北京邮电大学软件学院
School Of software Engineering Of BUPT

Operating Systems

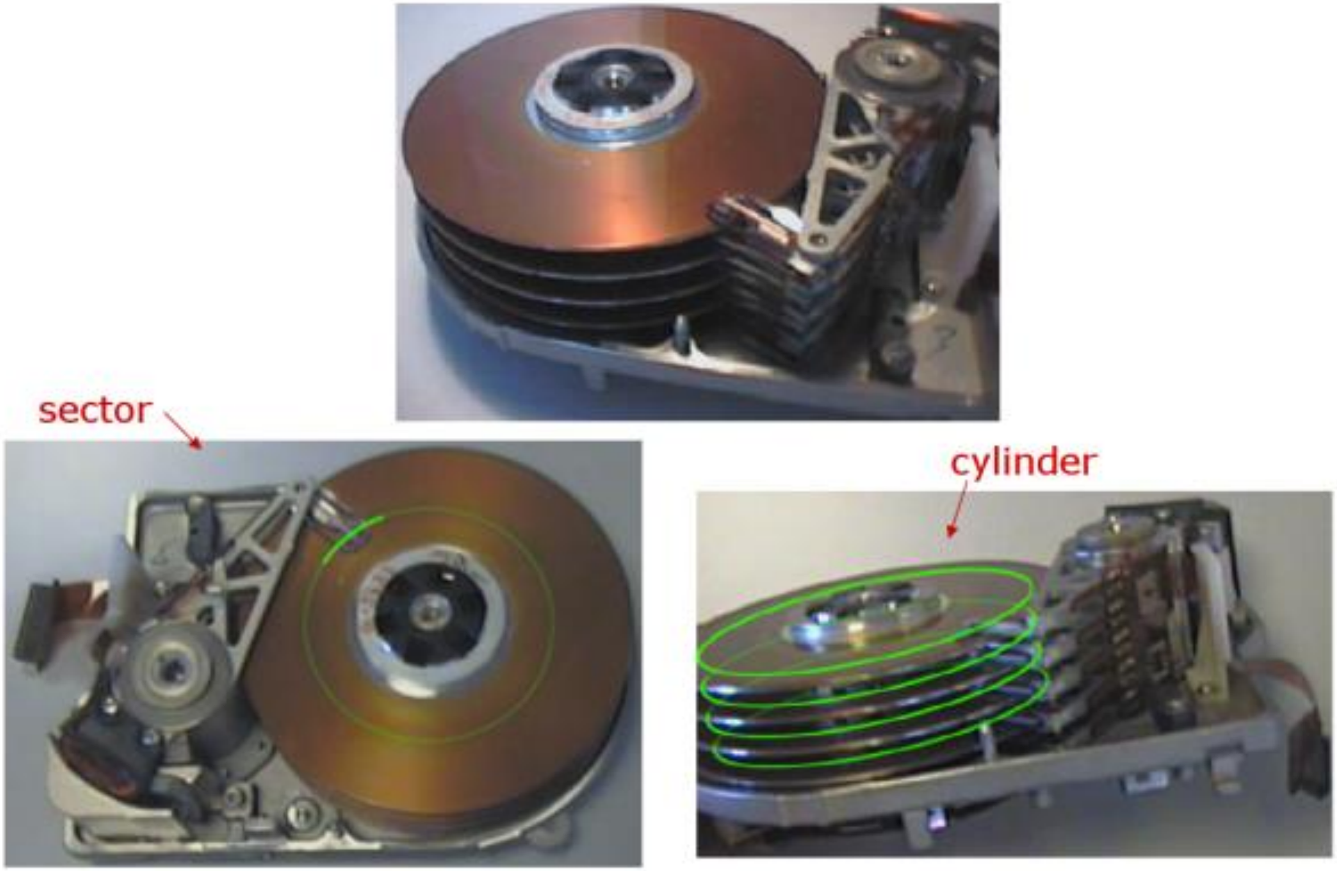
Lecture 11 ： Mass-Storage Systems

Jinpengchen
Email: jpchen@bupt.edu.cn

Catalog Description

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

Overview of Mass Storage Structure

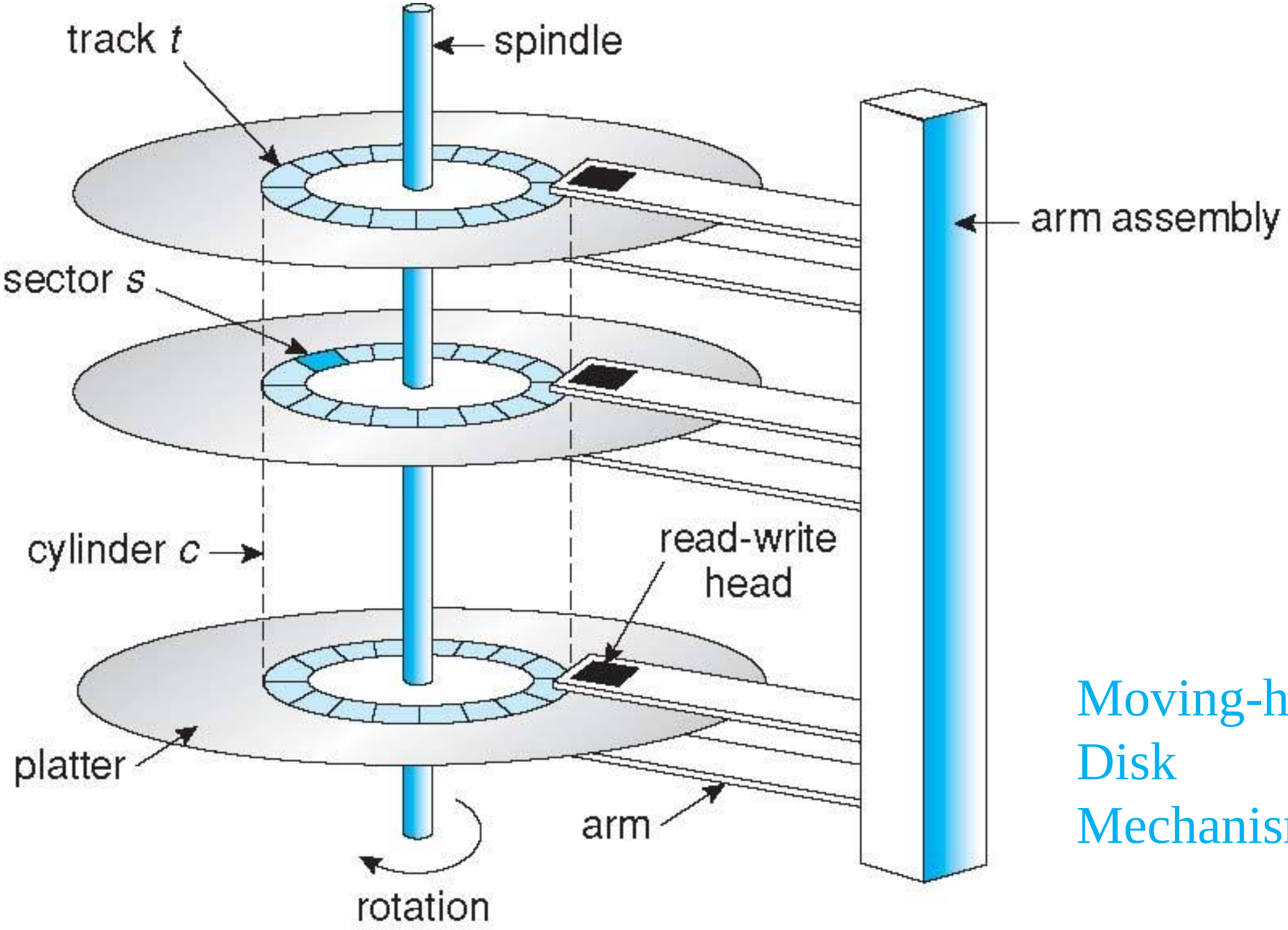


Overview of Mass Storage Structure

- Magnetic disks (磁盘) provide bulk of secondary storage of modern computers
  - Drives rotate at 60 to 200 times per second
  - Transfer rate (传输速率) 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, Fiber Channel, SCSI

Overview of Mass Storage Structure

- Host controller in computer uses bus to talk to disk controller built into drive or storage array



Moving-head Disk Mechanism

Catalog Description

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

Disk Structure

- Disk drives are addressed as large 1-dimensional arrays of logical blocks,
  - The logical block is the smallest unit of transfer.
  - Usually, 512B
- The 1-D array of logical blocks is mapped into the sectors of the disk sequentially.
  - Cylinder: track: sector
  - 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.
  - However, in practice, the mapping is difficult, because
    - ✓ Defective sectors
    - ✓ Sectors/track != constant → zones of cylinder

Catalog Description

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

Disk Scheduling（磁盘调度)

- The OS 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 to move the heads to the cylinder containing the desired sector.
    - ✓ Minimize seek time
    - ✓ Seek time ≈ seek distance
  - Rotational latency is the additional time waiting for the disk to rotate the desired sector to the disk head.
- Disk bandwidth (磁盘带宽) 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（磁盘调度)

- Request queue (请求队列)
  - empty or not
- How?

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

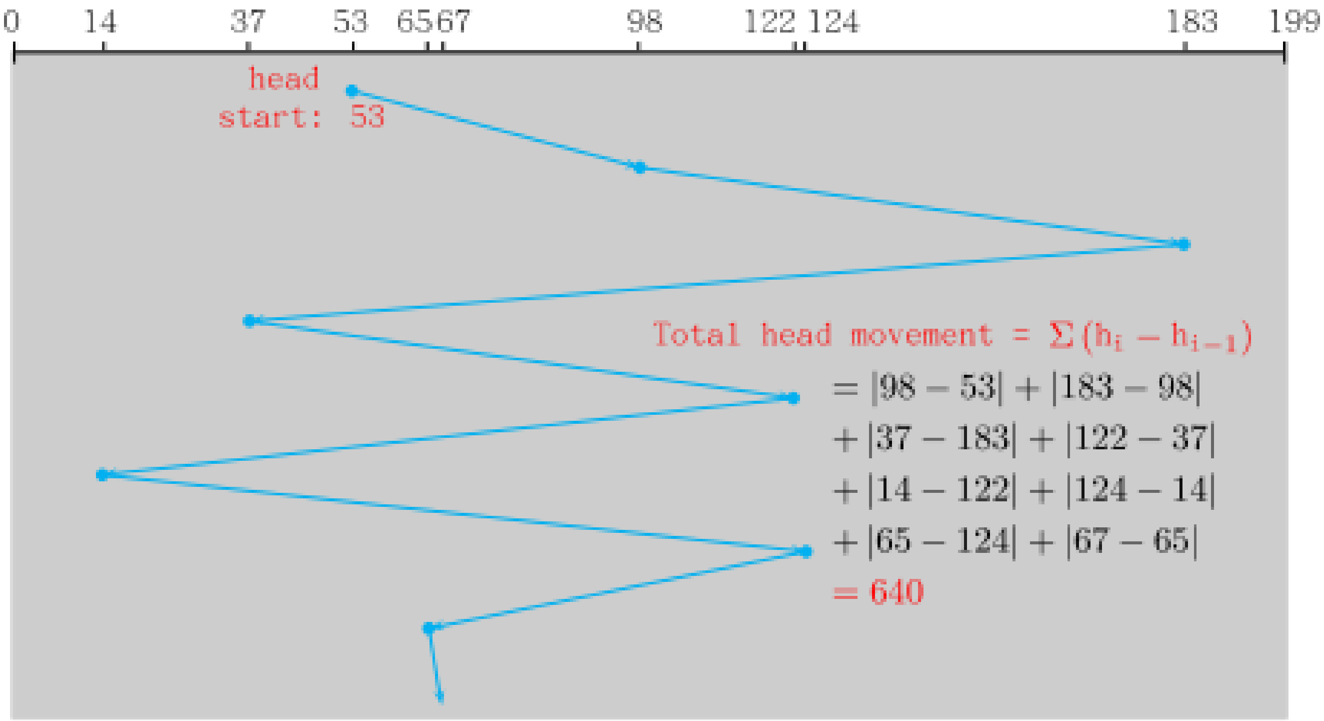
  - FCFS
  - SSTF (shortest-seek-time-first)
  - SCAN (elevator algorithm)
  - C-SCAN
  - C-LOOK
- We illustrate them with a request queue (0-199).

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

Head points to 53 initially

Disk Scheduling (磁盘调度)

- First Come, First Served (FCFS, 先来先服务)
  - Illustration shows total head movement of 640 cylinders

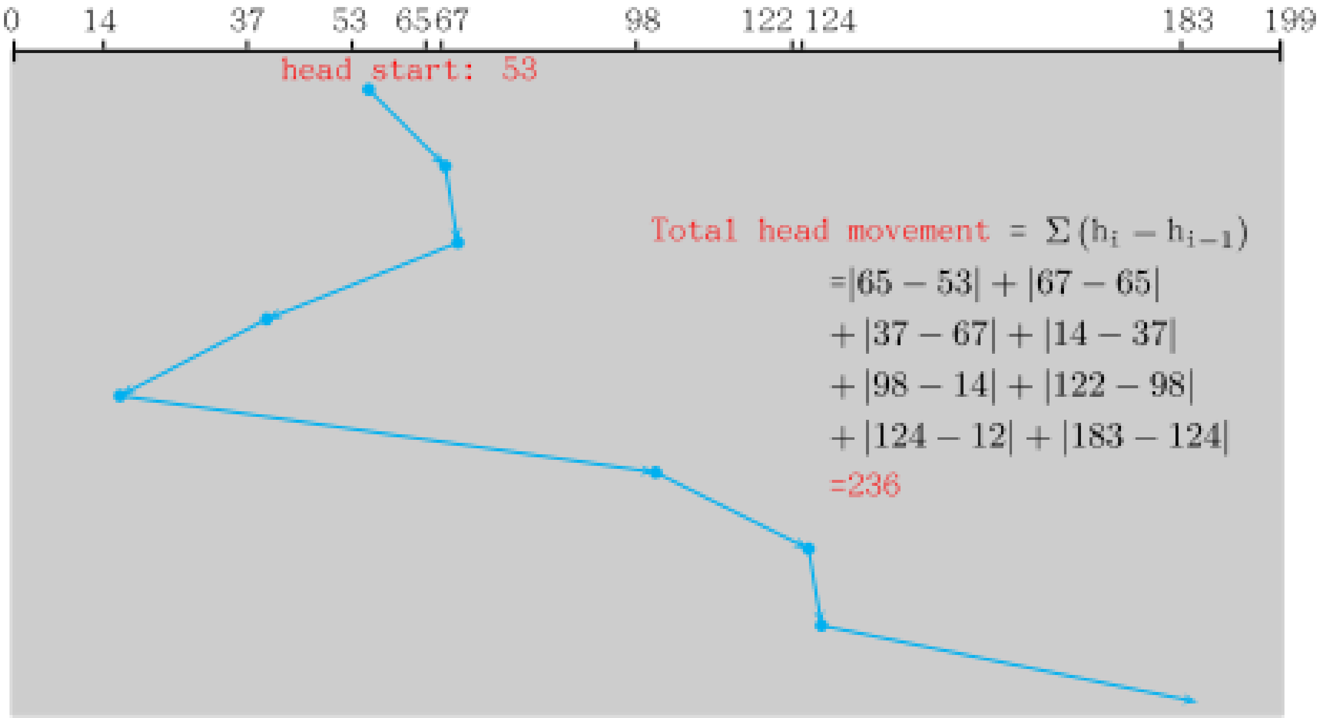


$$\text{Total head movement} = \Sigma(h_i - h_{i-1})$$
$$= |98 - 53| + |183 - 98|$$
$$+ |37 - 183| + |122 - 37|$$
$$+ |14 - 122| + |124 - 14|$$
$$+ |65 - 124| + |67 - 65|$$
$$= 640$$

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

Disk Scheduling（磁盘调度）

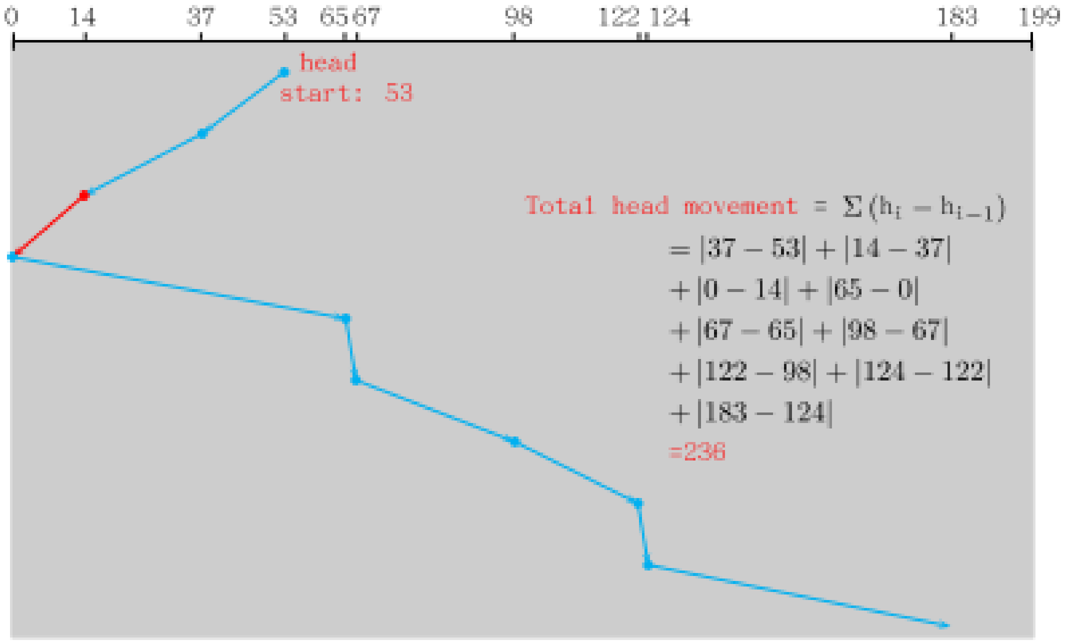
- SSTF (shortest-seek-time-first)
  - Selects the request with the minimum seek time from the current head position
  - SSTF scheduling is a form of SJF scheduling; may cause starvation of some requests
  - Illustration shows total head movement of 236 cylinders



$$\text{Total head movement} = \Sigma(h_i - h_{i-1})$$
$$= |65-53| + |67-65|$$
$$+ |37-67| + |14-37|$$
$$+ |98-14| + |122-98|$$
$$+ |124-12| + |183-124|$$
$$=236$$

Disk Scheduling（磁盘调度)

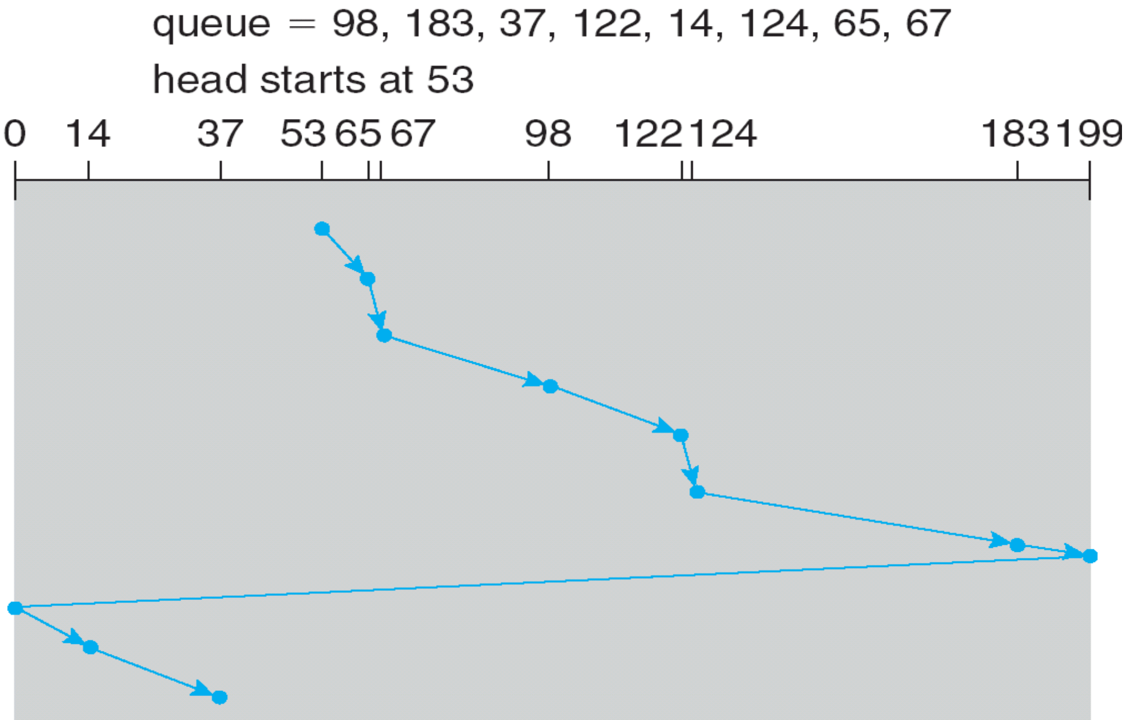
- SCAN (elevator algorithm)
  - The disk arm starts at one end of the disk, and moves toward the other end, servicing requests until it gets to the other end of the disk, where the head movement is reversed and servicing continues.



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

Disk Scheduling（磁盘调度)

- 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



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

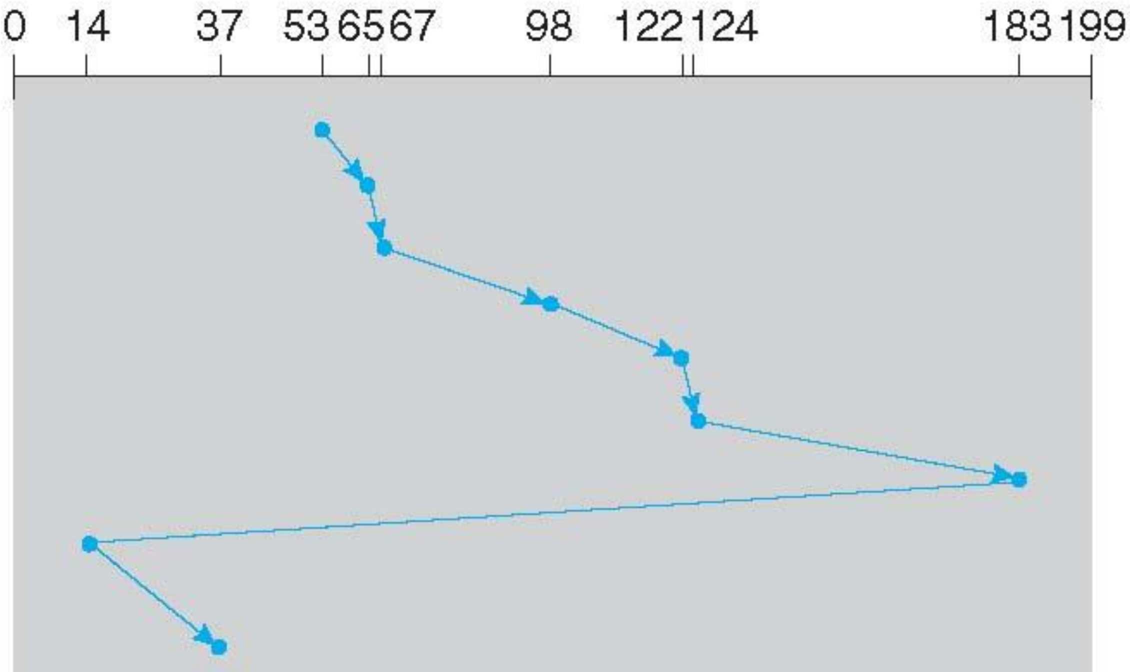
head starts at 53

Disk Scheduling (磁盘调度)

- 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.

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

head starts at 53



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, which can be influenced by

  - The file-allocation method
  - The location of directories and index blocks (caching?)

- Either SSTF or LOOK is a reasonable choice for the default algorithm.
- The disk-scheduling algorithm should be written as a separate module of the OS, allowing it to be replaced with a different algorithm if necessary.

Catalog Description

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

Disk Management

- Disk Formatting

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 OS 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".
- To increase efficiency, most file-systems group blocks together into larger chunks, frequently called clusters
  - Disk I/O done in blocks
  - File I/O done in clusters

Catalog Description

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

Swap-Space Management

- Swapping & paging
  - Entire processes
  - Paging √
- Swap-space (对换空间)

  Virtual memory uses disk space as an extension of main memory.
  - It can be carved out of the normal file system
    - ✓ A large file with the file system
  - Or, more commonly, it can be in a separate disk partition.

Catalog Description

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

RAID Structure

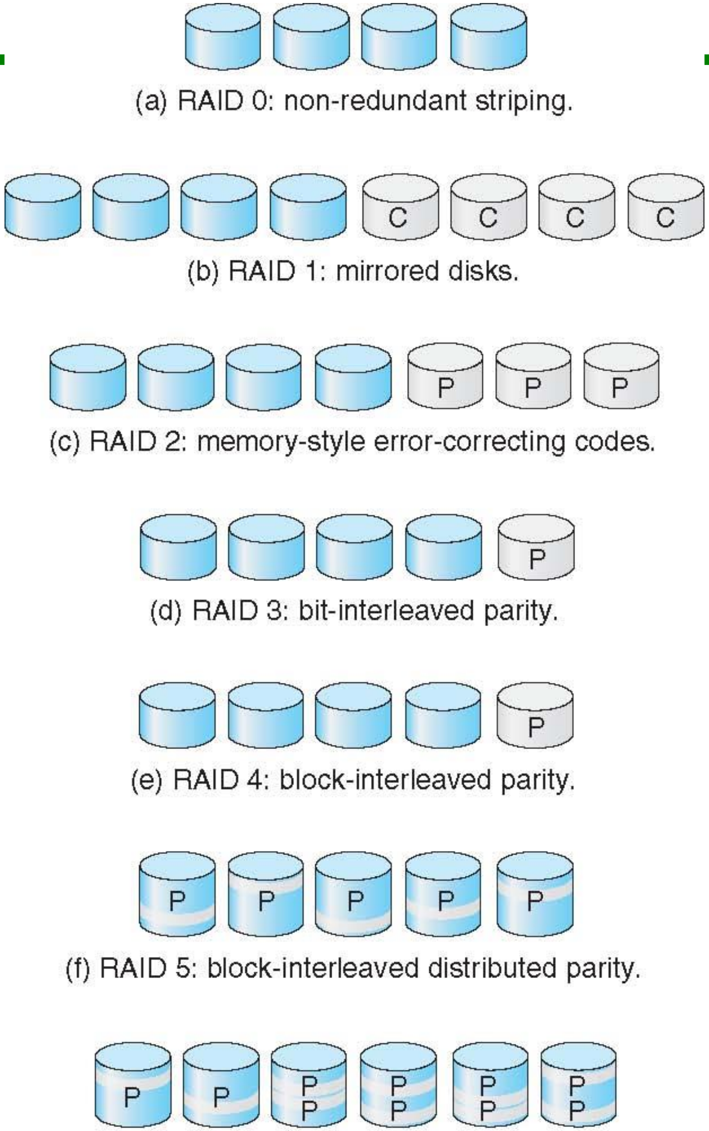
- Redundant arrays of inexpensive disks (RAIDs, 磁盘阵列) – multiple disk drives provides reliability via redundancy; higher data-transfer rate
- Frequently combined with NVRAM to improve write performance
- RAID is arranged into six different levels

RAID Structure

- 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 Structure

- RAID Levels



(a) RAID 0: non-redundant striping.

(b) RAID 1: mirrored disks.

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

(d) RAID 3: bit-interleaved parity.

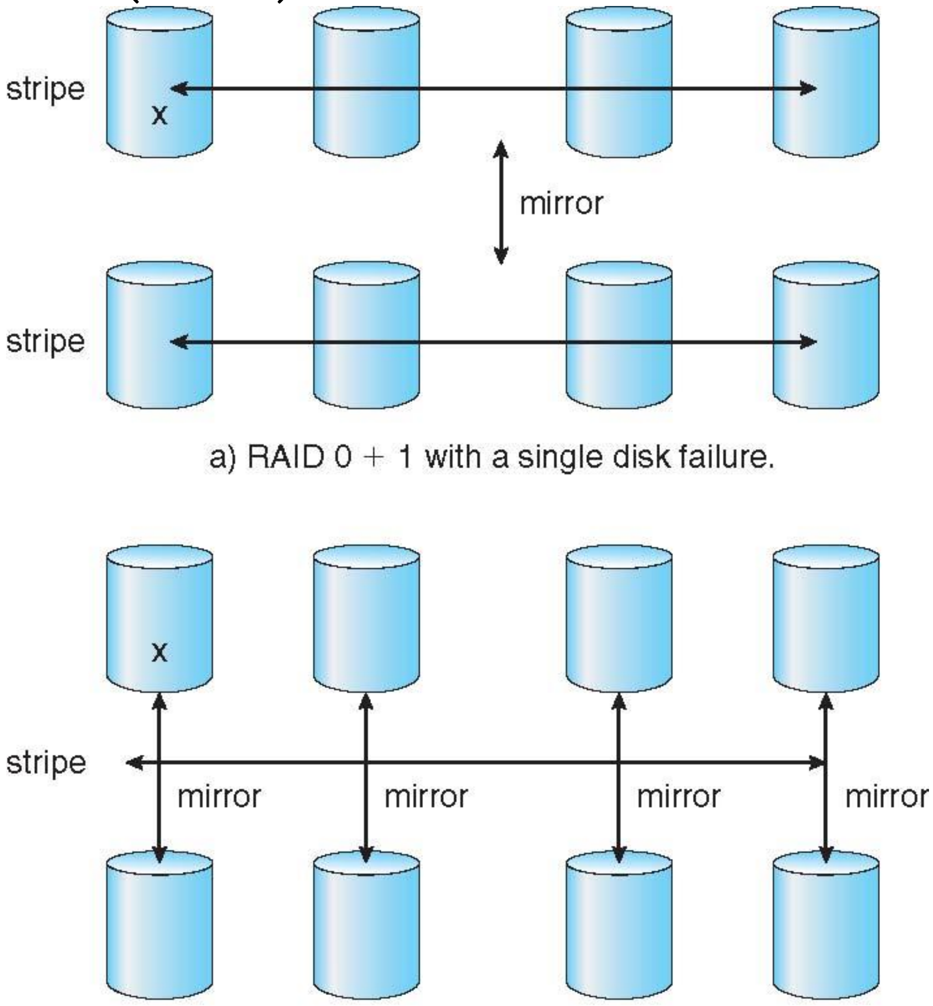
(e) RAID 4: block-interleaved parity.

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

(g) RAID 6: P + Q redundancy.

RAID Structure

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



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

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