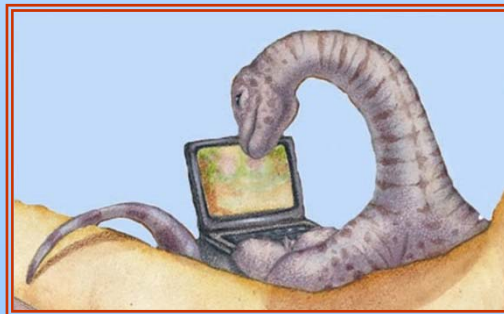


Chapter 12: Mass-Storage Systems





Chapter 12: Mass-Storage Systems

- Overview of Mass Storage Structure
- Disk Structure
- Disk Attachment
- Disk Scheduling
- Disk Management
- Swap-Space Management
- RAID Structure
- Disk Attachment
- Stable-Storage Implementation
- Tertiary Storage Devices
- Operating System Issues
- Performance Issues





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





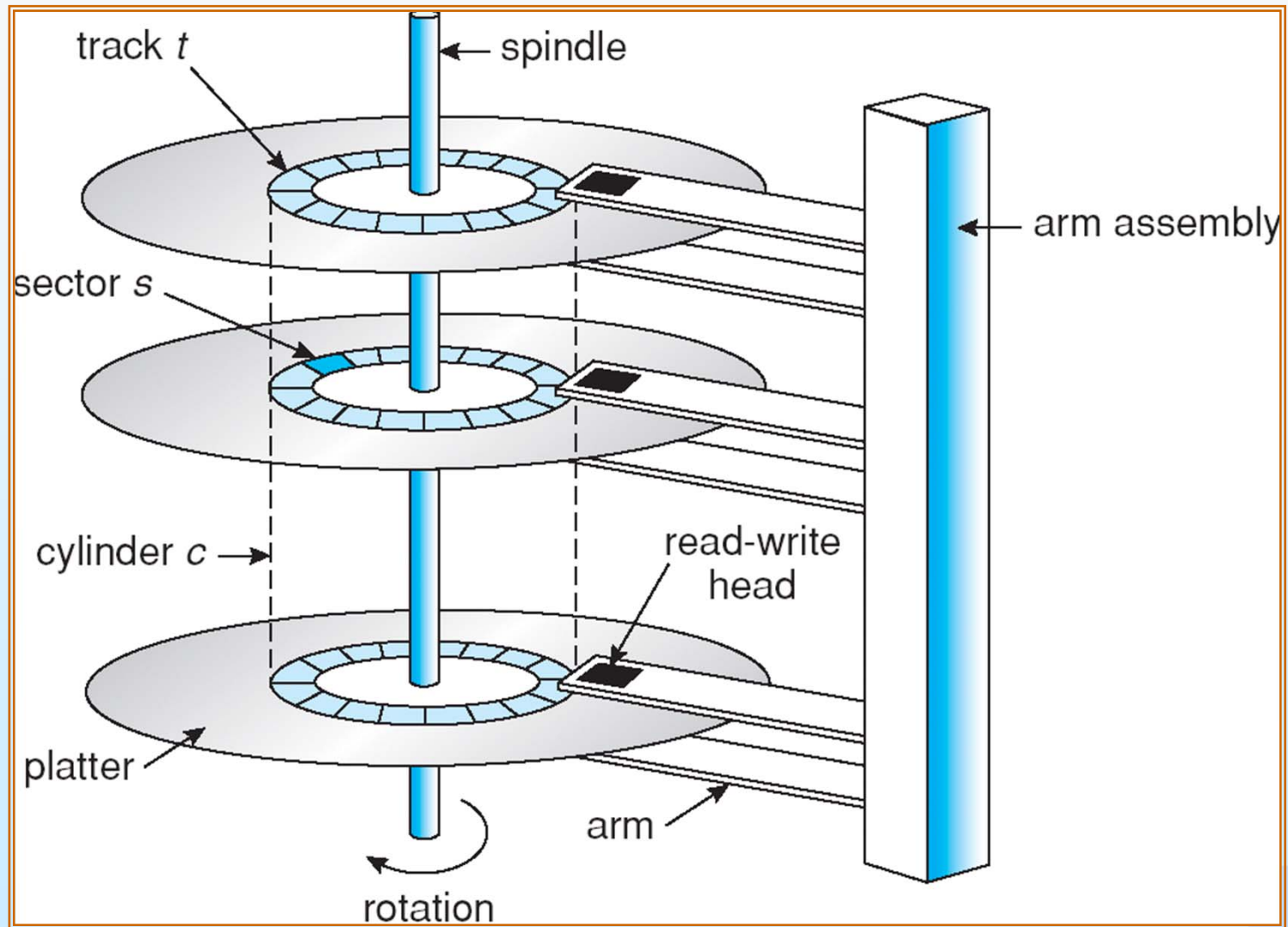
Overview of Mass Storage Structure

- Magnetic disks provide bulk of secondary storage of modern computers
 - Drives rotate at 60 to 200 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**)
 - **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





Moving-head Disk Mechanism



磁盘性能参数

为了读写，磁头必须定位于指定的磁道和该磁道中指定的扇区开始处

■ 寻道时间

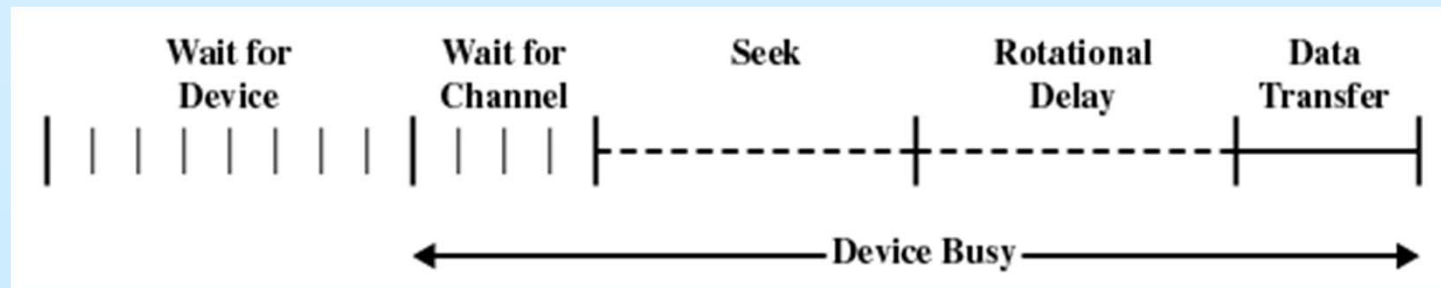
- 磁头定位到磁道所需要的时间；

■ 旋转延迟

- 磁头到达扇区开始位置的时间；

■ 传送时间

- 传送所需的时间；



时序比较

平均寻道时间为4ms,转速为7500r/m,每个磁道有500个扇区,读取2500个扇区

顺序访问?

随机访问?



时序比较

平均寻道时间为4ms,转速为7500r/m,每个磁道有500个扇区,读取2500个扇区

顺序访问

读第一个磁道4ms(平均寻址)+
4ms(旋转延迟)+
8ms(读500个扇区)=16ms

总时间=16+4*(4+8)=64ms

随机访问

读一个扇区4+4+0.016=8.016ms

总时间=2500*8.016=20040ms

显然, 磁盘读取扇区的顺序对I/O性能影响很大!





Overview of Mass Storage Structure (Cont.)

■ Magnetic tape

- Was **early** secondary-storage medium
- 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
- 20-200GB typical storage
- Common technologies are 4mm, 8mm, 19mm, LTO-2 and SDLT





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 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.





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 two major components
 - **Seek time** is the time for the disk are 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.
- Minimize seek time
- **Seek time \approx seek distance**
- 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.)

- Several algorithms exist to schedule the servicing of disk I/O requests.
- We illustrate them with a request queue (0-199).

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

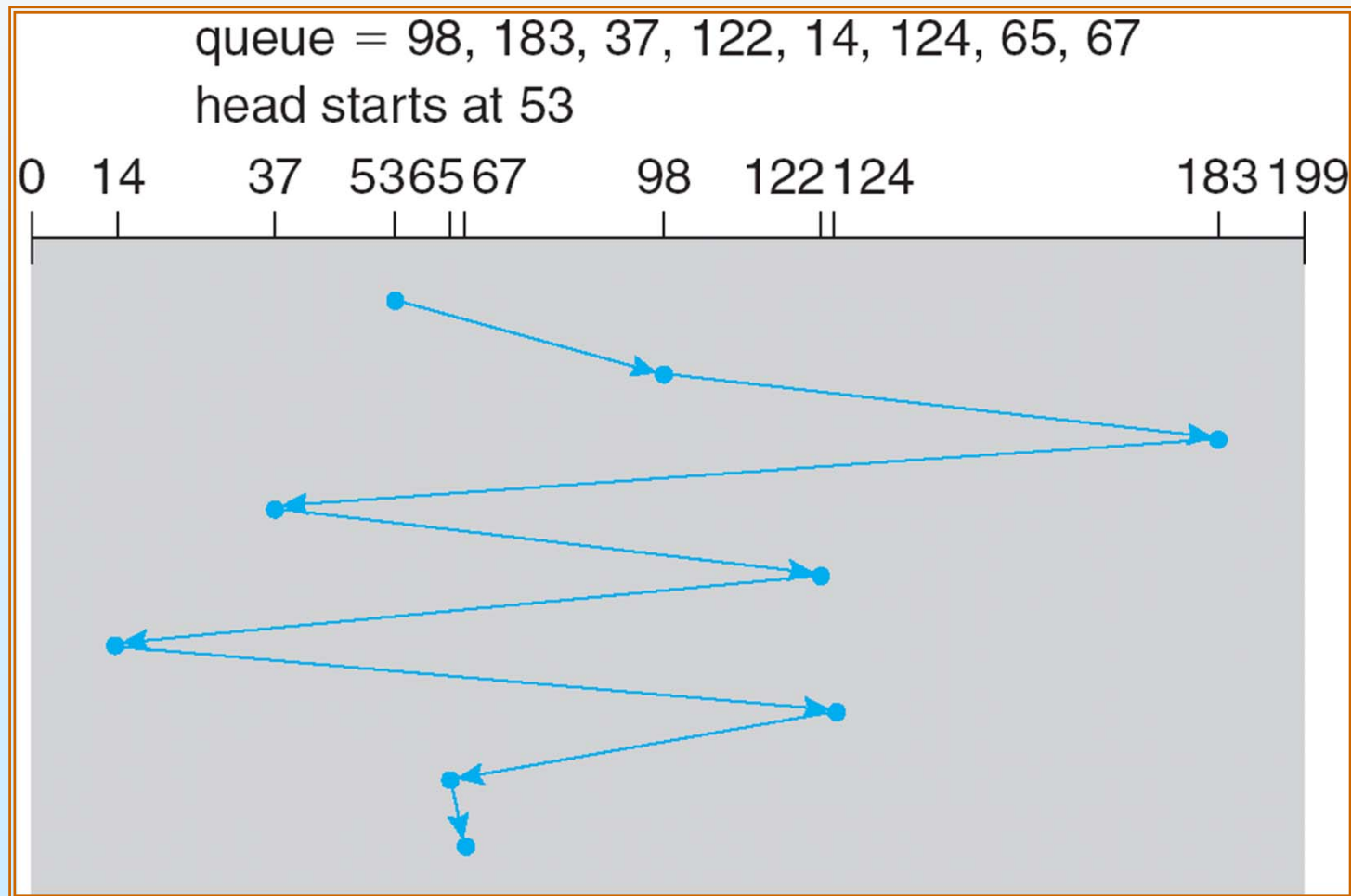
Head pointer 53





FCFS

Illustration shows total head movement of 640 cylinders.





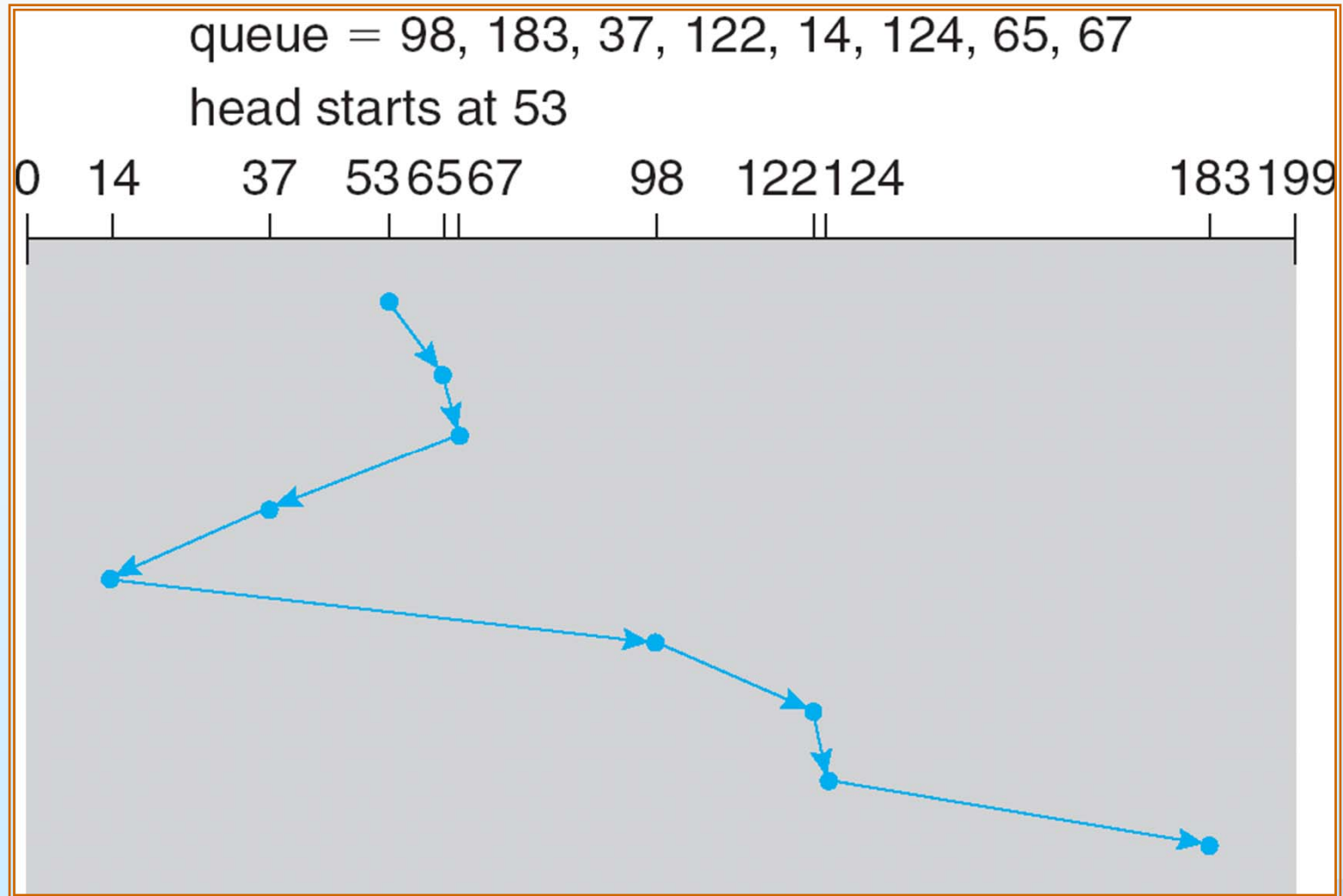
SSTF

- 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.)





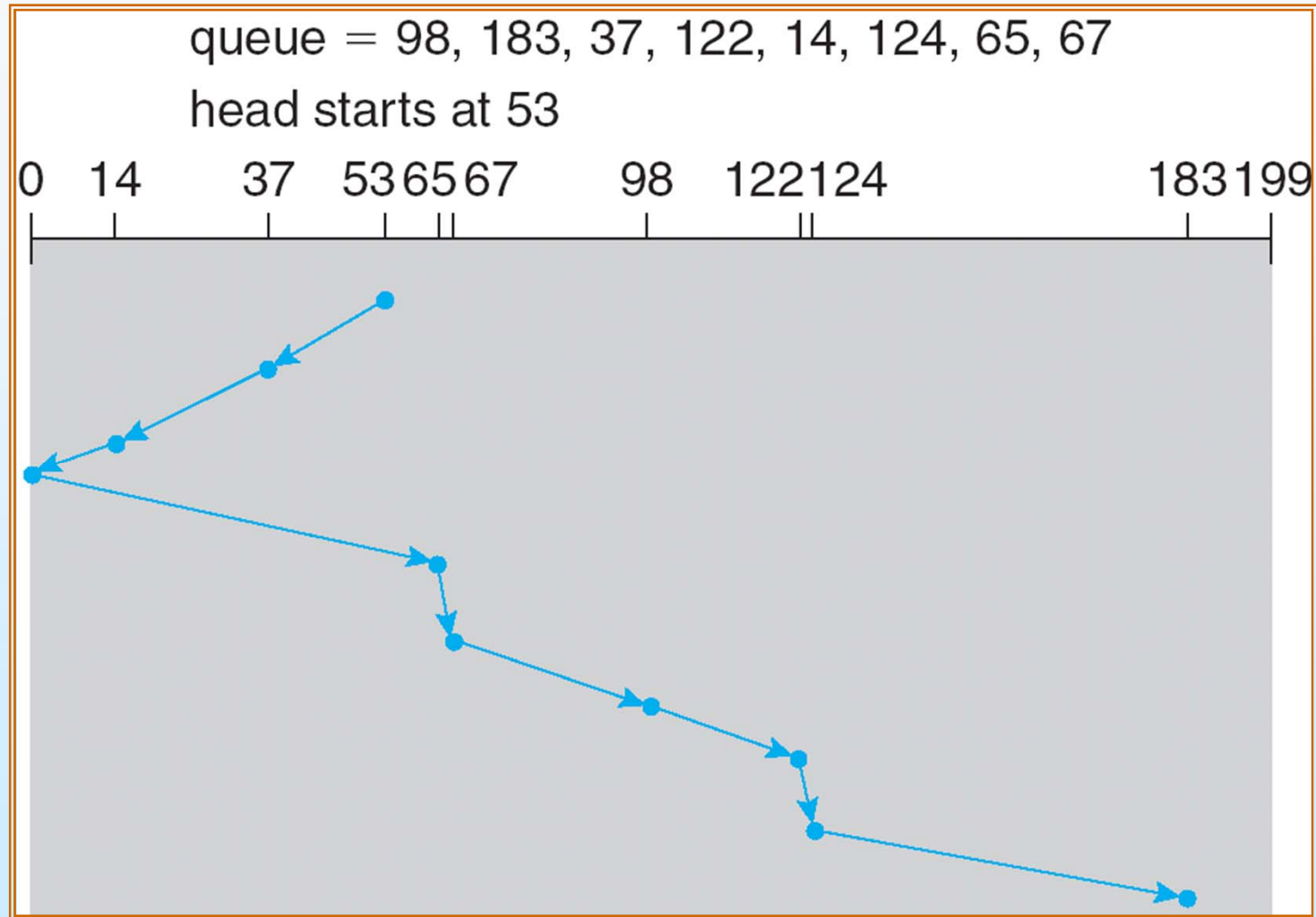
SCAN

- 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 208 cylinders.





SCAN (Cont.)





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.

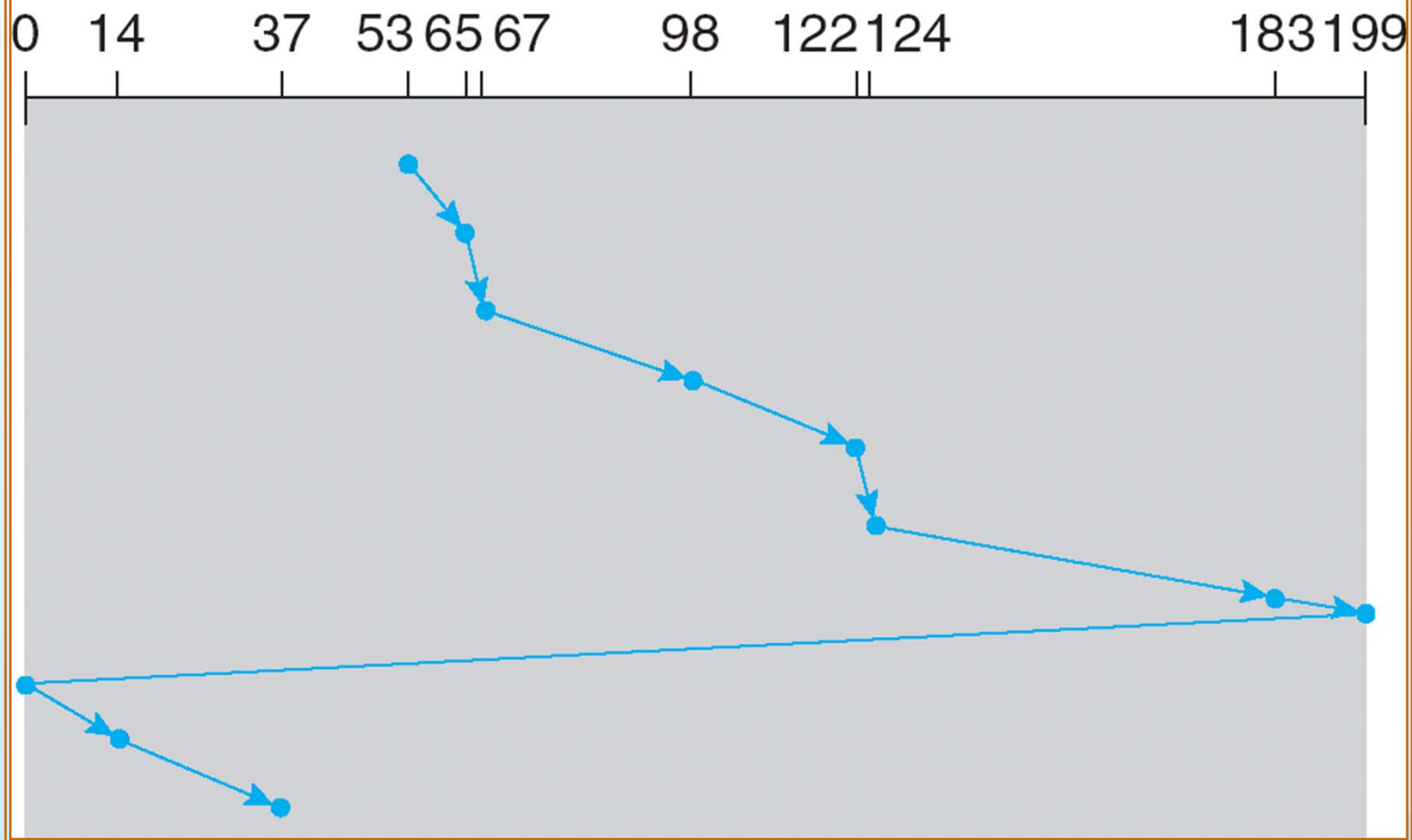




C-SCAN (Cont.)

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

head starts at 53





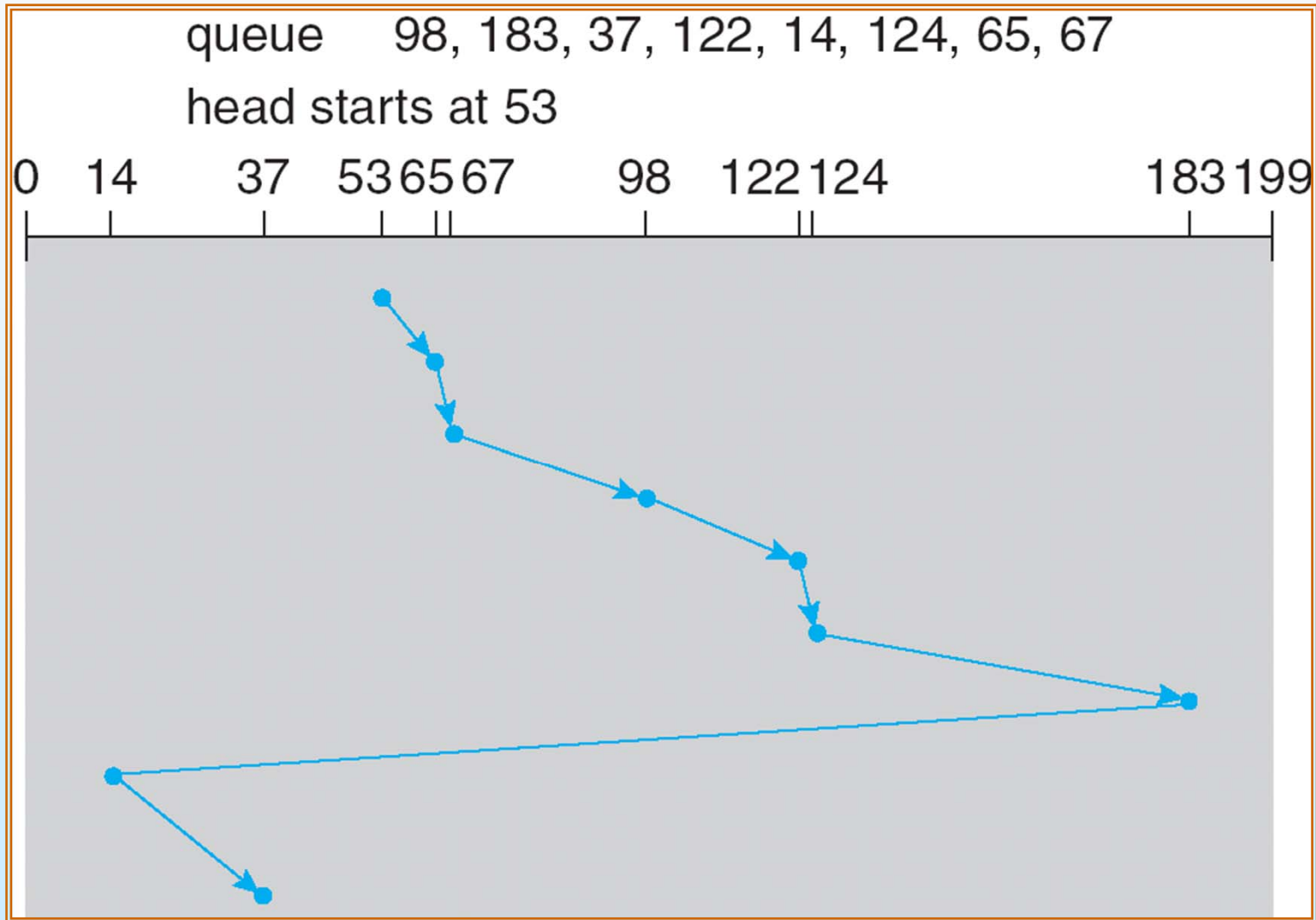
C-LOOK

- 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.)





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.





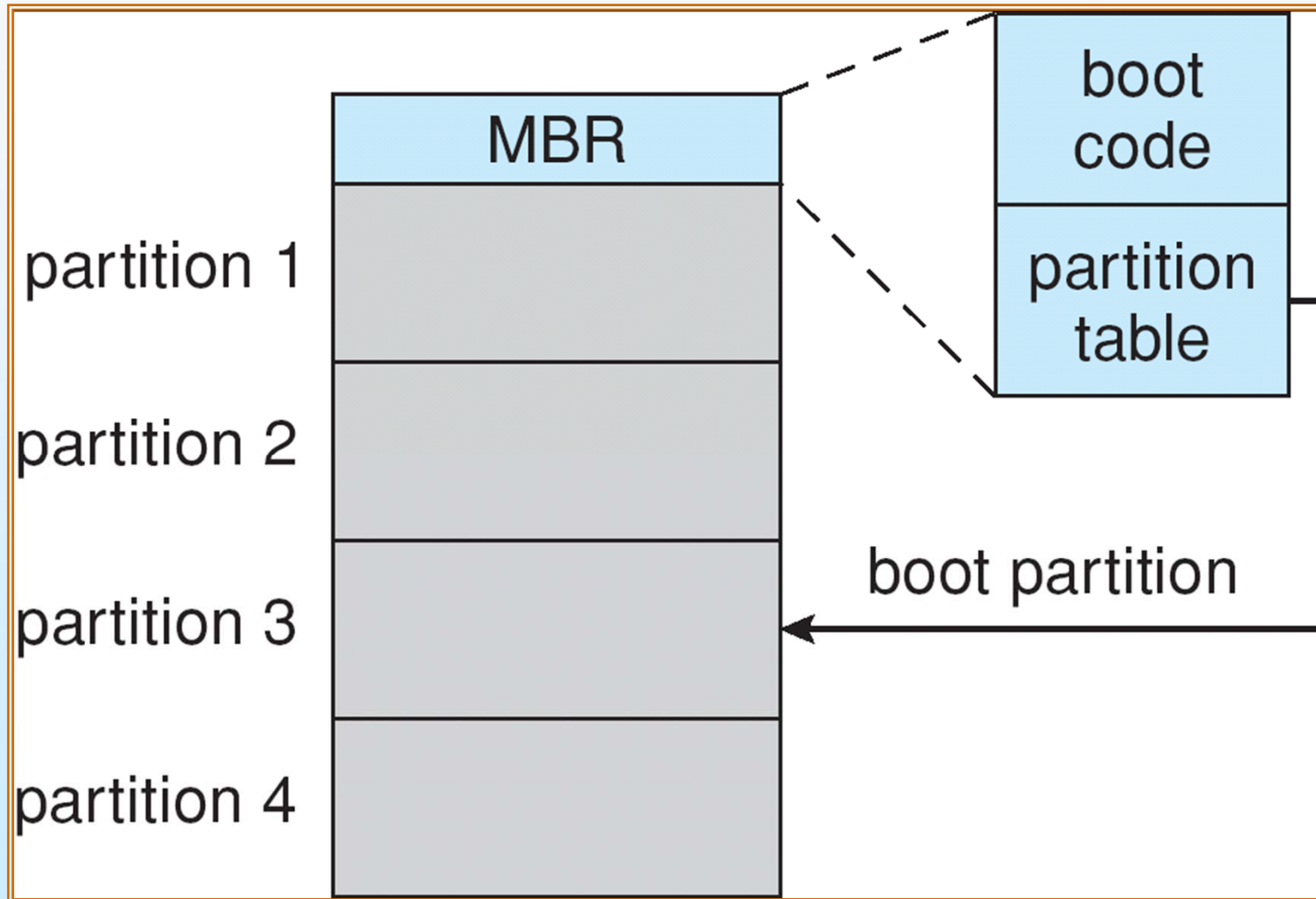
Disk Management

- *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”.
- Boot block initializes system.
 - The bootstrap is stored in ROM.
 - *Bootstrap loader* program.
- Methods such as *sector sparing* used to handle bad blocks.





Booting from a Disk in Windows 2000





Swap-Space Management

- Swap-space — Virtual memory **uses disk space as an extension of main memory.**
- Swap-space can be carved out of the normal file system, or, more commonly, it can be in a separate disk partition.
- 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.





RAID Structure

- **RAID** – multiple disk drives provides **reliability** via **redundancy**.
- RAID is arranged into six different levels.

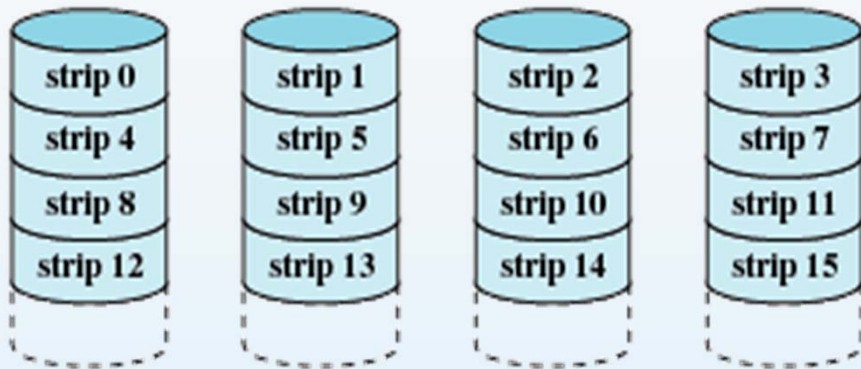


RAID

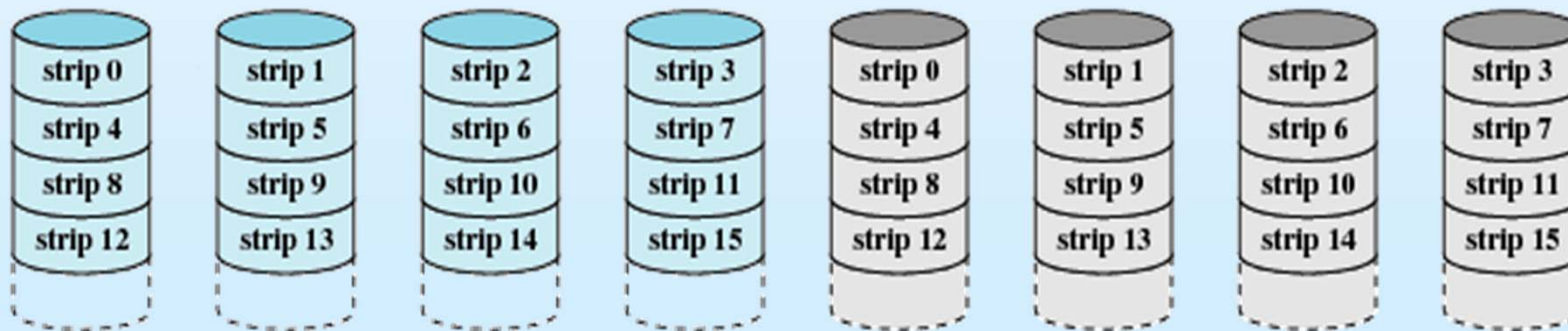
- **独立磁盘冗余阵列**是利用**一台磁盘阵列控制器**统一管理和控制一组磁盘驱动器，组成一个速度快、可靠性高的大容量磁盘系统。
 - RAID是一组物理磁盘驱动器，操作系统把它看作是一个单独的逻辑驱动器；
 - 数据分布在物理驱动器阵列中；
 - 使用冗余的磁盘容量保存奇偶校验信息，从而保证当一个磁盘失败时，数据具有可恢复性；



RAID



(a) RAID 0 (non-redundant)



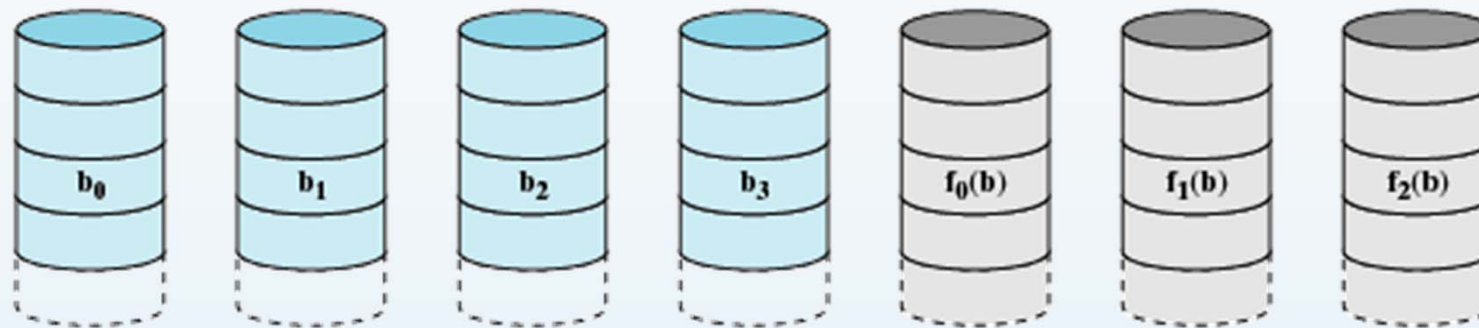
(b) RAID 1 (mirrored)

RAID0
无冗余

RAID1
镜像

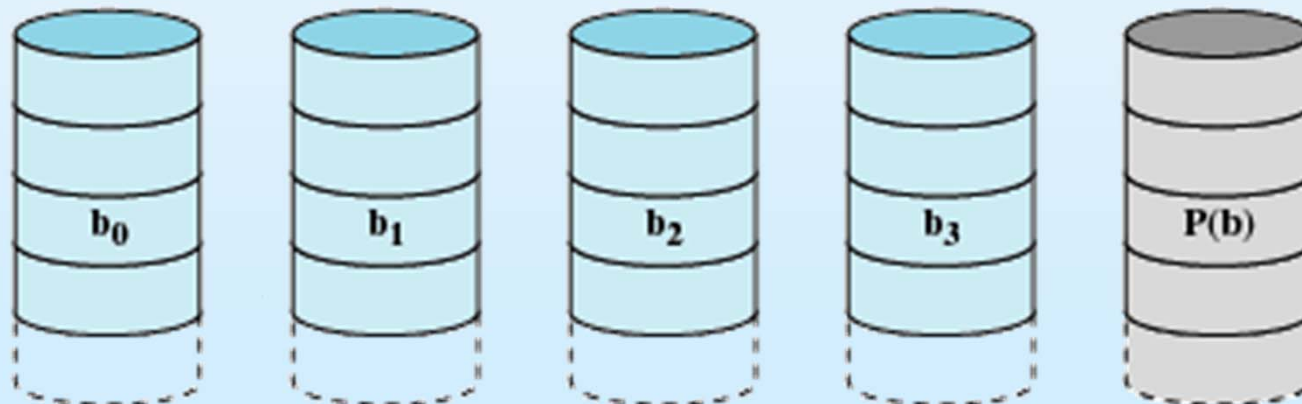


RAID



RAID2
通过汉明码冗余

(c) RAID 2 (redundancy through Hamming code)



RAID3
交错位奇偶校验

(d) RAID 3 (bit-interleaved parity)





SSD

- SSD = solid state disk, 固态硬盘, 用固态电子存储芯片阵列而制成的硬盘。
- 固态硬盘的接口规范和定义、功能及使用方法上与普通硬盘的相同, 在产品外形和尺寸上也与普通硬盘一致。其芯片的工作温度范围很宽 ($-40\sim 85^{\circ}\text{C}$) ;
- 由于固态硬盘技术与传统硬盘技术不同, 所以产生了不少新兴的存储器厂商。厂商只需购买NAND存储器, 再配合适当的控制芯片, 就可以制造固态硬盘了。新一代的固态硬盘普遍采用SATA-2接口。





SSD

- 固态硬盘的存储介质分为两种，一种是采用闪存（**FLASH**芯片）作为存储介质，另外一种是采用**DRAM**作为存储介质。
- 基于闪存的固态硬盘：采用**FLASH**芯片作为存储介质，这也是我们通常所说的**SSD**。这种**SSD**固态硬盘最大的优点就是可以移动，而且数据保护不受电源控制，能适应于各种环境，但是使用年限不高，适合于个人用户使用。在基于闪存的固态硬盘中，存储单元又分为两类：**SLC（Single Layer Cell 单层单元）**和**MLC（Multi-Level Cell 多层单元）**。**SLC**的特点是成本高、容量小、但是速度快，而**MLC**的特点是容量大成本低，但是速度慢。**MLC**的每个单元是**2bit**的，相对**SLC**来说整整多了一倍。不过，由于每个**MLC**存储单元中存放的资料较多，结构相对复杂，出错的几率会增加，必须进行错误修正，这个动作导致其性能大幅落后于结构简单的**SLC**闪存。此外，**SLC**闪存的优点是复写次数高达**100000**次，比**MLC**闪存高**10**倍。此外，为了保证**MLC**的寿命，控制芯片都校验和智能磨损平衡技术算法，使得每个存储单元的写入次数可以平均分摊，达到**100**万小时故障间隔时间(MTBF)。
- 基于**DRAM**的固态硬盘
- 基于**DRAM**的固态硬盘：采用**DRAM**作为存储介质，目前应用范围较窄。它仿效传统硬盘的设计、可被绝大部分操作系统的文件系统工具进行卷设置和管理，并提供工业标准的**PCI**和**FC**接口用于连接主机或者服务器。应用方式可分为**SSD**硬盘和**SSD**硬盘阵列两种。它是一种高性能的存储器，而且使用寿命很长，美中不足的是需要独立电源来保护数据安全。



End of Chapter 12

