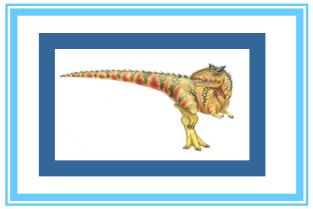
Chapter 12: Mass-Storage Systems

大容量存储系统



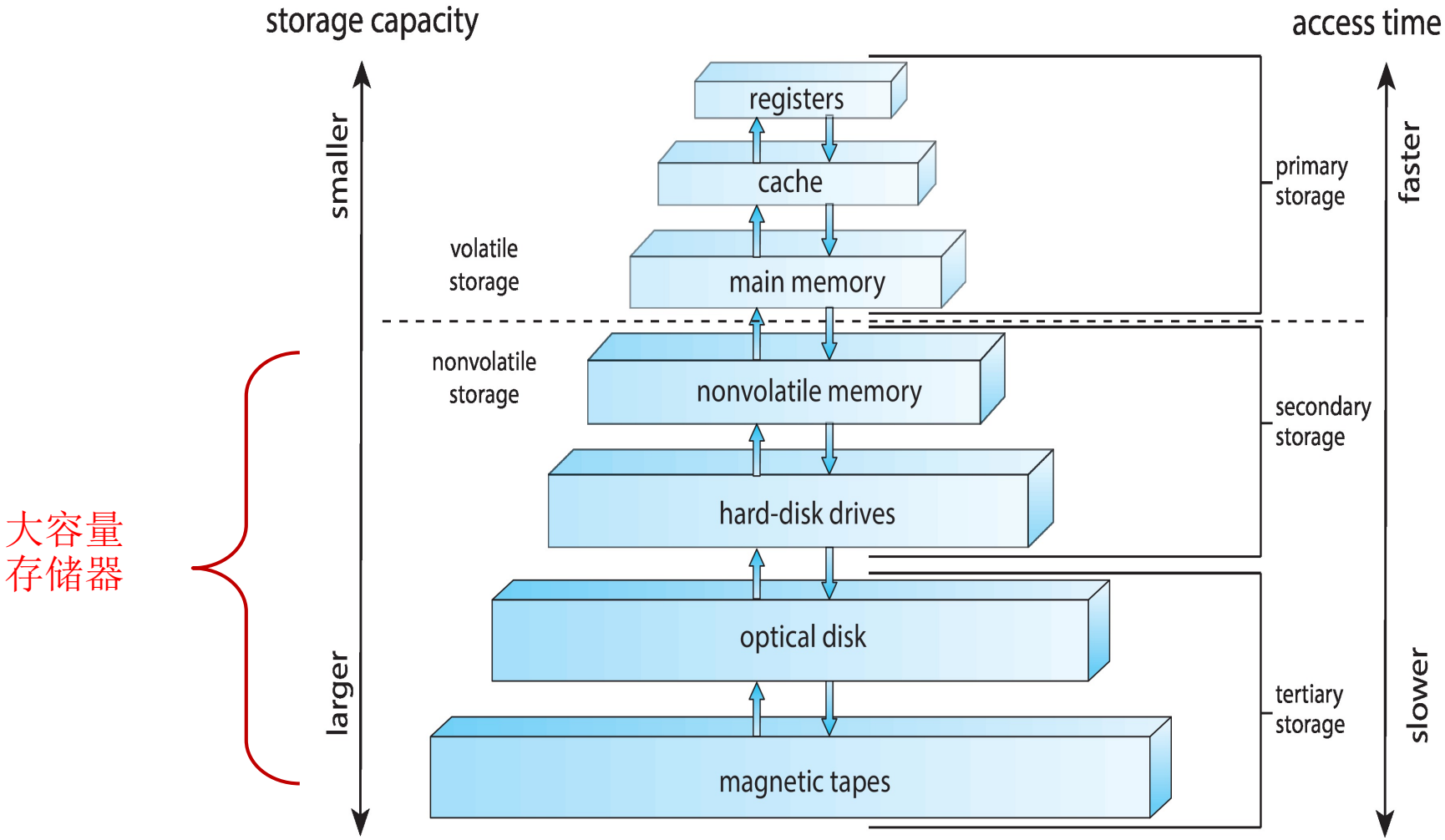
Chapter 12: Mass-Storage Systems

- 12.1 Overview of Mass Storage Structure
- 12.2 Disk Structure
- 12.3 Disk Attachment
- 12.4 Disk Scheduling
- 12.5 Disk Management
- 12.6 Swap-Space Management
- 12.7 RAID Structure
- 12.8 Stable-Storage Implementation
- 12.9 Tertiary Storage Devices
- 12.10 Summary

Objectives

- Describe the physical structure of secondary and tertiary storage devices and the resulting effects on the uses of the devices
- Explain the performance characteristics of mass-storage devices
- Discuss operating-system services provided for mass storage, including RAID and HSM

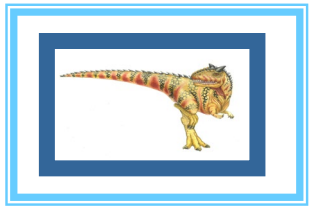
Hierarchical Storage Architecture



Hierarchical Storage Architecture

- 智能时代，数据增长的速度将超过摩尔定理
- 存储场景
  - 全世界总人口超过73亿，亚洲总人口超过 40亿。
  - **互联网用户超过40亿;70亿手机用户。**
  - 2016年全球数据总量12ZB，预计到2020年将到达44ZB(相当于全球每人5.2TB)，2025年甚至达到163ZB。
  - 所有数据中约42%是重复的（至少重复一次）。
  - 所有数据中约33%是压缩的（2次）。
  - 所有数据中约5%是瞬态的（临时的）。
  - 所有数据中7%是非结构化结构（很难浏览）。

12.1Overview of Mass Storage Structure



Overview of Mass Storage Structure

- 分层存储体系

Data Aging Profile

Age in days	Probability of Re-use	Avg. Data By Tier	Key Applications
1	70-80%	5%	Very high performance apps, critical data and OLTP.
3	40-60%	15%	Mission-critical, OLTP, revenue generating, high performance apps.
7	20-25%	20%	Backup/recovery, vital and sensitive data, moderate performance and big data.
+30	1-5%	60%	Backup, recovery, archive, long-term retention, big data and disaster recovery

Archive-as-a-Service (AaaS),
Disaster-Recovery-as-a-Service (DRaaS),
Backup Recovery-as-a-Service (BRaaS).

Overview of Mass Storage Structure

- Magnetic disks（磁盘） provide bulk of secondary storage of modern computers
  - 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)
  - Head crash results from disk head making contact with the disk surface
    - That's bad
- Disks can be removable
- Drive attached to computer via I/O bus
  - Busses vary, including EIDE, ATA, SATA, USB, Fibre Channel, SCSI
  - Host controller in computer uses bus to talk to disk controller built into drive or storage array

Overview of Mass Storage Structure (Cont)

■ **固态驱动器（Solid State Drives）**，称固态硬盘，固态硬盘用固态电子存储芯片阵列制成的硬盘，由控制单元和存储单元（**FLASH芯片**、DRAM芯片）组成。

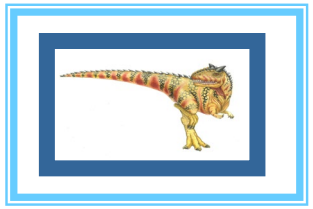
- 第一只SSD出现在1978年（STK 4305，每MB售价8800美元，DRAM）。
- 全闪存阵列（AFAS）和混合闪存阵列（HFA）呈爆发式增长。
- 现在SSD的容量是15.36TB（SAS）。
- 非易失性、低功耗（只有HDD的三分之一）。
- 无活动部件、可靠性高——位误码率（BER）1 x 1017
- 读取存取时间：0.2毫秒，存取时间比HDD大概快 50倍。

Overview of Mass Storage Structure (Cont)

Magnetic tape （磁带）

- 出货的磁带驱动器中超过85%是LTO（Linear Tape Open）。
- 磁带驱动器的可靠性、数据传输速率和容量已超过磁盘。
- 磁带的原生容量为10TB，压缩容量超过25TB。(LTO-10:48TB)
- 磁带的原生数据传输速率为360MB/s。
- **LTFS(Liner Tape File System)为磁带提供了一种通用、开放的文件系统。**
- 由于总体拥有成本，云采用磁带解决方案用于归档服务。
- 对企业级磁带和LTO而言，磁带介质的寿命至少是30年。

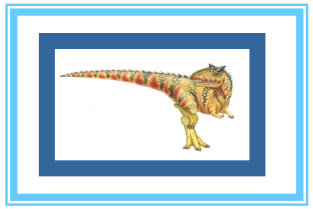
12.2 Disk Structure



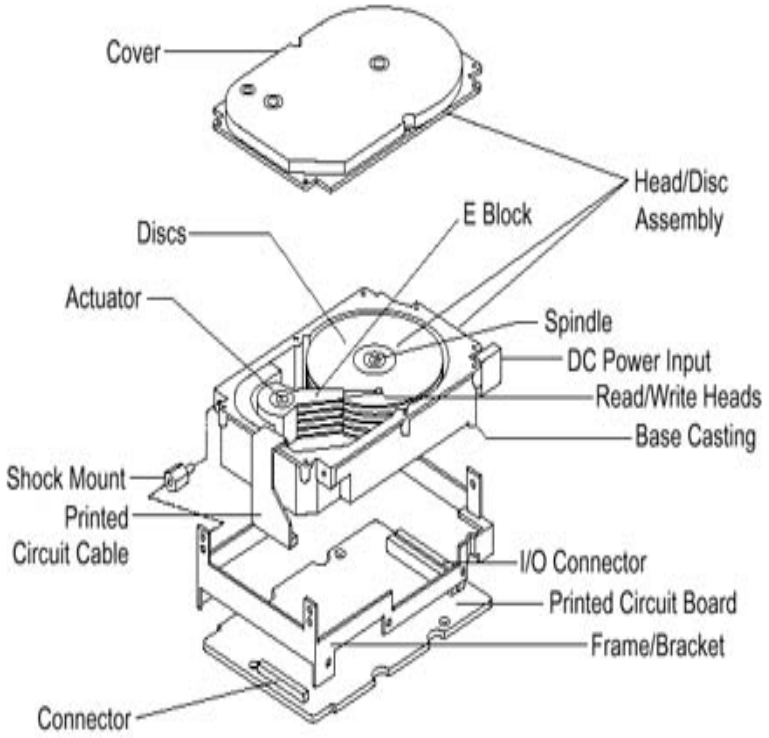
Disk Structure

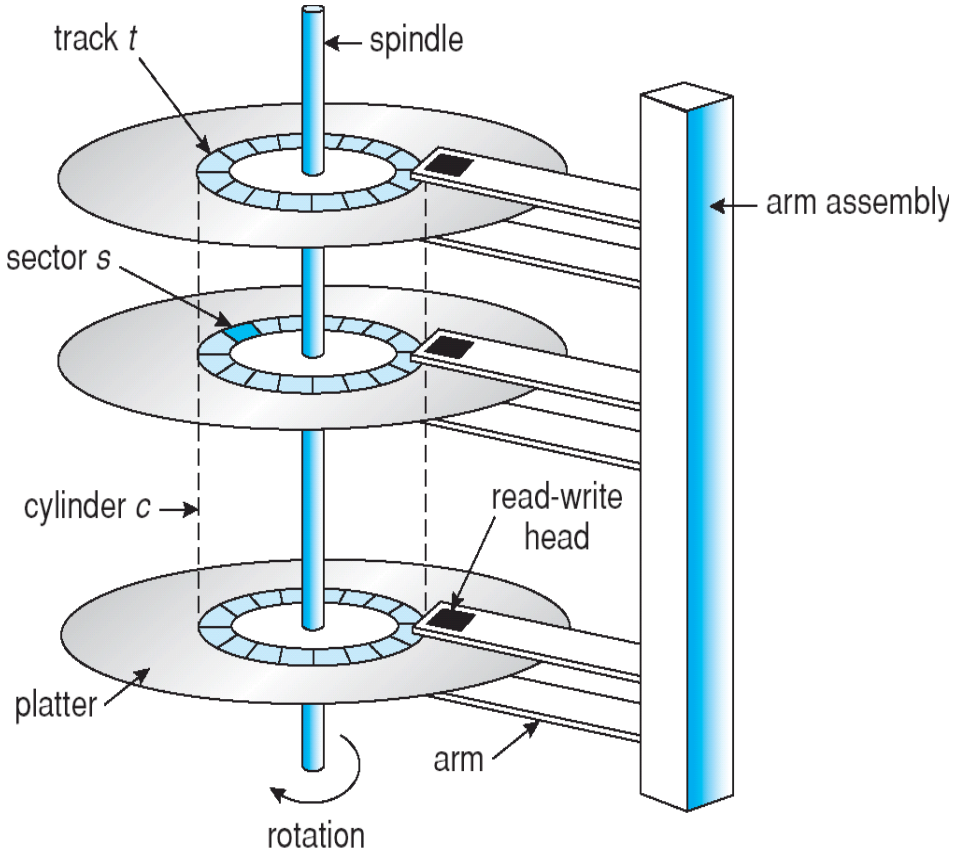
- Disk drives are addressed as large 1-dimensional arrays of logical blocks, where the logical block is the smallest unit of transfer
- The size of the logical block is usually 512 bytes
- The 1-dimensional array of logical blocks is mapped into the sectors of the disk sequentially
  - Sector 0 is the 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, and then through the rest of the cylinders from outermost to innermost

12.3 Disk Attachment



Moving-head Disk Mechanism



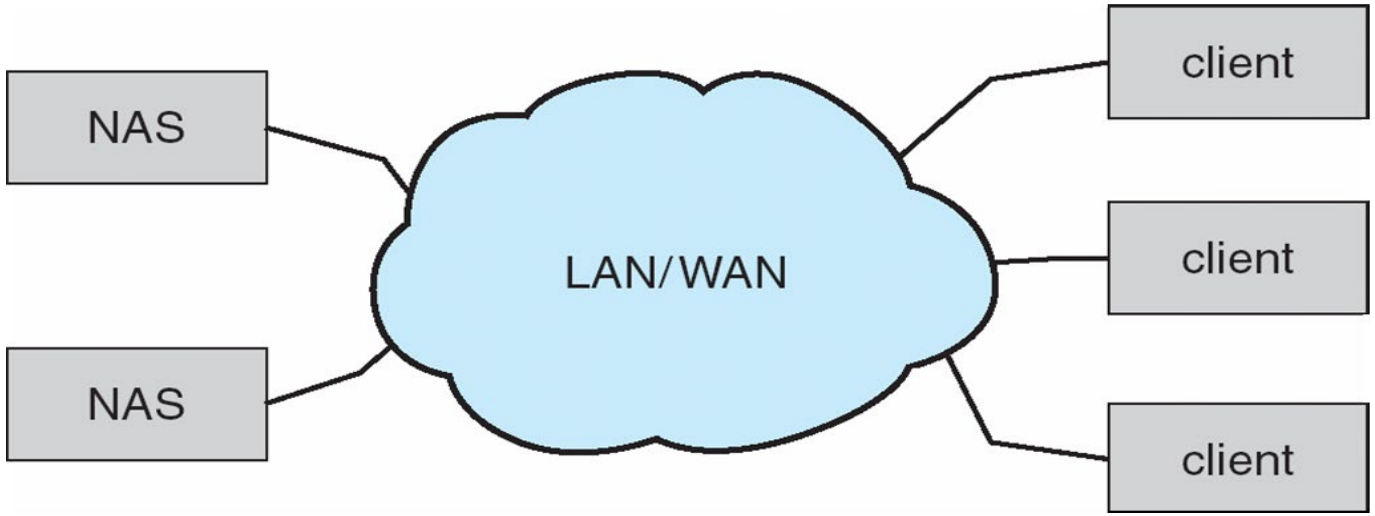


Host-attached storage

- Host-attached storage accessed through I/O ports talking to I/O busses
- I/O bus like IDE
  - a maximum of 2 drives per I/O bus
- 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（Fibre Channel，光纤通道） 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
  - Can be arbitrated loop (FC-AL) of 126 devices

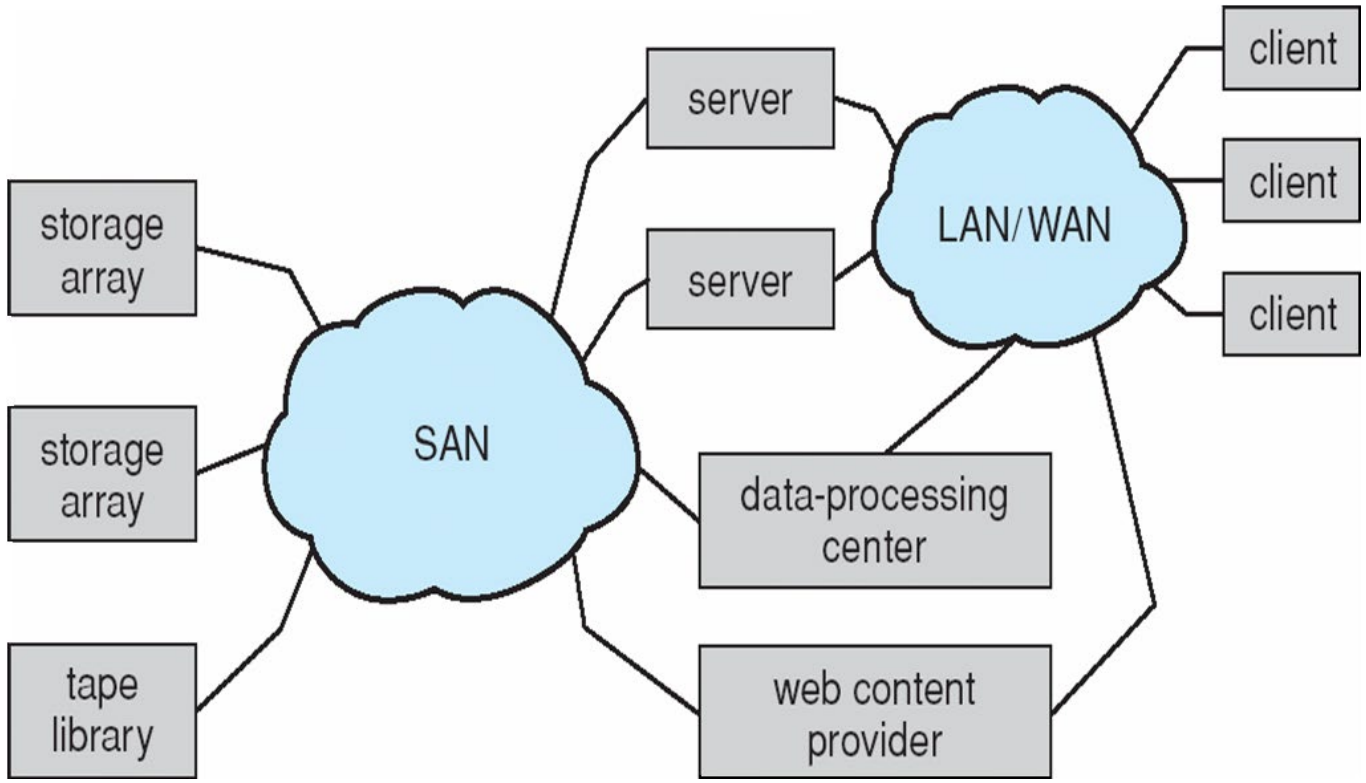
Network-Attached Storage

- Network-attached storage (NAS) is storage made available over a network rather than over a local connection (such as a bus)
- NFS and CIFS are common protocols
- Implemented via remote procedure calls (RPCs) between host and storage
- New iSCSI protocol uses IP network to carry the SCSI protocol



Storage Area Network（SAN）

- Common in large storage environments (and becoming more common)
- Multiple hosts attached to multiple storage arrays - flexible



存储虚拟化技术

- SNIA（Storage Networking Industry Association，存储网络联合会）官方对于Virtualization（存储虚拟化技术）的定义，如下：
  - 是将存储（子）系统内部功能与具体应用、主机及通用网络资源分离、隐藏及抽象的行为。以期达到存储或数据管理的网络无关性。
  - 对于存储服务及设备的虚拟化应用，以期达到整合设备功能、隐藏复杂细节以及向已经存在的底层存储资源添加新的应用。

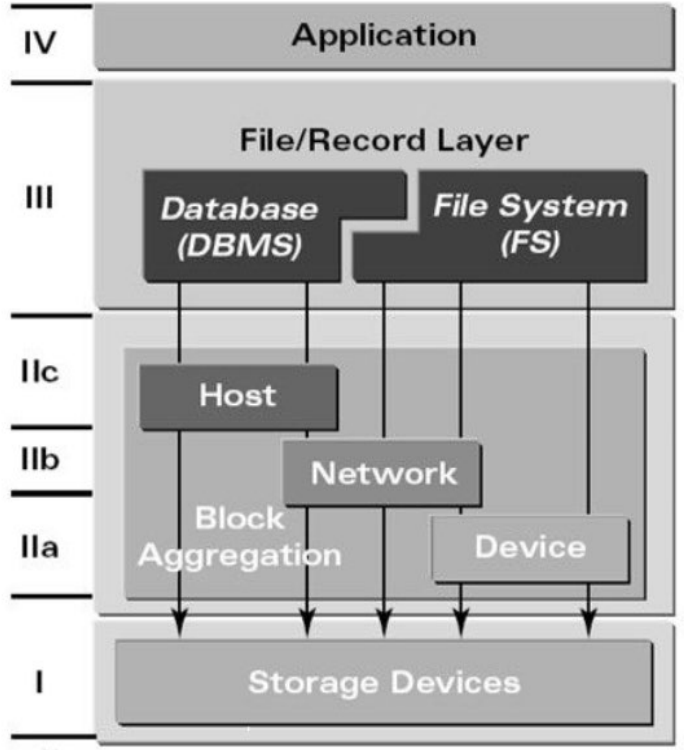
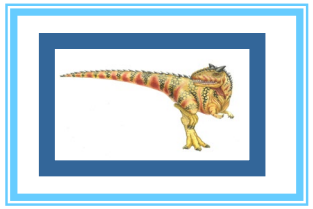


图 1 SNIA 共享存储模型

12.4 Disk Scheduling



Disk Scheduling

- The operating system is responsible for using hardware efficiently — for the disk drives, this means having a fast access time and disk bandwidth
- Access time has three major components
  - **Seek time（寻道时间）** is the time for the disk are to move the heads to the cylinder containing the desired sector
  - **Rotational latency（旋转延迟）**is the additional time waiting for the disk to rotate the desired sector to the disk head
  - **Transfer time（传输时间）**
- Minimize seek time
- **Seek time ≈ seek distance 寻道时间 ≈ 寻道距离**
- 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

数据计算

- 7200(转／每分钟)的硬盘，每旋转一周所需时间为60×1000（毫秒）÷7200＝8.33毫秒，
  则平均旋转延迟时间为8.33÷2＝4.17毫秒(平均情况下，需要旋转半圈)。

- 7200转机械硬盘的寻道时间一般为12-14毫秒，固态硬盘可以达到0.1毫秒甚至更低。
- 固态硬盘持续读写速度超过500MB/s
- 机械硬盘读写速度超过50~200MB/s（接口不同）
- 磁带的原生数据传输速率为360MB/s。

Disk Scheduling (Cont)

- Several algorithms exist to schedule the servicing of disk I/O requests
- 常用的磁盘调度算法有：先来先服务(FCFS)、最短寻道时间优先(SSTF)、扫描(SCAN)算法和循环扫描(C-SCAN)算法等

- We illustrate them with a request queue (0-199)

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

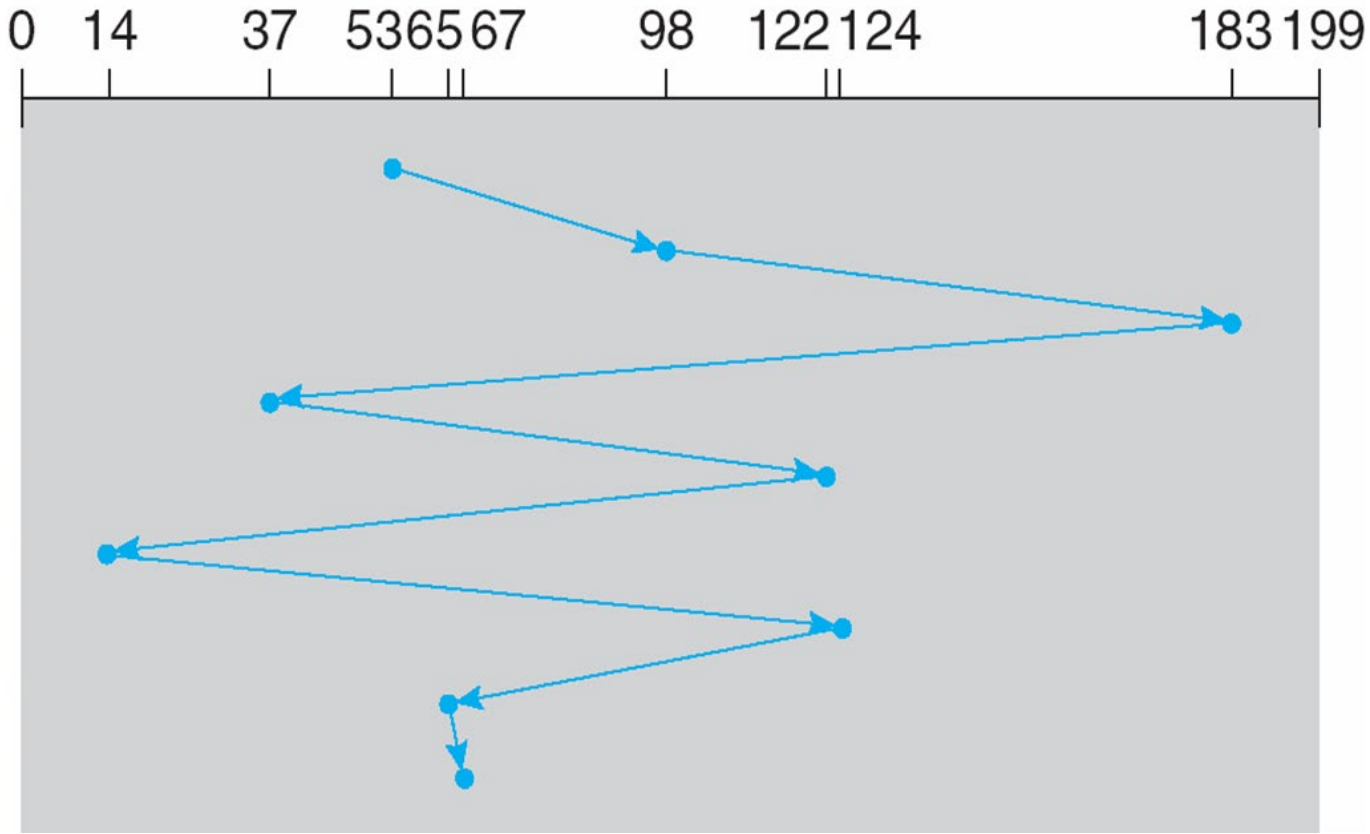
  Head pointer 53

FCFS 先来先服务

Illustration shows total head movement of 640 cylinders

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

head starts at 53



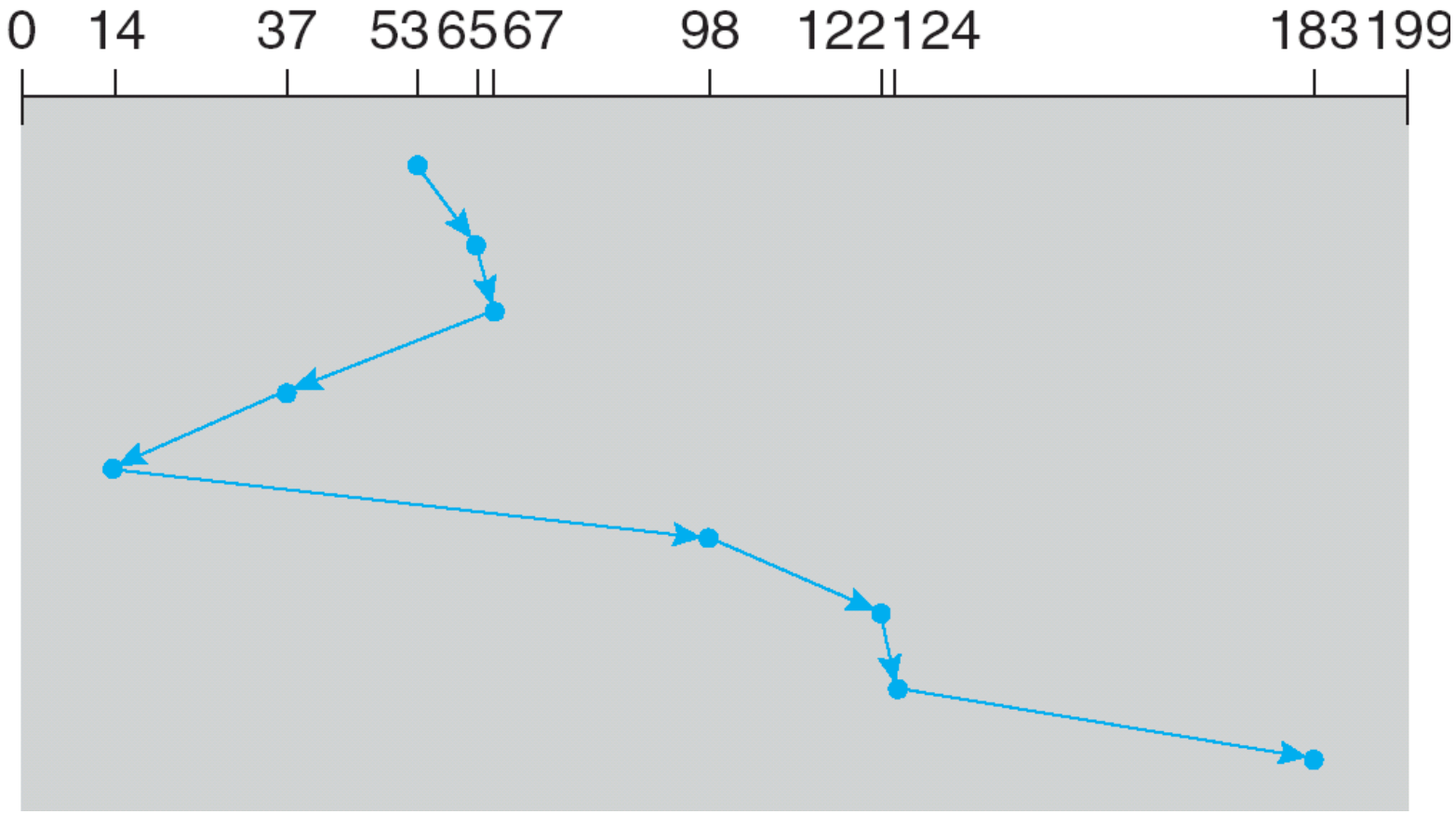
SSTF 最短寻道时间优先

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



total head movement of 236 cylinders

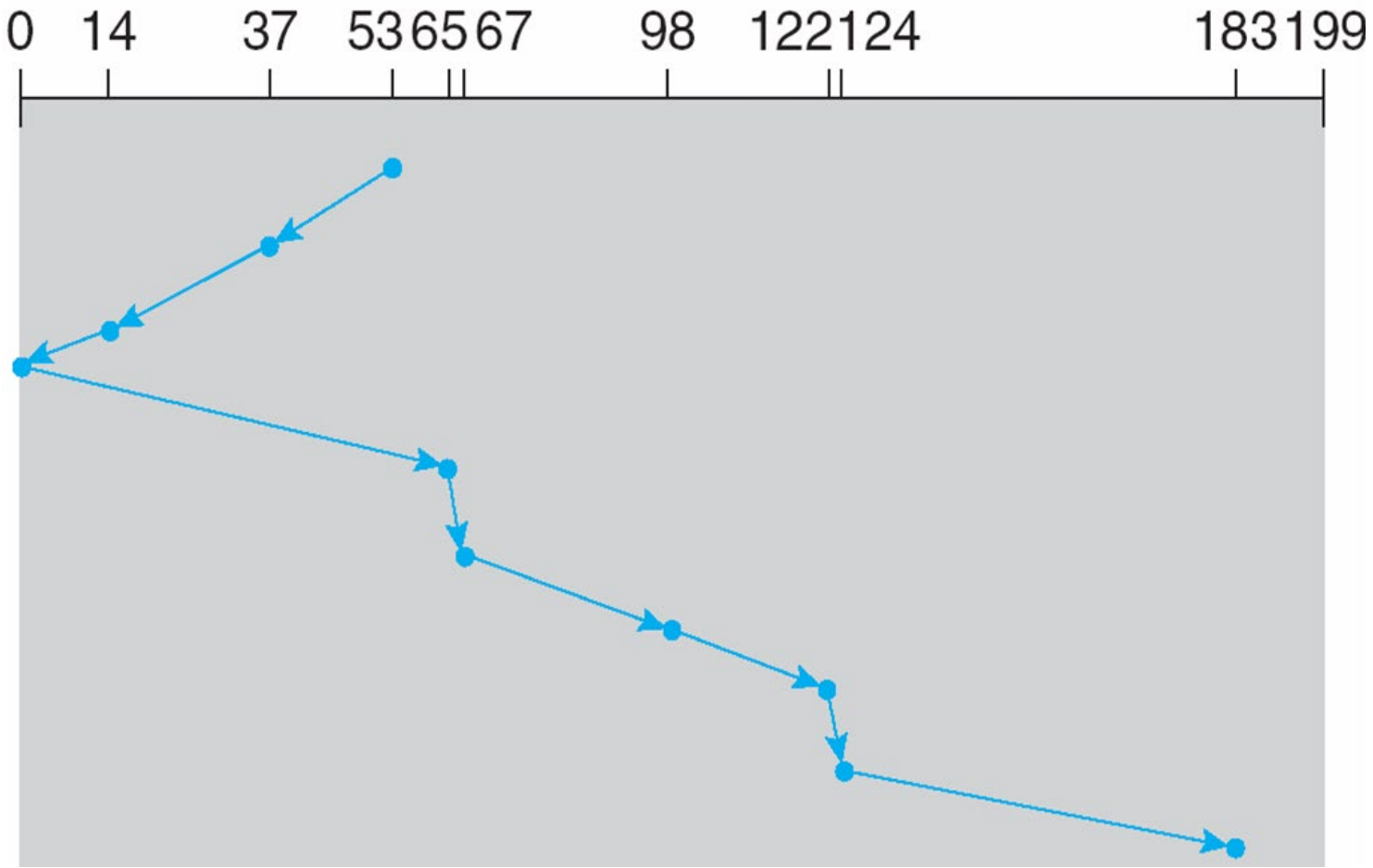
SCAN扫描

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

SCAN (Cont.)

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

head starts at 53



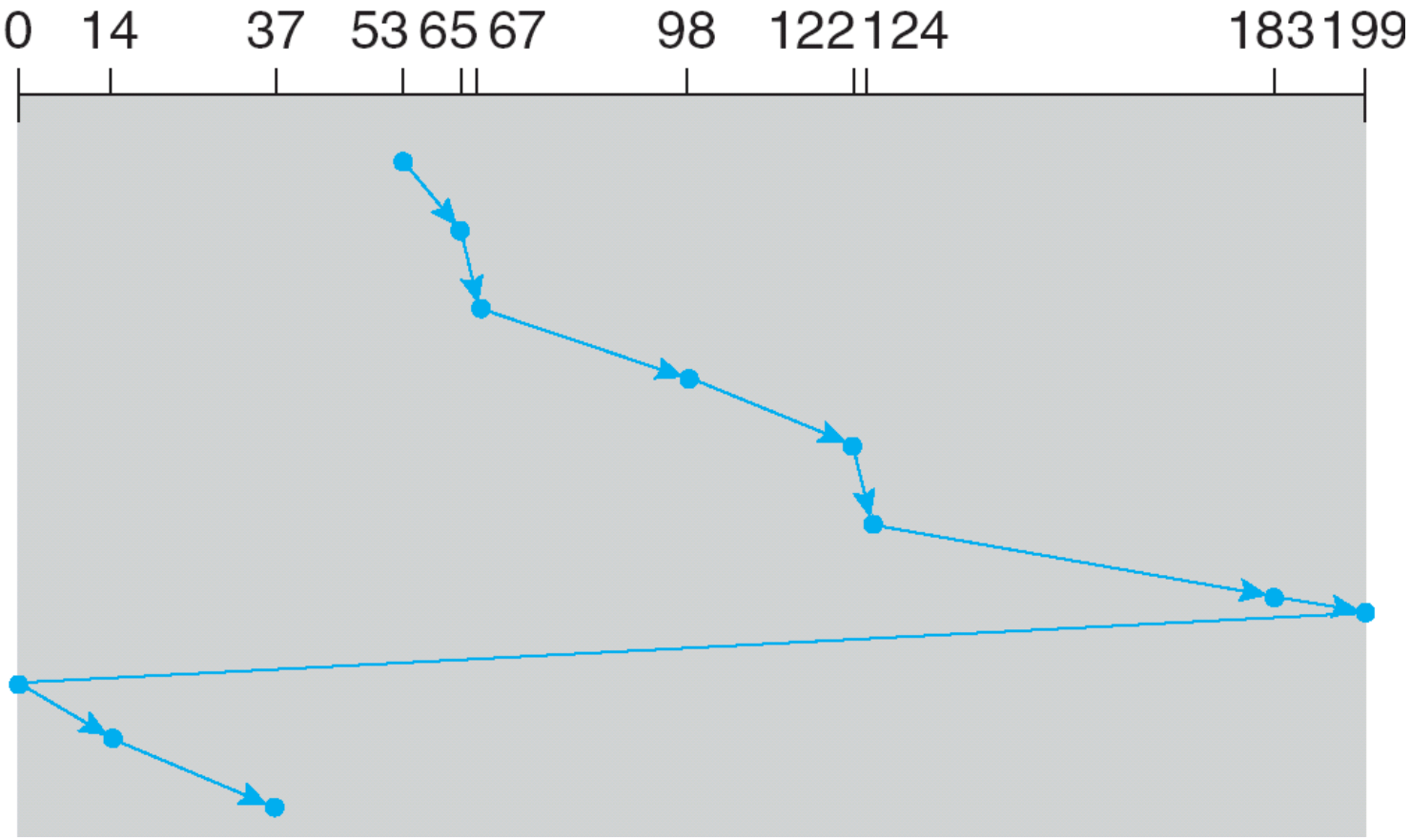
C-SCAN

- 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, 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)

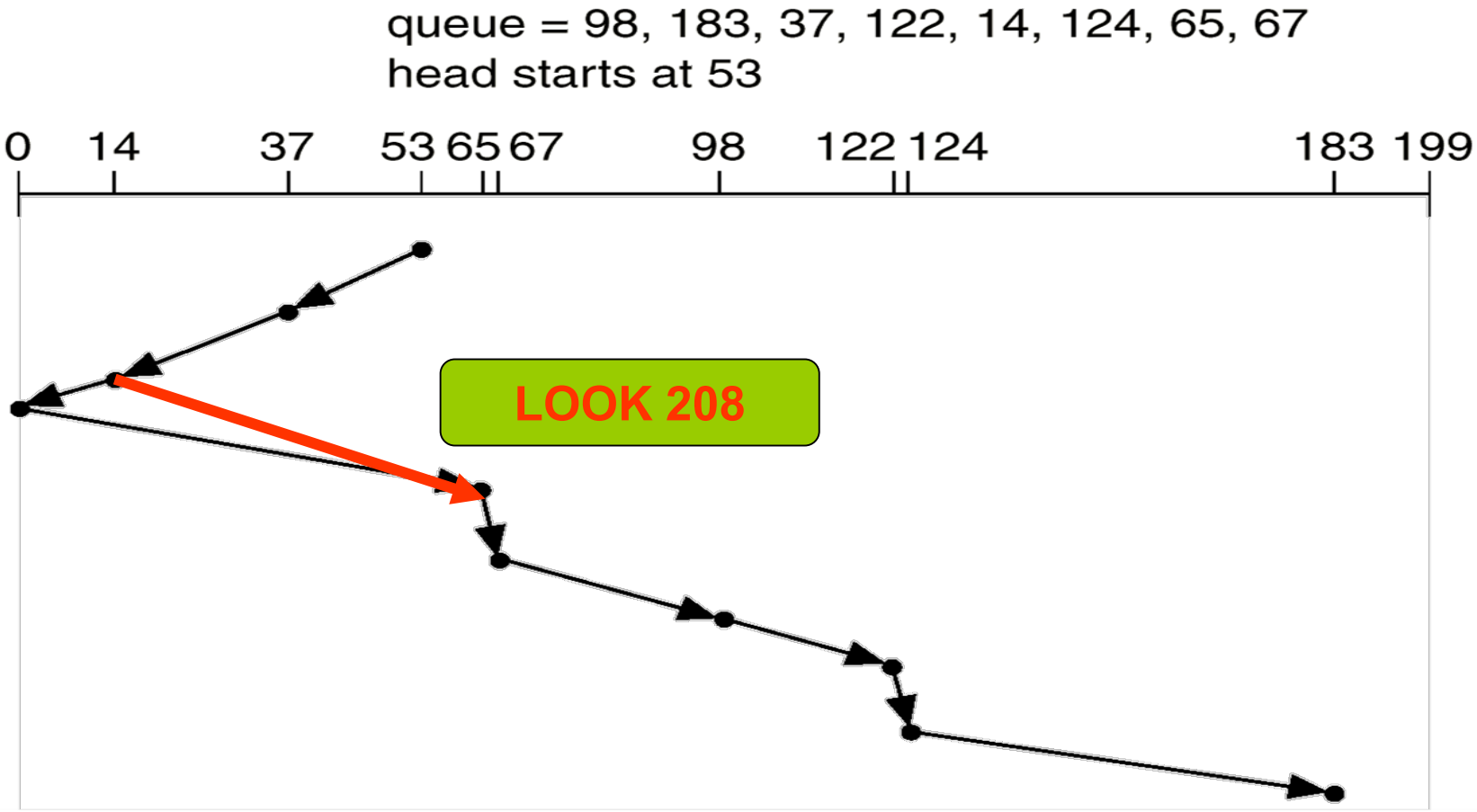
queue = 98, 183, 37, 122, 14, 124, 65, 67

head starts at 53



total head movement of 382 cylinders

LOOK-- Version of SCAN

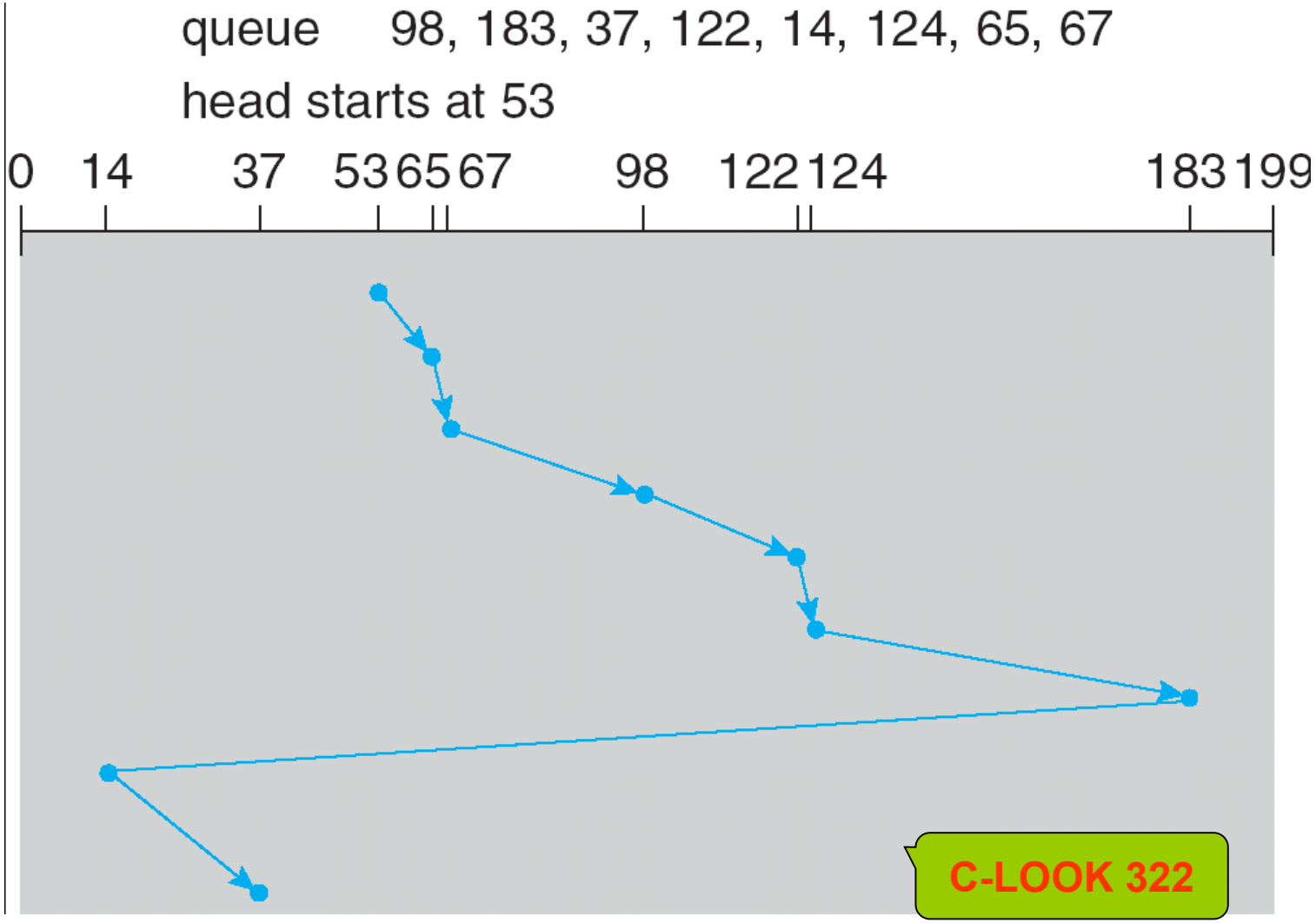


total head movement of 208 cylinders

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

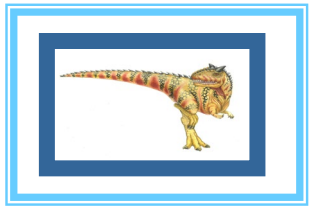
C-LOOK (Cont)


queue 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0 14 37 53 65 67 98 122 124 183 199
C-LOOK 322

Selecting 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
- Performance depends on the number and types of requests
- Requests for disk service can be influenced by the file-allocation method
- 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

12.5 Disk Management



Disk Management

- Low-level formatting（低级格式化）, 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 （逻辑格式化）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

mkfs.ext2

Boot Block 启动块

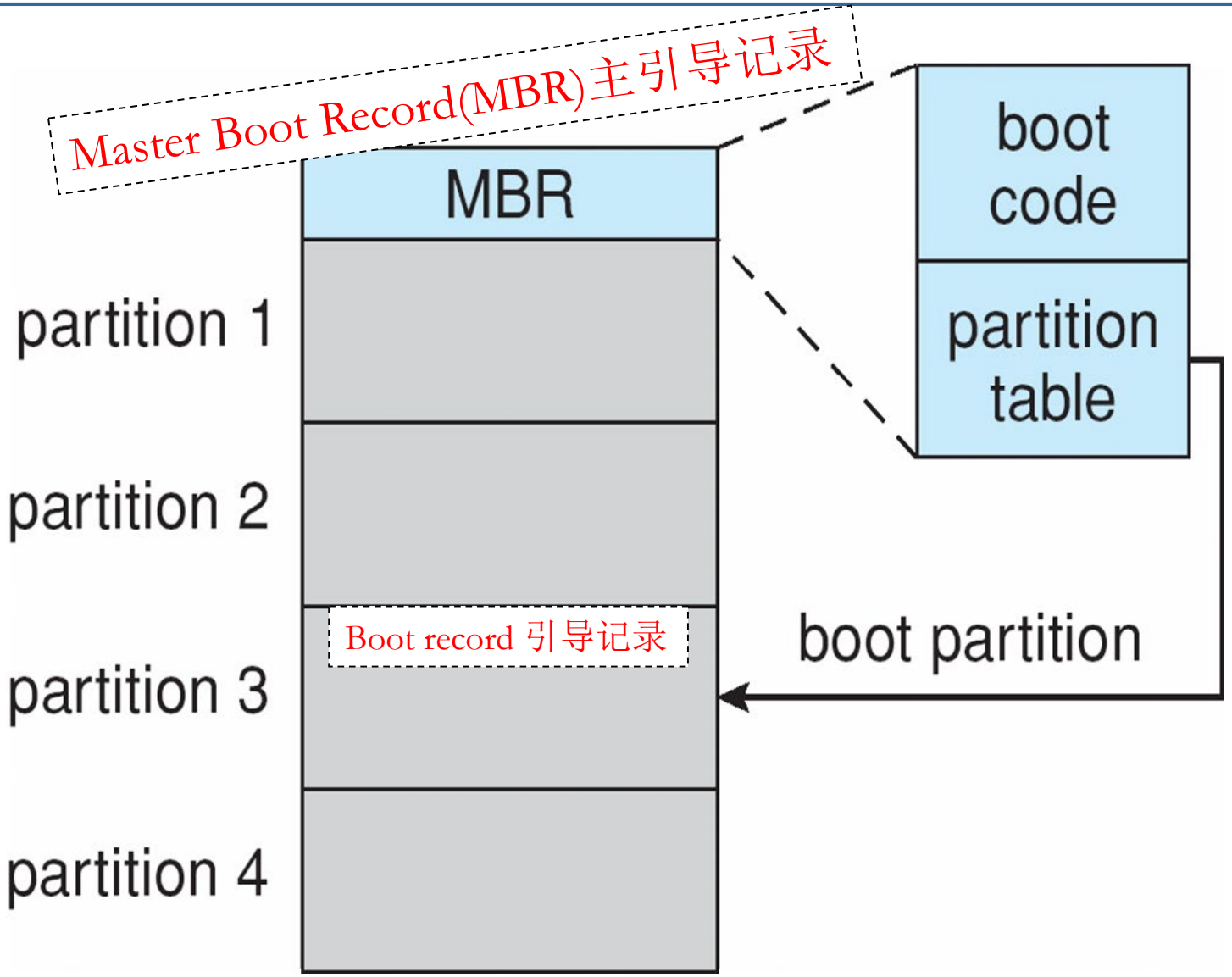
- Boot block initializes system
  - The bootstrap is stored in ROM
  - Bootstrap loader program

- Typical boot sequence
  - code (simple bootstrap) in ROM
  - code (full bootstrap) in boot block
    - the bootstrap loader, e.g. Grub or LILO
  - the entire kernel of the operating systems

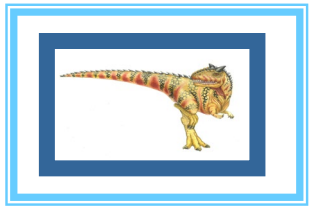
Bad Block 坏块

- Disks frequently have defective blocks or bad blocks
- 坏块的处理方法
- MS-DOS的处理方法: format,chkdsk命令

Booting from a Disk in Windows 2000



12.6 Swap-Space Management



Swap-Space Management

- Swap-space — Virtual memory uses disk space as an extension of main memory
- Swap-space can be carried out in two forms：
  - in the normal file system
    e.g. Windows family

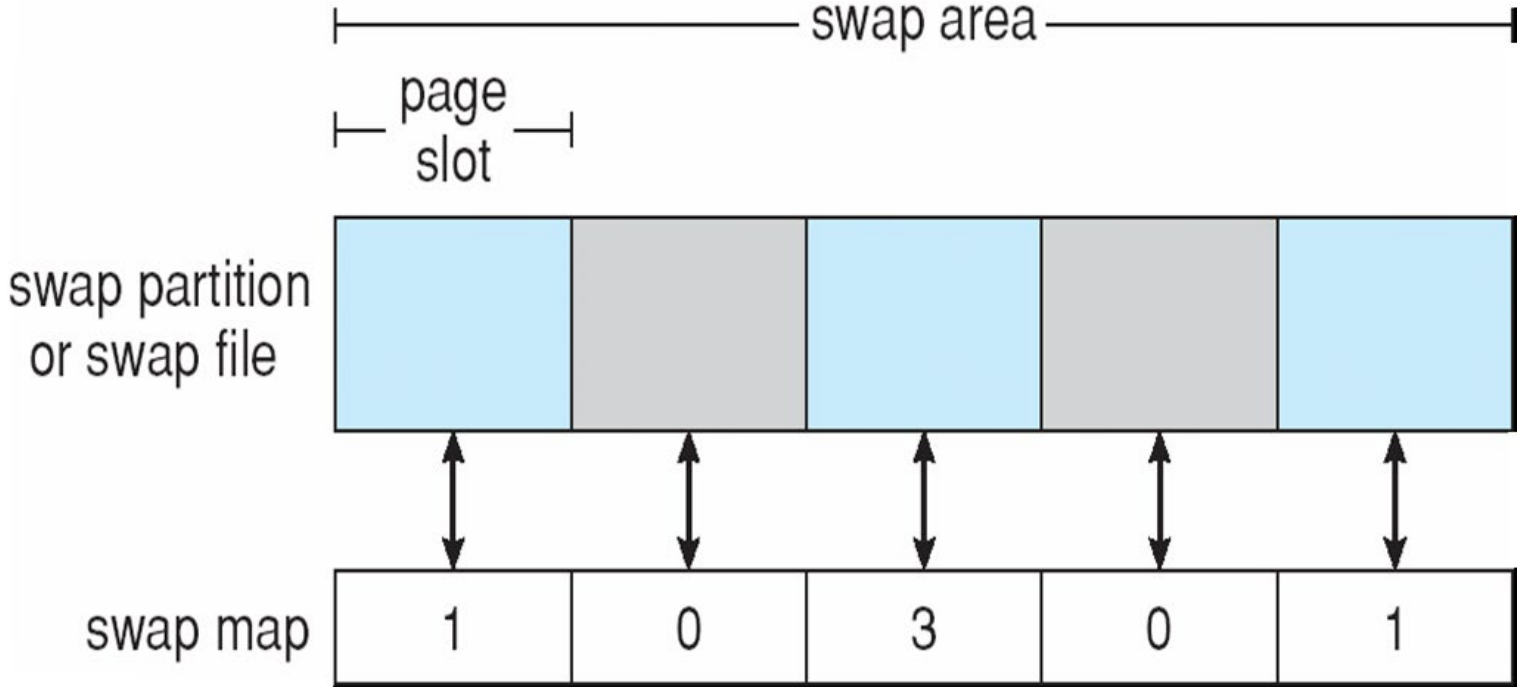
    pagefile.sys文件
  - in a separate disk partition
    e.g. Linux、Unix、solaris

    SWAP分区

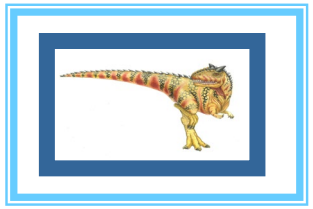
Swap-space management

- 4.3BSD allocates swap space when process starts; holds text segment (the program) and data segment
- Kernel uses swap maps to track swap-space use
- Solaris 2 allocates swap space only when a page is forced out of physical memory, not when the virtual memory page is first created

Data Structures for Swapping on Linux Systems



12.7 RAID Structure



RAID Structure

- **RAID**：Redundant Arrays of Inexpensive (independent) Disks（冗余廉价磁盘阵列）.RAID是一种把多块独立的硬盘（物理硬盘）按不同的方式组合起来形成一个硬盘组（逻辑硬盘），从而提供比单个硬盘更高的存储性能和提供数据备份技术。
- Inexpensive -> Independent
- RAID – multiple disk drives provides reliability via redundancy
- Increases the mean time to failure
- Frequently combined with NVRAM to improve write performance
- RAID is arranged into six different levels（较早）

RAID (Cont)

- Several improvements in disk-use techniques involve the use of multiple disks working cooperatively
- Disk **striping**（条带化） 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（镜像） 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 (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

RAID Levels

(a) RAID 0: non-redundant striping.

P:纠错位

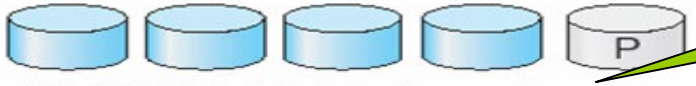
C:数据的第二拷贝

(b) RAID 1: mirrored disks.

Hamming码

RAID2、3按字节或位striping

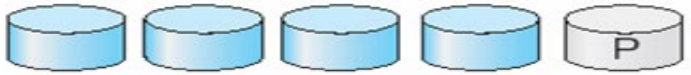
(c) RAID 2: memory-style error-correcting codes.



奇偶校验

(d) RAID 3: bit-interleaved parity.

按块striping



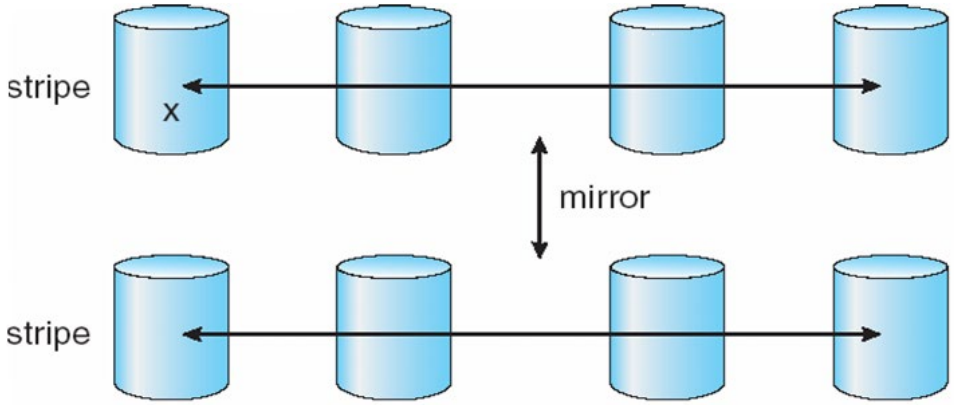
(e) RAID 4: block-interleaved parity.

(f) RAID 5: block-interleaved distributed parity.

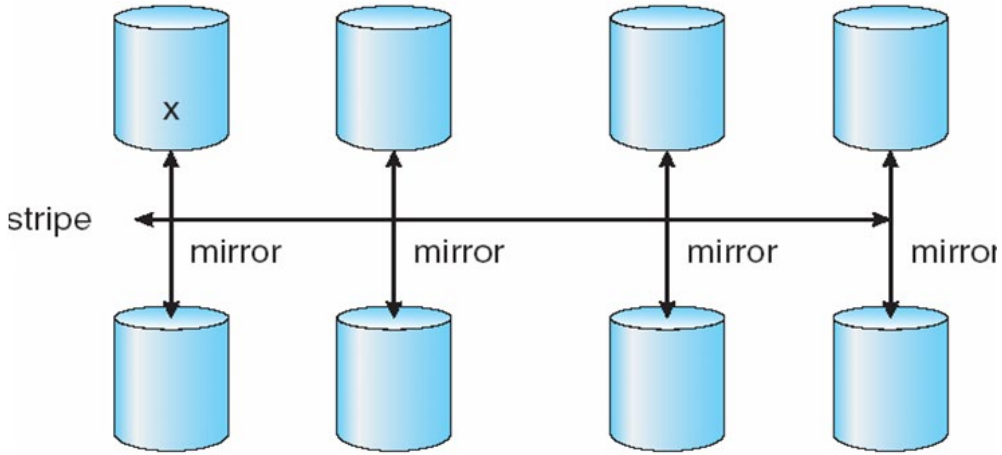
P+Q冗余，差错纠正码

(g) RAID 6: P + Q redundancy.

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



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

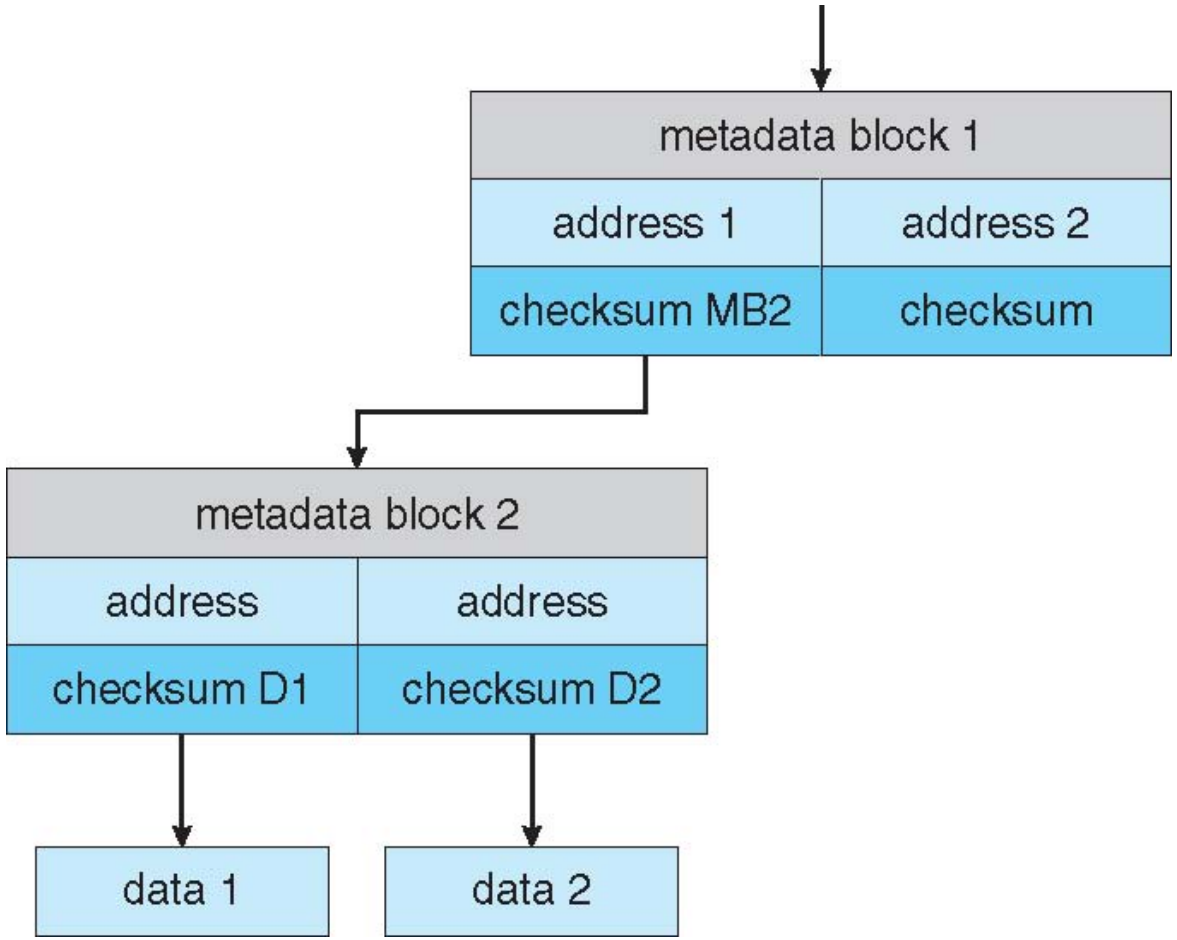


b) RAID 1 + 0 with a single disk failure.

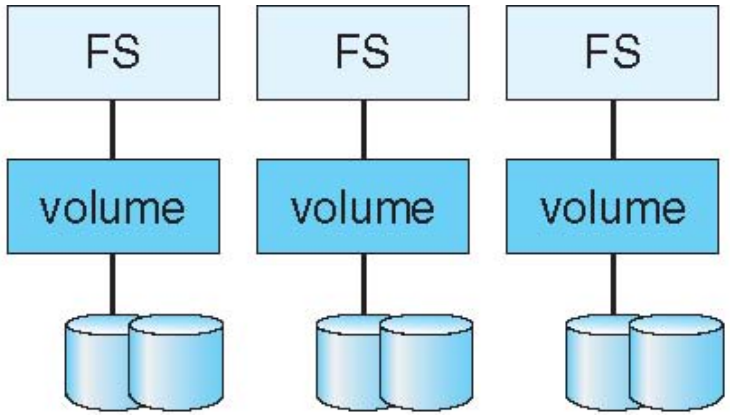
Extensions

- 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, partititions
  - Disks allocated in pools
  - Filesystems with a pool share that pool, use and release space like “malloc” and “free” memory allocate / release calls

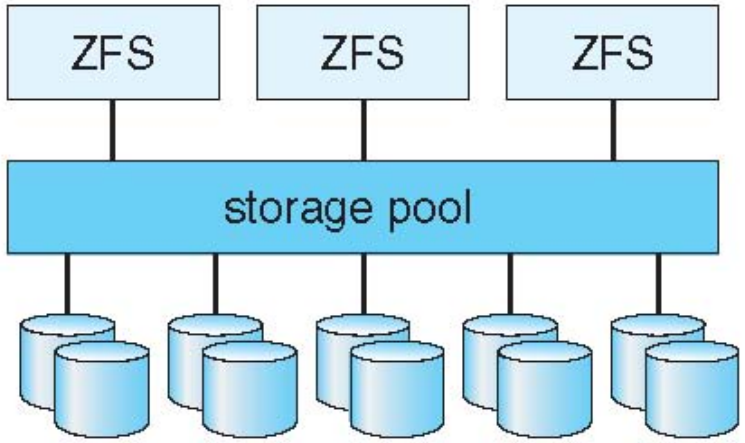
ZFS Checksums All Metadata and Data



Traditional and Pooled Storage



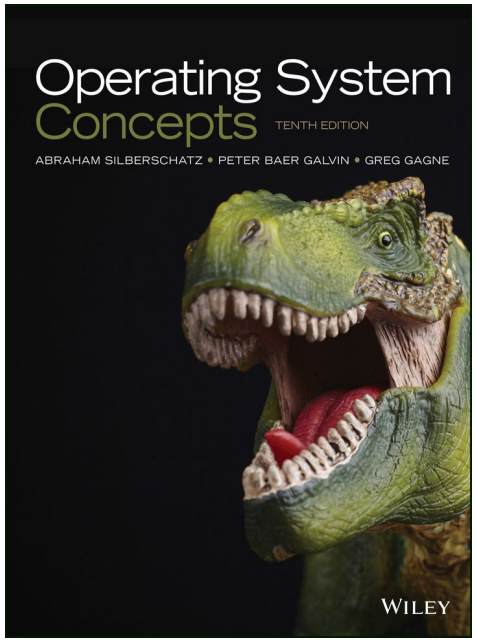
(a) Traditional volumes and file systems.

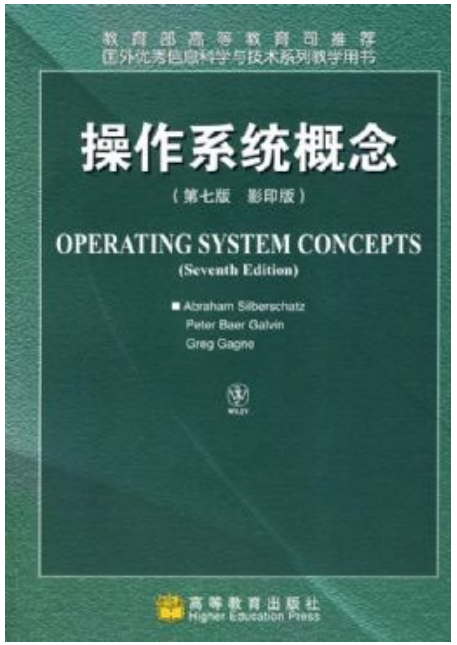
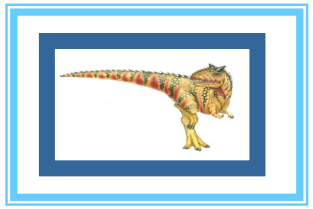


(b) ZFS and pooled storage.

Reading Assignments

- Read for this week:
  - Chapters 12
    of the text book:
- Read for next week:
  - Chapters 13
    of the text book:





End of Chapter 12