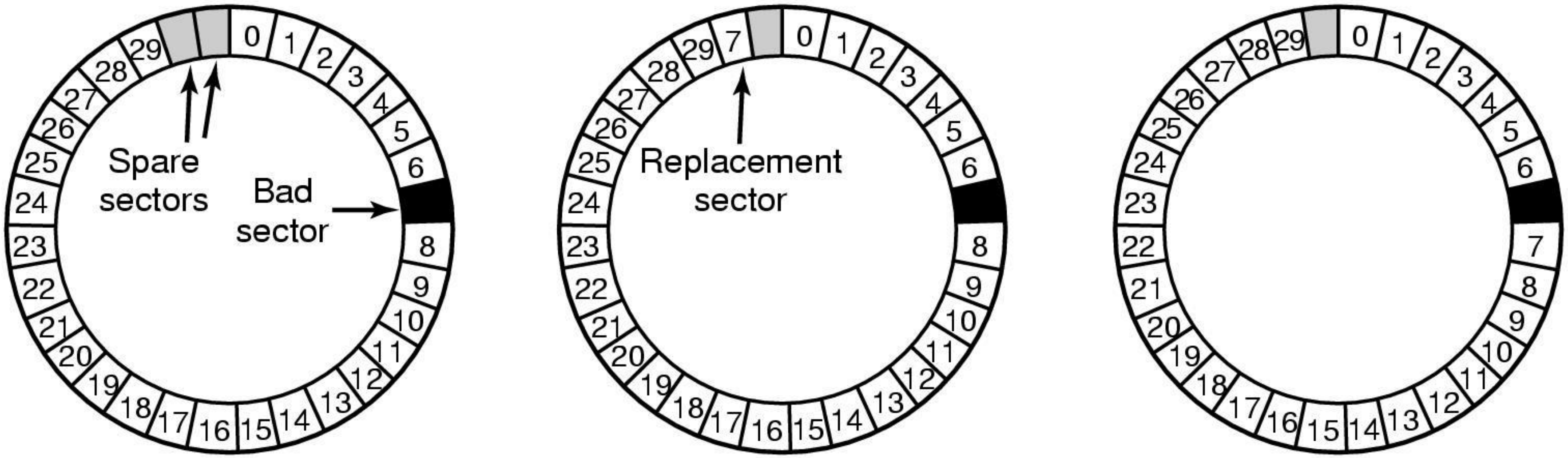
An illustration of cylinder skew

Disk Formatting (3)



- No interleaving
- Single interleaving
- Double interleaving

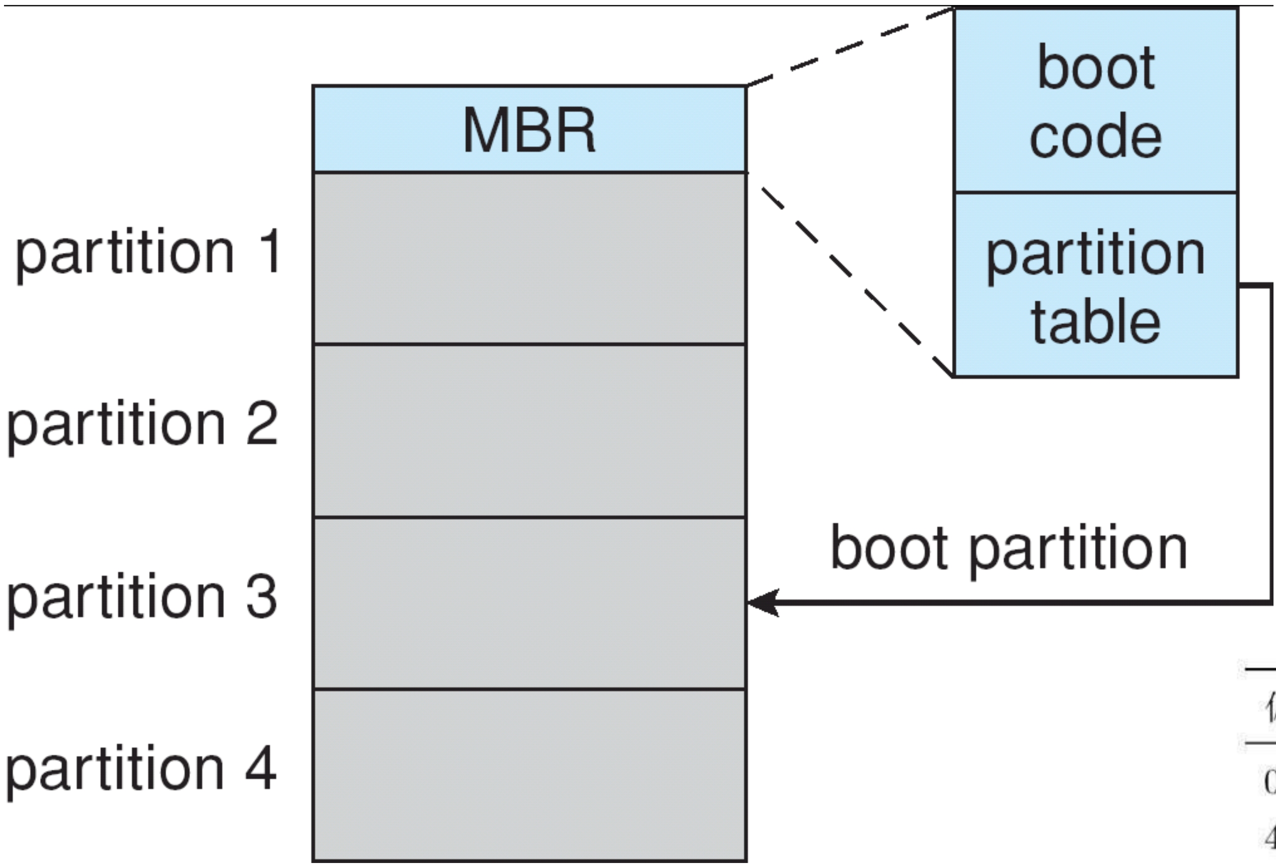
Error Handling



(a) (b) (c)

- A disk track with a bad sector
- Substituting a spare for the bad sector
- Shifting all the sectors to bypass the bad one

Booting from a Disk in Windows 2000



主引导记录（MBR，Master Boot Record）是采用MBR分区表的硬盘的第一个扇区，即C/H/S地址的0柱面0磁头1扇区，也叫做MBR扇区。

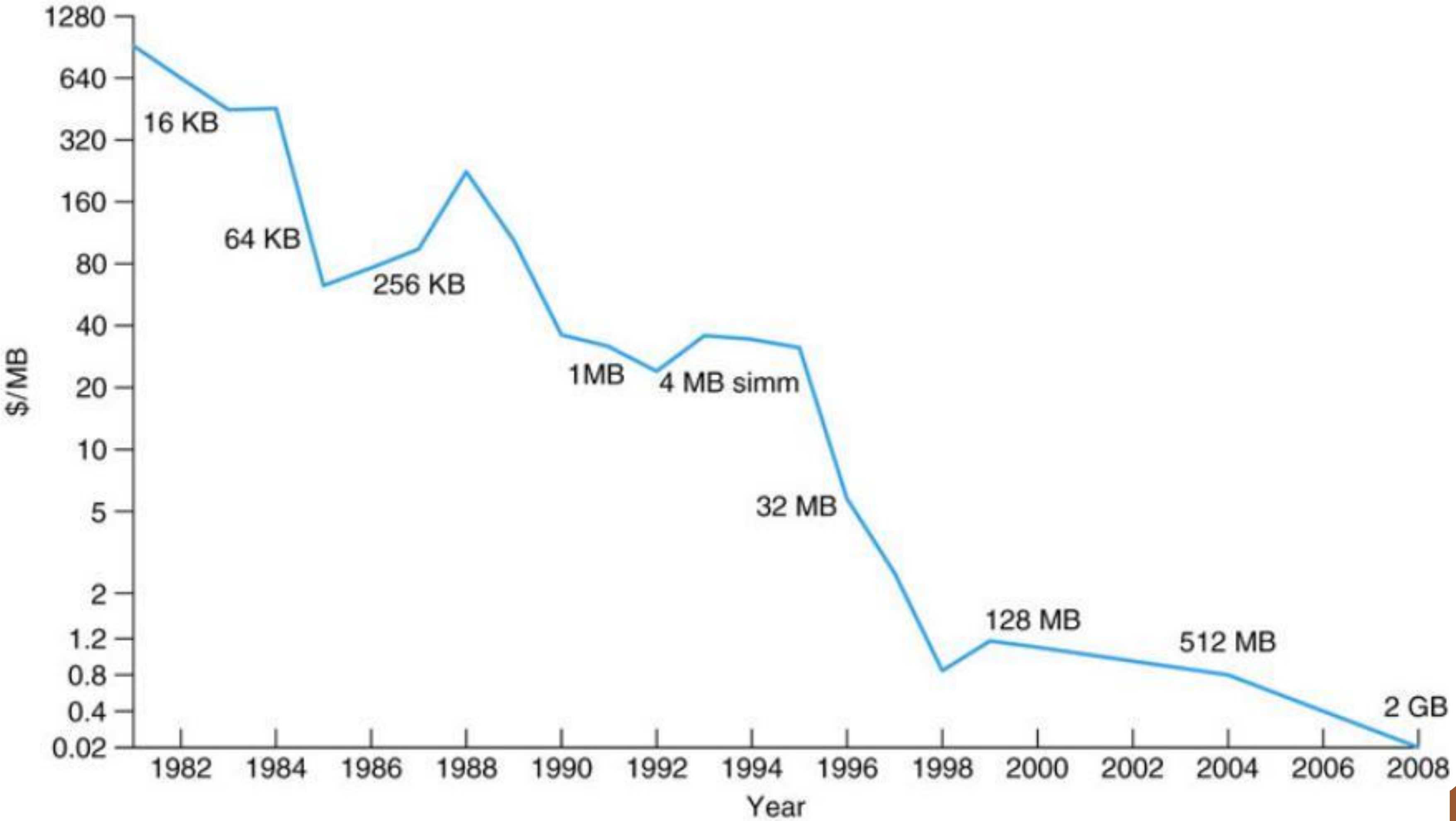
由于历史原因，硬盘的一个扇区大小是512字节，包含最多446字节的启动代码、4个硬盘分区表项（每个表项16字节，共64字节）、2个签名字节（0x55,0xAA），如图所示

偏移地址	字节数	内容
00	446	启动代码
446	16	第 1 分区表项
462	16	第 2 分区表项
478	16	第 3 分区表项
494	16	第 4 分区表项
510	2	签名 (0x55 0xAA)

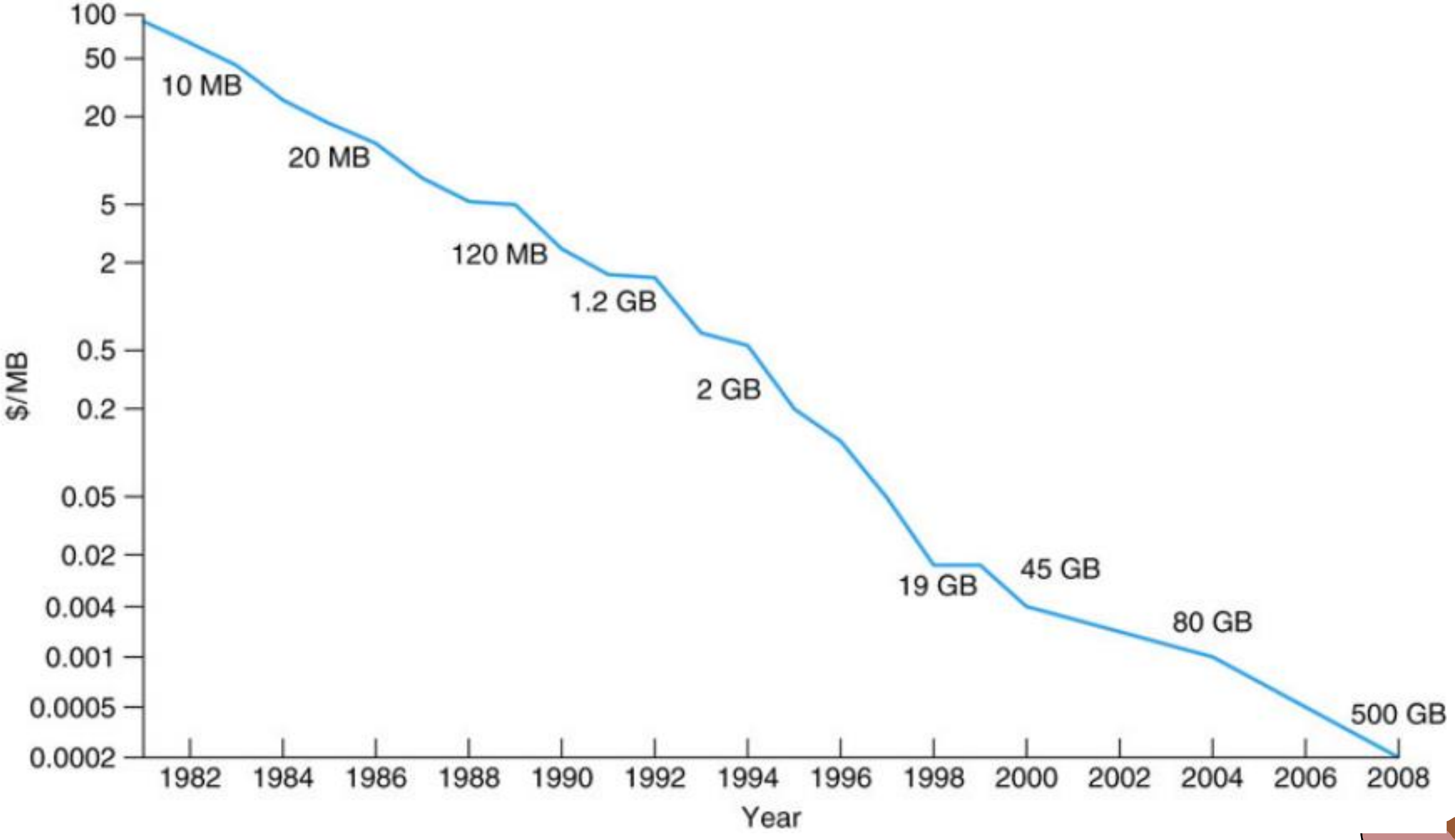
Chapter 12: Mass-Storage Systems

- Overview of Mass Storage Structure
- Disk Structure
- Disk Attachment
- Disk Scheduling
- Disk Management
- Swap-Space Management
- RAID Structure
- Stable-Storage Implementation

DRAM Price (1981 – 2008)



Magnetic Hard Disk Price (1981 – 2008)

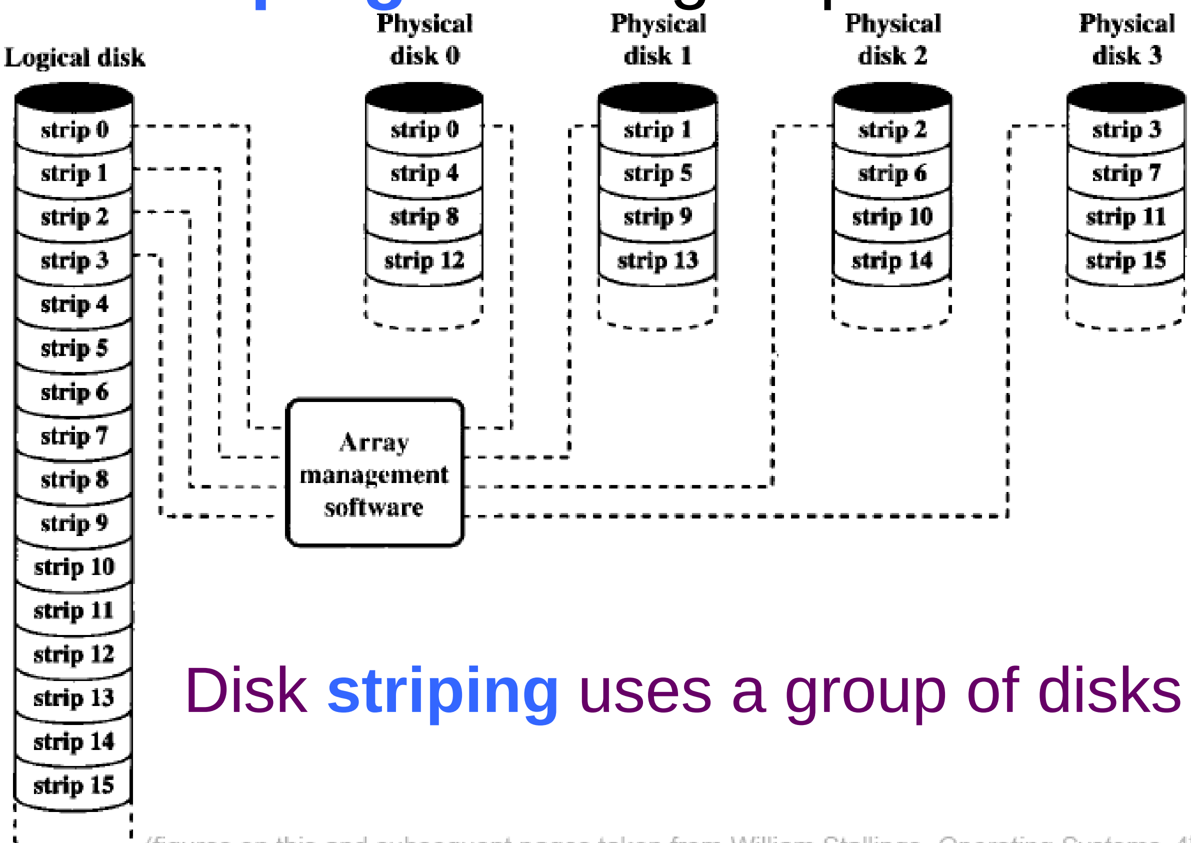


RAID Structure

- **RAID** = **R**edundant **A**rrays of **I**nexpensive **D**isks
  - multiple disk drives provides reliability via **redundancy**
- Increases the **mean time to failure**
- **Mean time to repair –** exposure time when another failure could cause data loss
- **Mean time to data loss (MTTDL)** based on above factors. MTTDL is an estimate of the expected time that it would take a given storage system to exhibit enough failures such that at least one block of data cannot be retrieved or reconstructed.
- If mirrored disks fail independently, consider disk with 100,000 hour mean time to failure and 10 hour mean time to repair
  - MTTDL is $100,000^2 / (2 * 10) = 500 * 10^6$ hours= 57,000 years!

RAID Structure

- RAID is arranged into six different levels
- 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

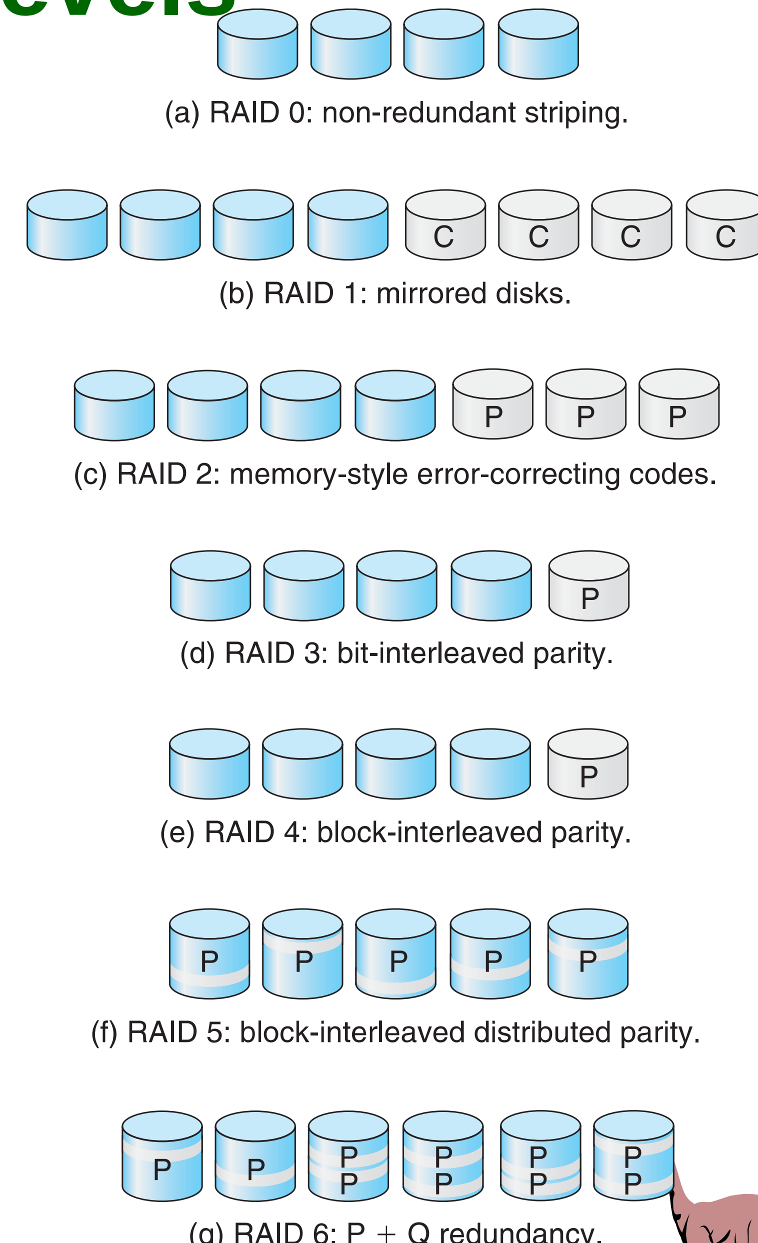


Disk **striping** uses a group of disks as one storage unit

(figures on this and subsequent pages taken from William Stallings, *Operating Systems*, 4th ed)

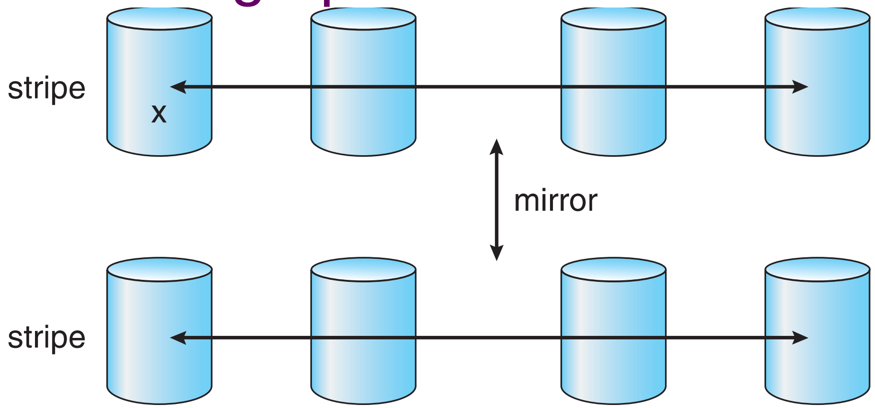
RAID Levels

- 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
  - **Block interleaved parity** (**RAID 4, 5, 6**) uses much less redundancy

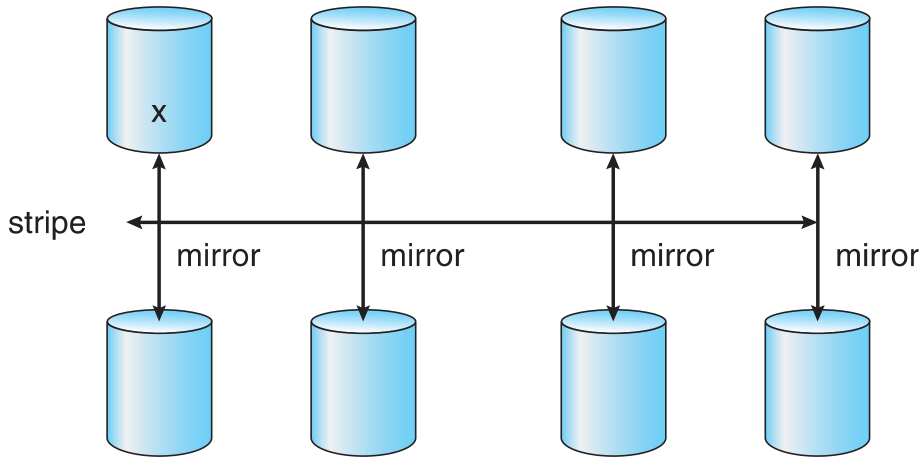


RAID (0 + 1) and (1 + 0)

Striped mirrors (**RAID 1+0**) or mirrored stripes (**RAID 0+1**) provides high performance and high reliability



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



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

RAID (Cont.)

- 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

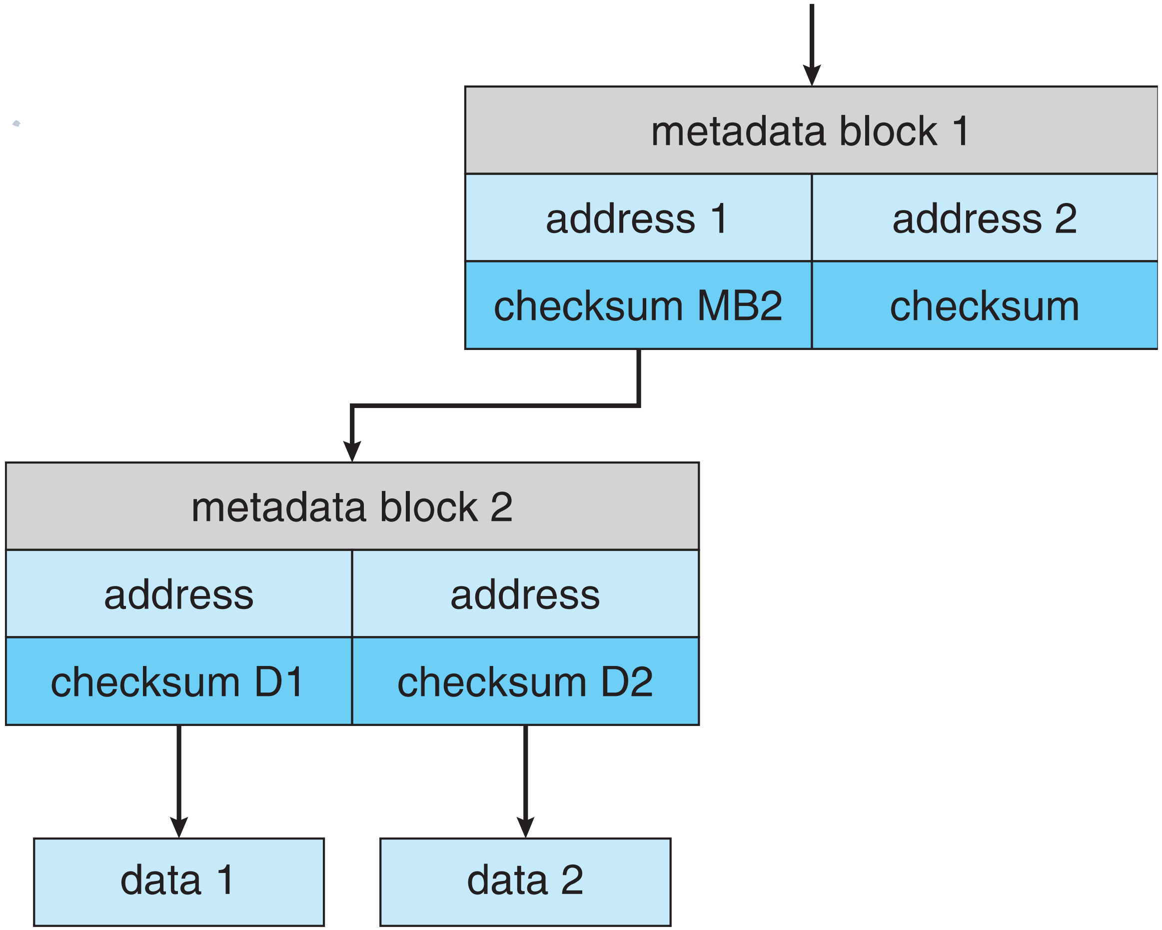
Other Features

- Regardless of where RAID implemented, other useful features can be added
- **Snapshot** is a view of file system before a set of changes take place (i.e. at a point in time)
  - More in Ch 12
- Replication is automatic duplication of writes between separate sites
  - For redundancy and disaster recovery
  - Can be synchronous or asynchronous
- Hot spare disk is unused, automatically used by RAID production if a disk fails to replace the failed disk and rebuild the RAID set if possible

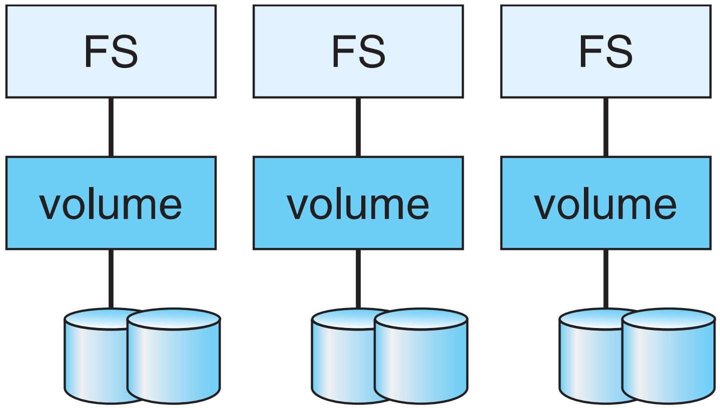
Extensions

- RAID alone does not prevent or detect data corruption or other errors, just disk failures
- Solaris ZFS adds **checksums** of all data and metadata
- Checksums kept with pointer to object, to detect if object is the right one and whether it changed
- Can detect and correct data and metadata corruption
- ZFS also removes volumes, partitions
  - Disks allocated in **pools**
  - Filesystems with a pool share that pool, use and release space like **`malloc()`** and **`free()`** memory allocate / release calls

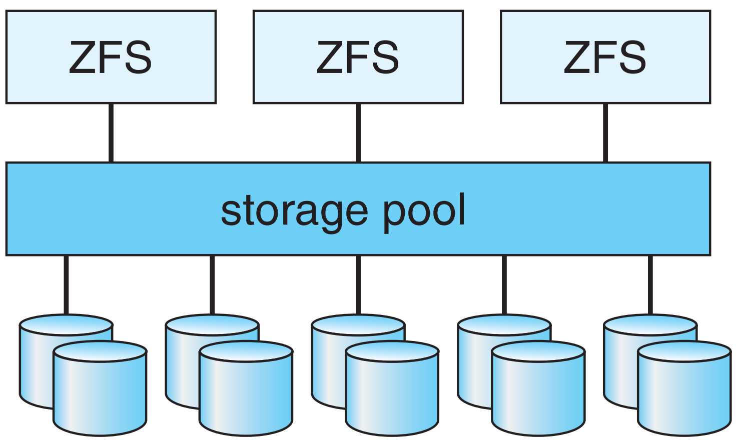
ZFS Checksums All Metadata and Data



Traditional and Pooled Storage



(a) Traditional volumes and file systems.



(b) ZFS and pooled storage.