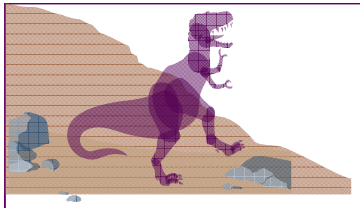
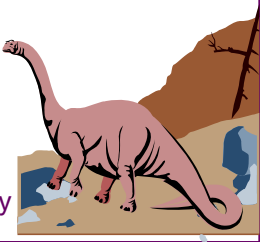
Chapter 12: Mass Storage Structure

肖 卿 俊
办公室：计算机楼212室
电邮：csqjxiao@seu.edu.cn
主页：http://cse.seu.edu.cn/PersonalPage/csqjxiao
电话：025-52091022

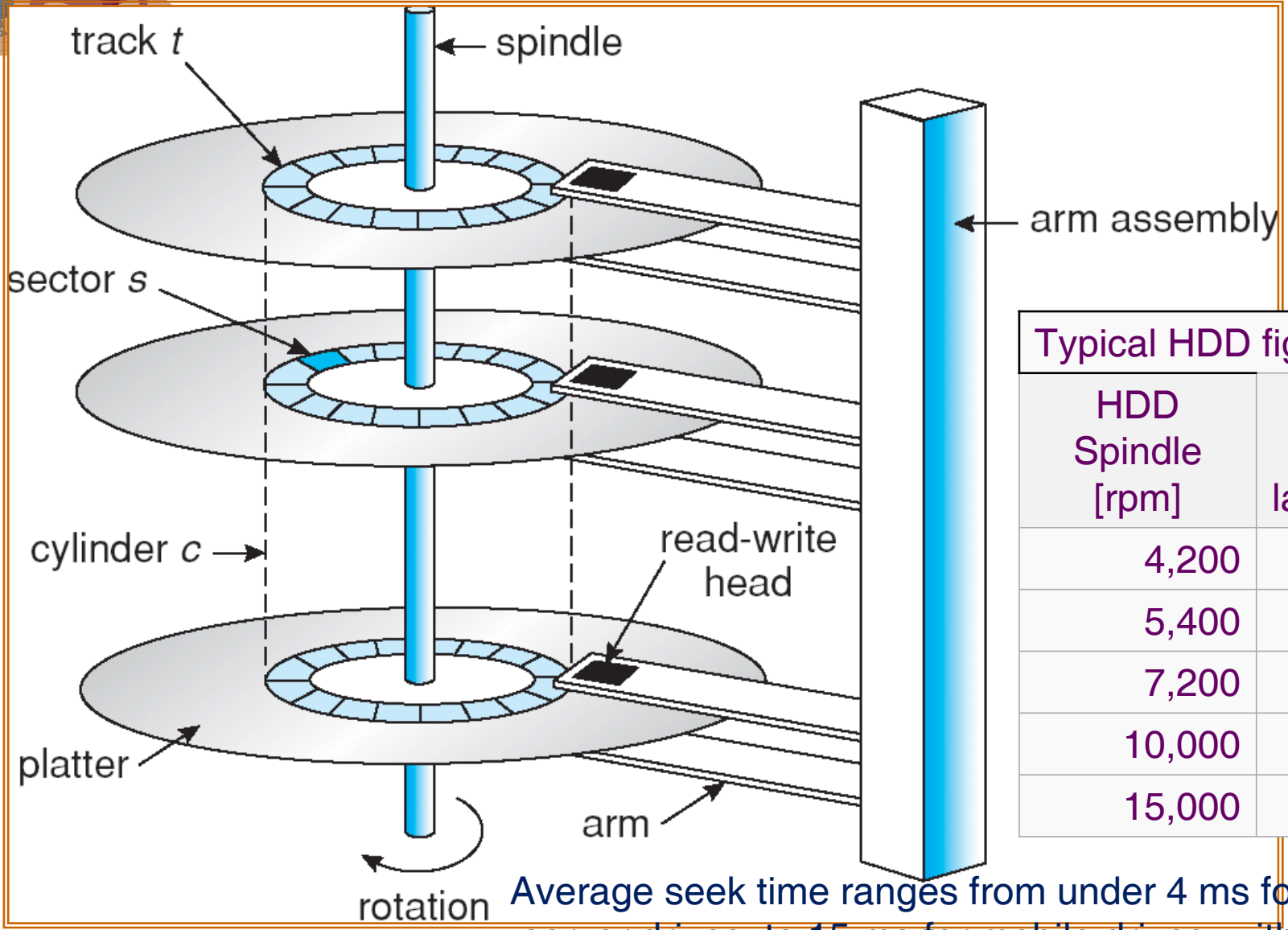


Chapter 12: Mass-Storage Systems

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

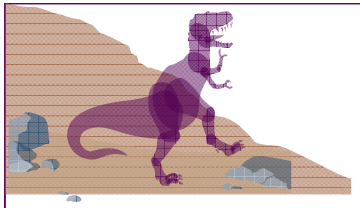


Disk Hardware



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

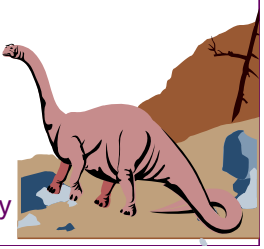
Average seek time ranges from under 4 ms for high-end server drives, to 15 ms for mobile drives, with the most common mobile drives at about 12 ms and the most common desktop drives typically being around 9 ms.

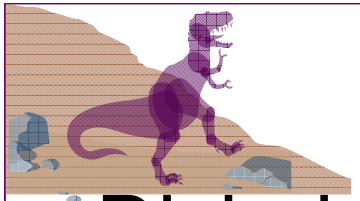


Disk Hardware (Cont.)

Parameter	IBM 360-KB floppy disk	WD 18300 hard disk
Number of cylinders	40	10601
Tracks per cylinder	2	12
Sectors per track	9	281 (avg)
Sectors per disk	720	35742000
Bytes per sector	512	512
Disk capacity	360 KB	18.3 GB
Seek time (adjacent cylinders)	6 msec	0.8 msec
Seek time (average case)	77 msec	6.9 msec
Rotation time	200 msec	8.33 msec
Motor stop/start time	250 msec	20 sec
Time to transfer 1 sector	22 msec	17 µsec

Disk parameters for the original IBM PC floppy disk and a Western Digital WD 18300 hard disk

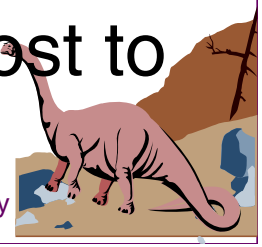


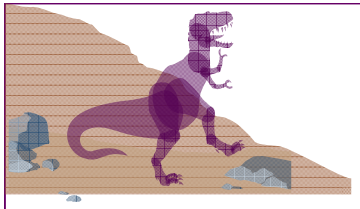


Disk Structure

- 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.

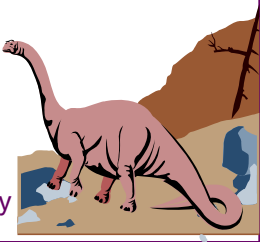


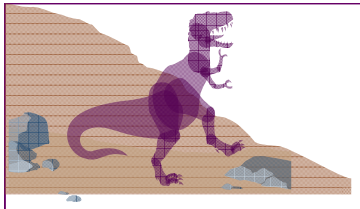


Disk Scheduling

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

- 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.

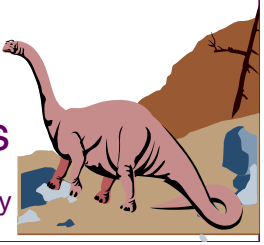


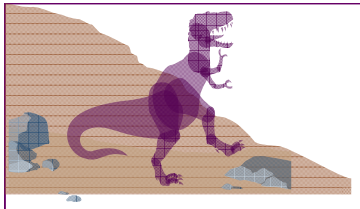


Disk Scheduling (cont.)

- 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

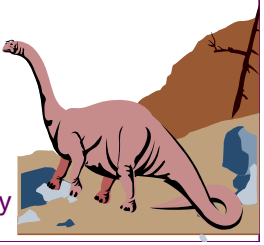


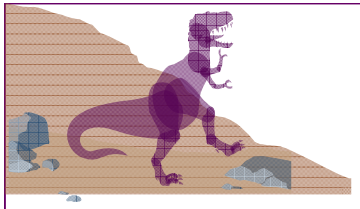


Disk Scheduling (cont.)

- 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





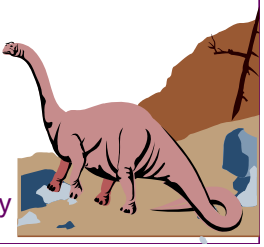
Disk Scheduling (cont.)

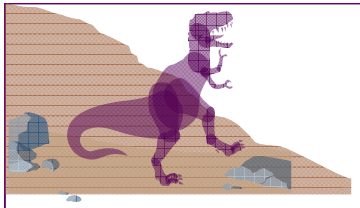
- Several algorithms exist to schedule the servicing of disk I/O requests.

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

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

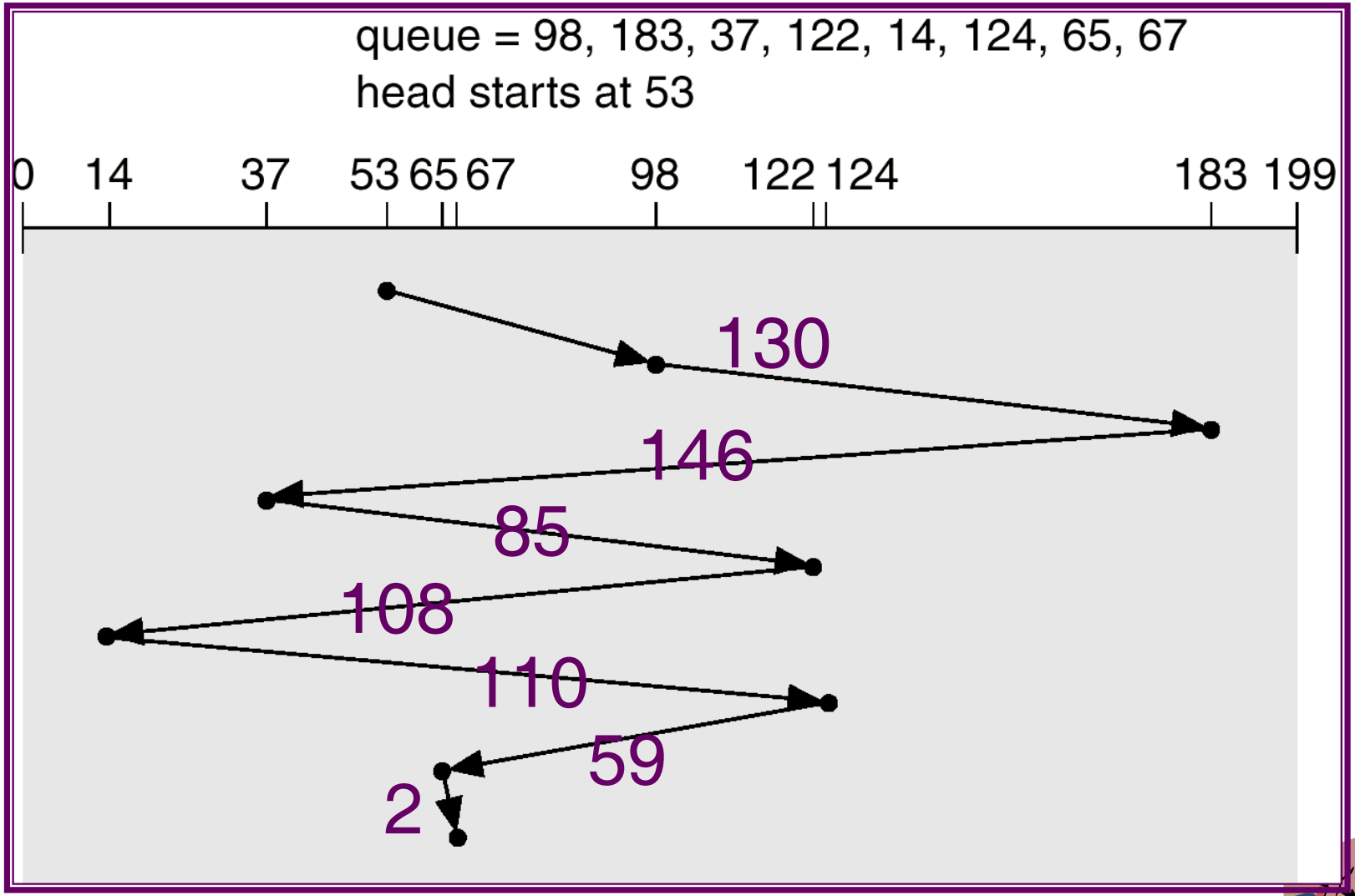
  Head pointer 53

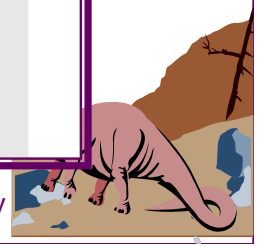




FCFS

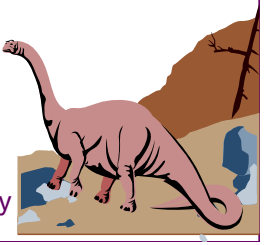
Illustration shows total head movement of 640 cylinders.

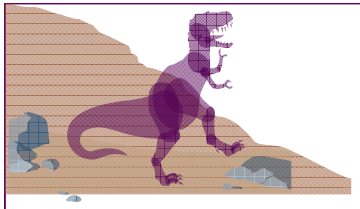




SSTF (Shortest Seek Time First)

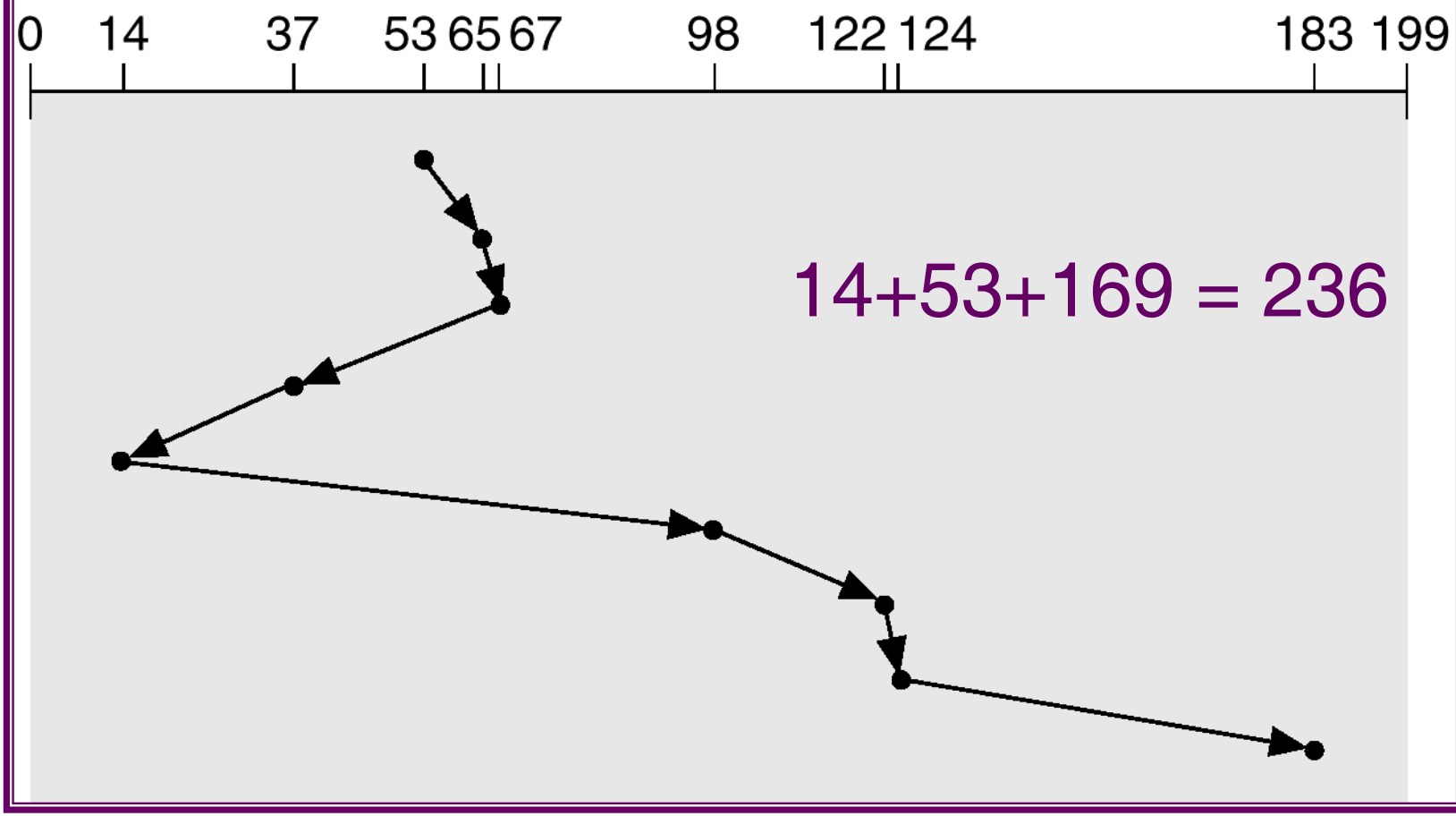
- Selects the request with the minimum seek time from the current head position.
- SSTF scheduling is a form of SJF scheduling; may cause starvation of some requests.
- Illustration shows total head movement of 236 cylinders.

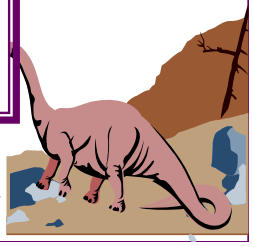


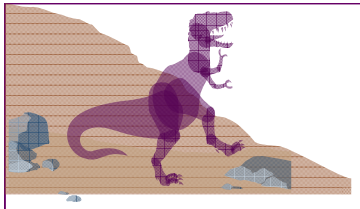


SSTF (cont.)

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

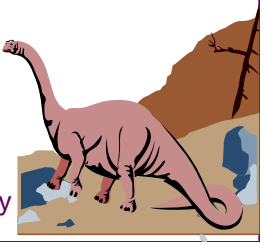


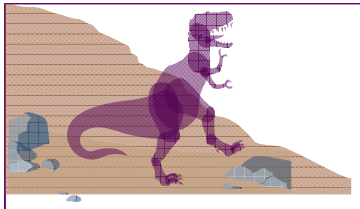




SCAN

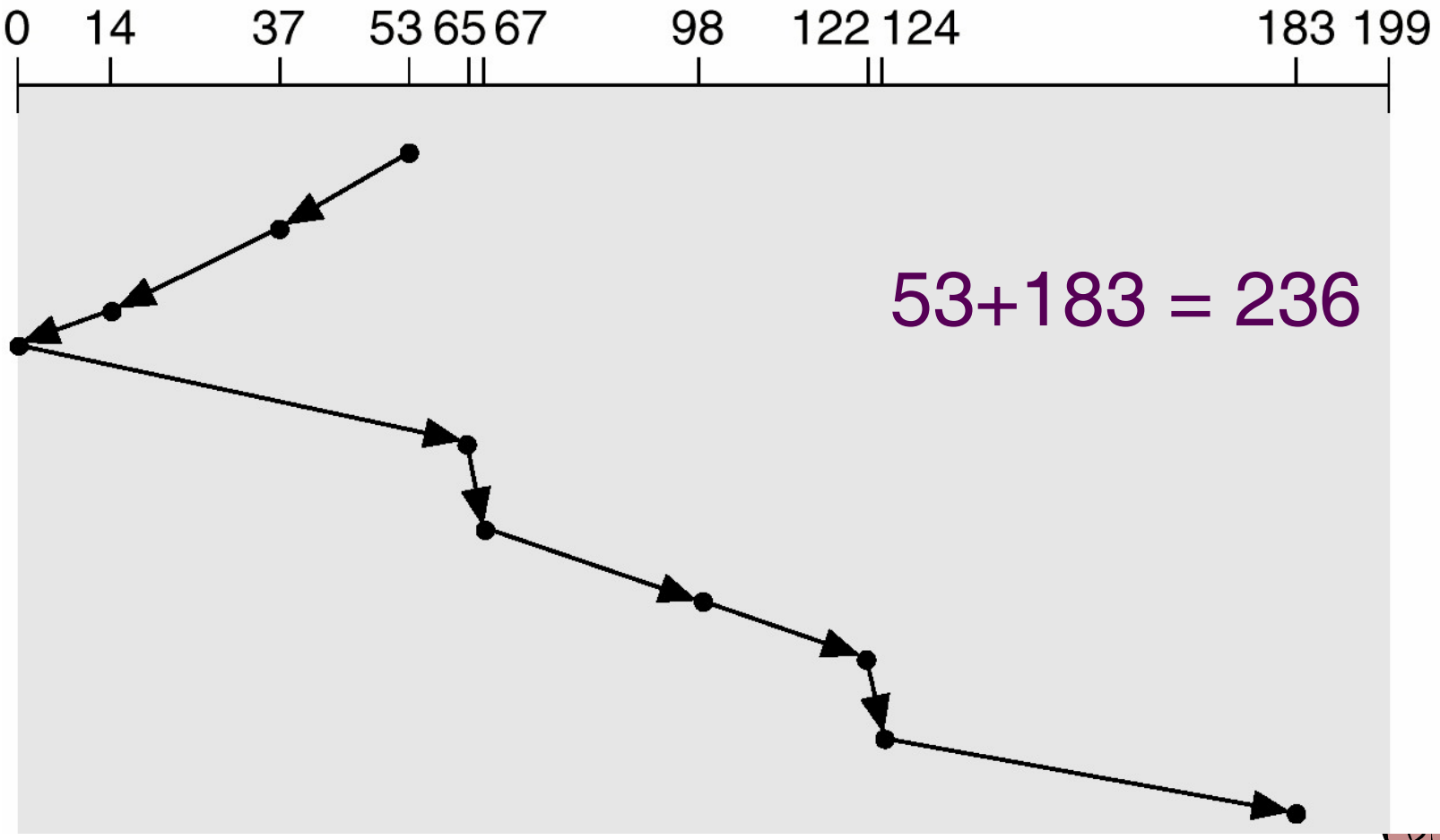
- The disk arm starts at one end of the disk, and moves toward the other end, servicing requests until it gets to the other end of the disk, where the head movement is reversed and servicing continues.
- Sometimes called the *elevator algorithm*.
- Illustration shows total head movement of 236 cylinders.

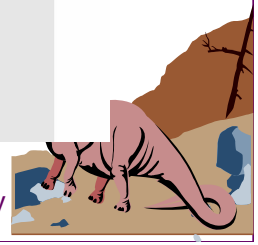


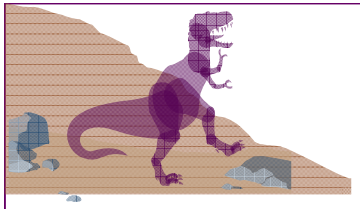


SCAN (cont.)

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

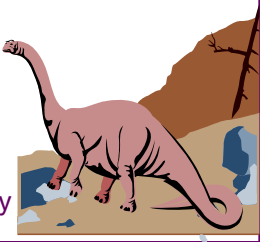


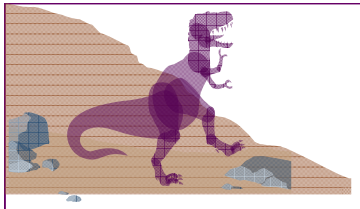




LOOK

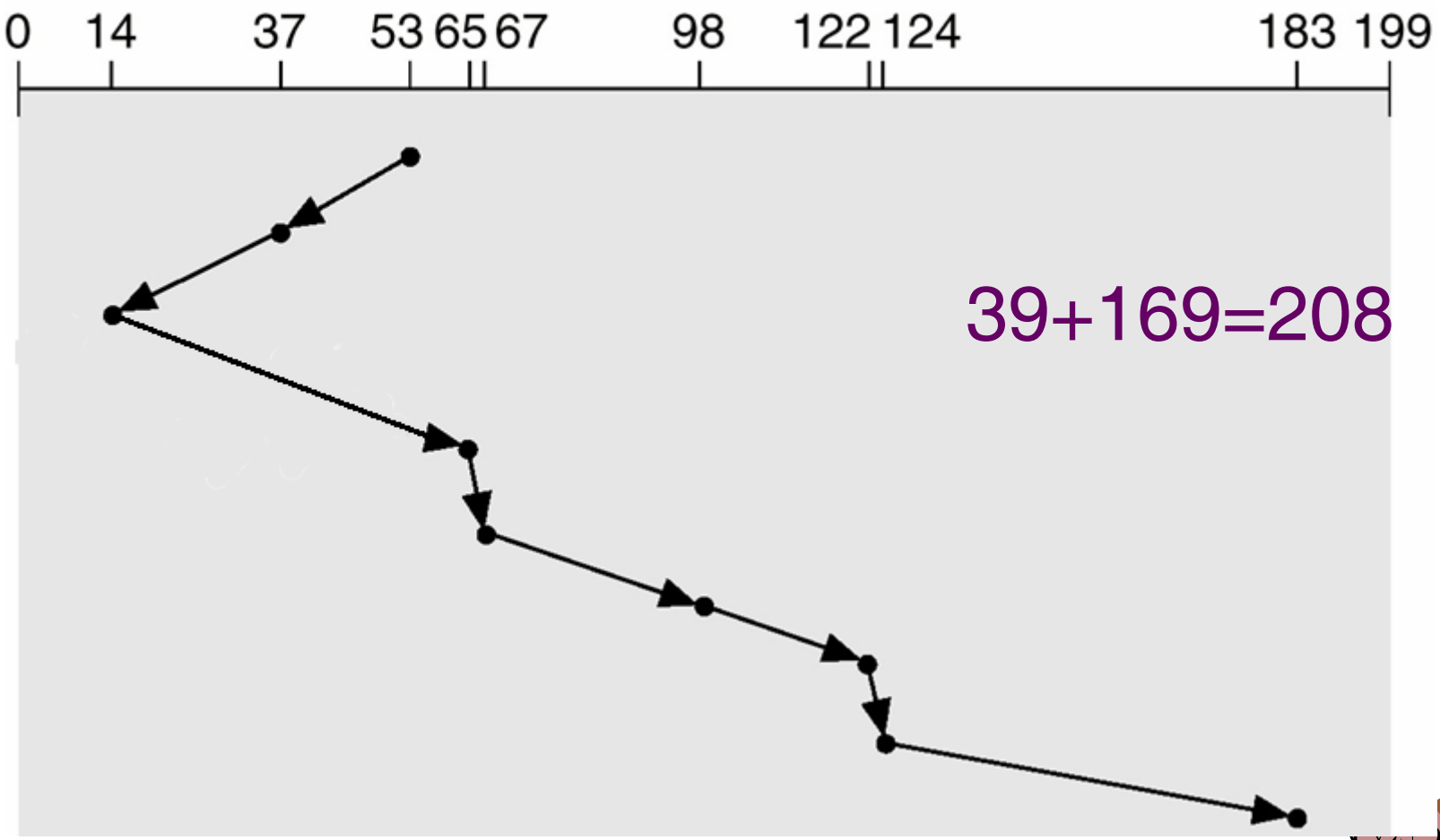
- Arm only goes as far as the last request in each direction, then reverses direction immediately, without first going all the way to the end of the disk.
- Illustration shows total head movement of 208 cylinders.



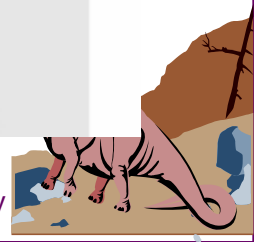


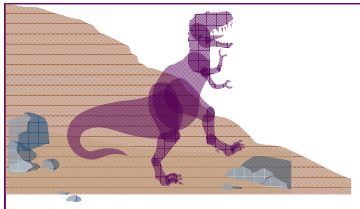
LOOK (cont.)

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



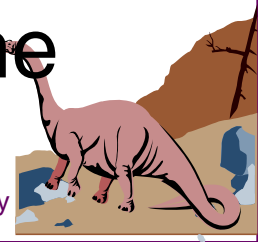
39+169=208

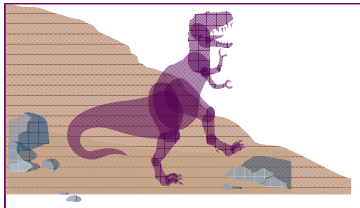




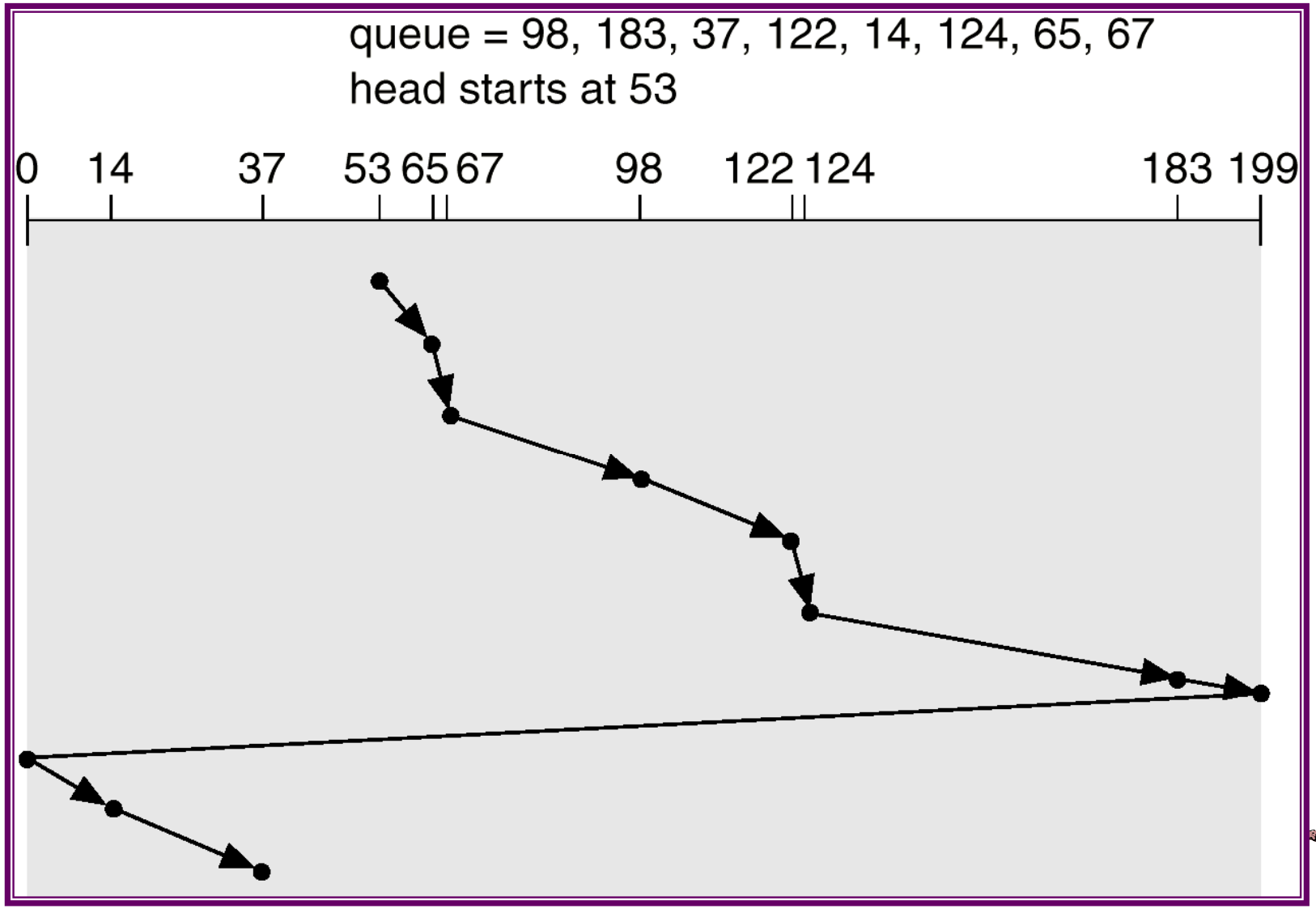
C-SCAN (Circular-SCAN)

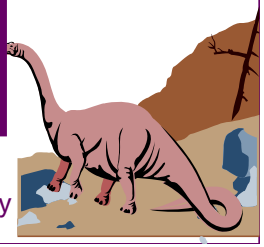
- Provides a more uniform wait time than SCAN.
- The head moves from one end of the disk to the other. Servicing requests as it goes. When it reaches the other end, however, it immediately returns to the beginning of the disk, without servicing any requests on the return trip.
- Treats the cylinders as a circular list that wraps around from the last cylinder to the first one.

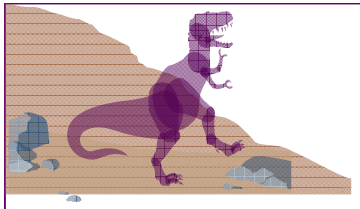




C-SCAN (cont.)

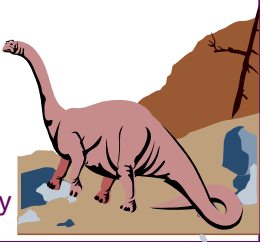

queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0
14
37
53
65
67
98
122
124
183
199

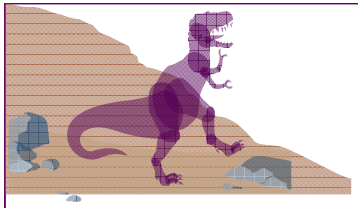




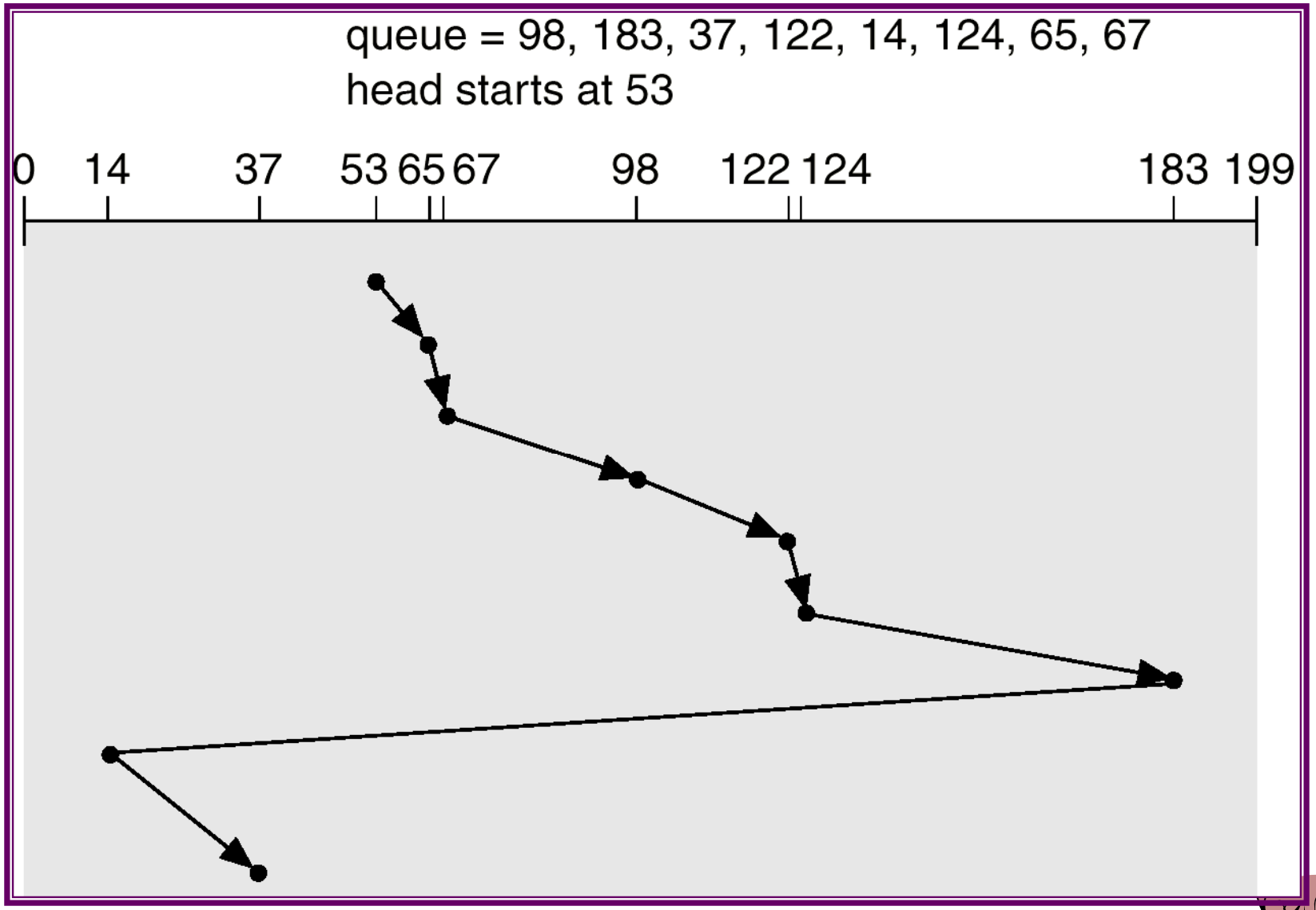
C-LOOK

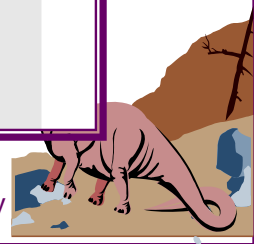
- Version of C-SCAN
- Arm only goes as far as the last request in each direction, then reverses direction immediately, without first going all the way to the end of the disk.

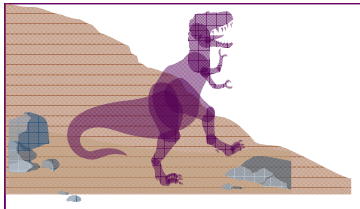




C-LOOK (cont.)

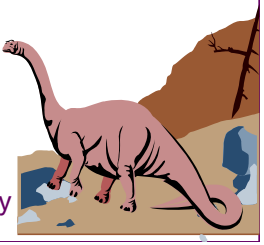

queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0
14
37
53
65
67
98
122
124
183
199

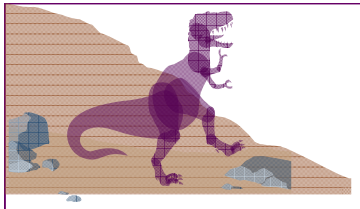




Select a Disk-Scheduling Algorithm

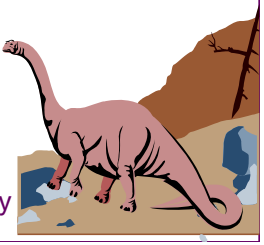
- SSTF is common and has a natural appeal
- SCAN and C-SCAN perform better for systems that place a heavy load on the disk.
- Performance depends on the number and types of requests.
- Requests for disk service can be influenced by the file-allocation method.

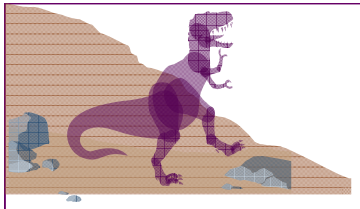




Select a Disk-Scheduling Algorithm (cont.)

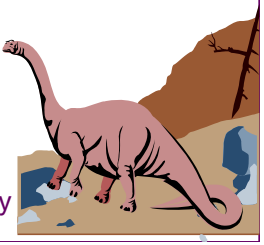
- The disk-scheduling algorithm should be written as a separate module of the operating system, allowing it to be replaced with a different algorithm if necessary.
  - Separation of policy from mechanism
- Either SSTF or LOOK is a reasonable choice for the default algorithm.

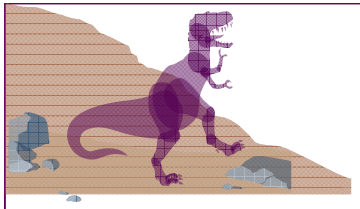




Disk Formatting

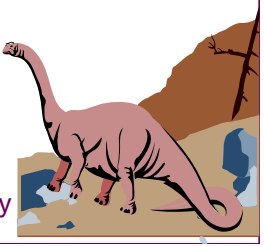
- *Low-level formatting*, or *physical formatting* — Dividing a disk into sectors that the disk controller can read and write.
- To use a disk to hold files, the operating system still needs to record its own data structures on the disk.
  - *Partition* the disk into one or more groups of cylinders.
  - *Logical formatting* or “making a file system”.

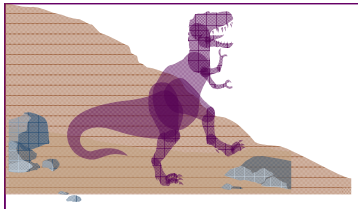




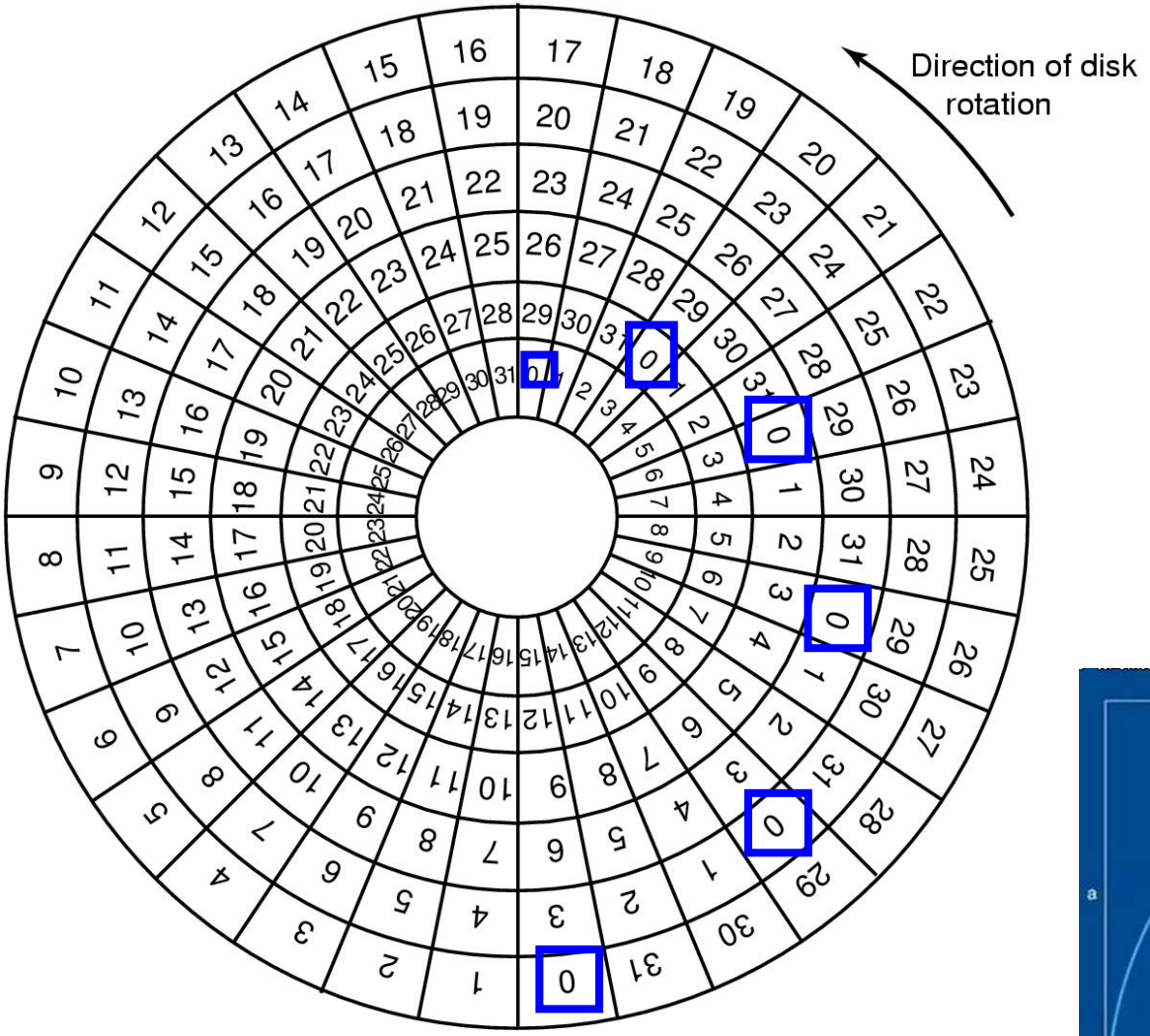
Disk Formatting (1)

A disk sector

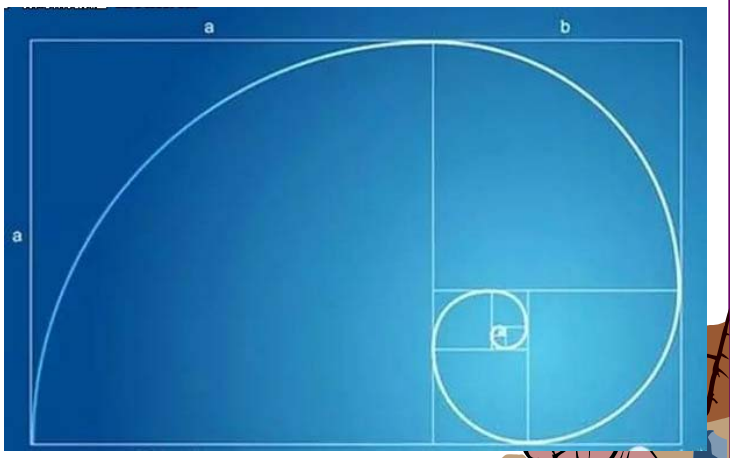


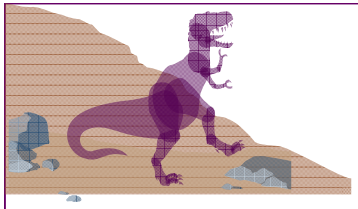


Disk Formatting (2)

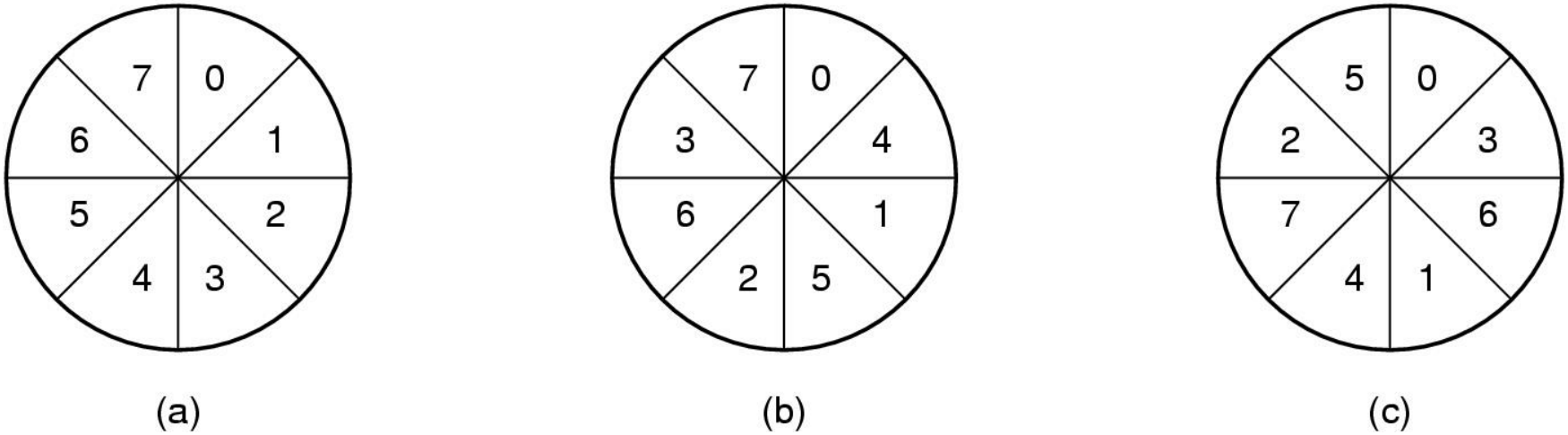


An illustration of cylinder skew

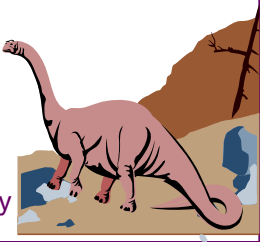




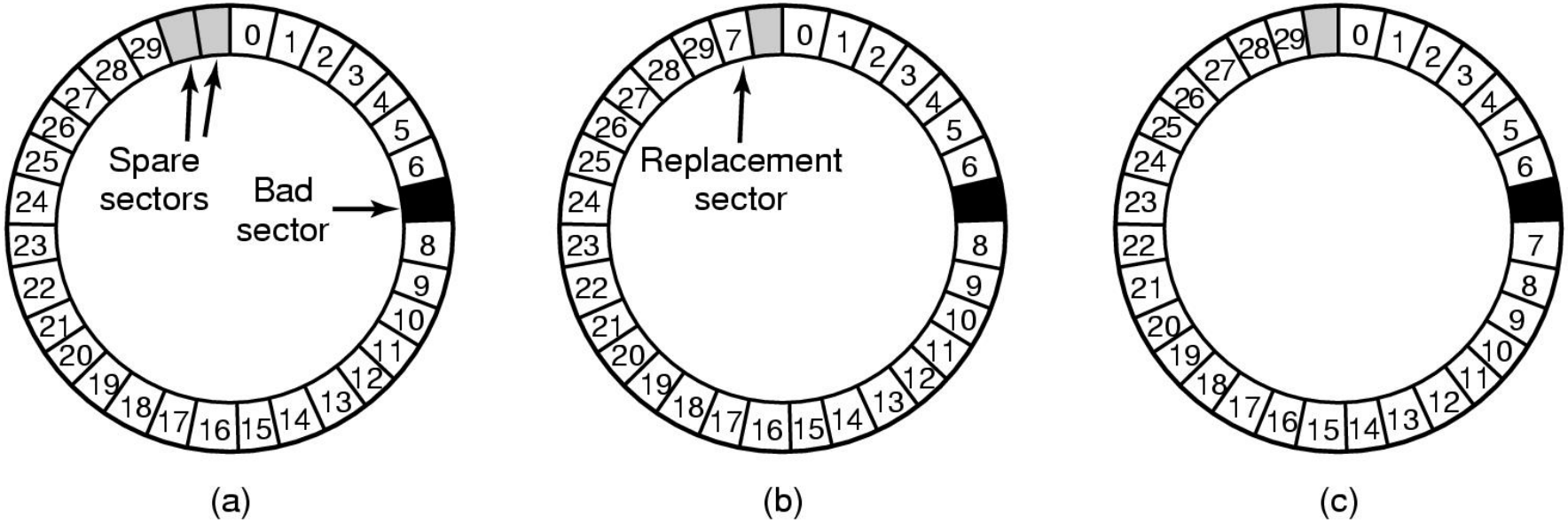
Disk Formatting (3)



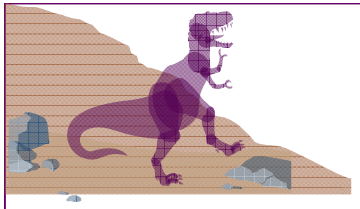
- No interleaving
- Single interleaving
- Double interleaving



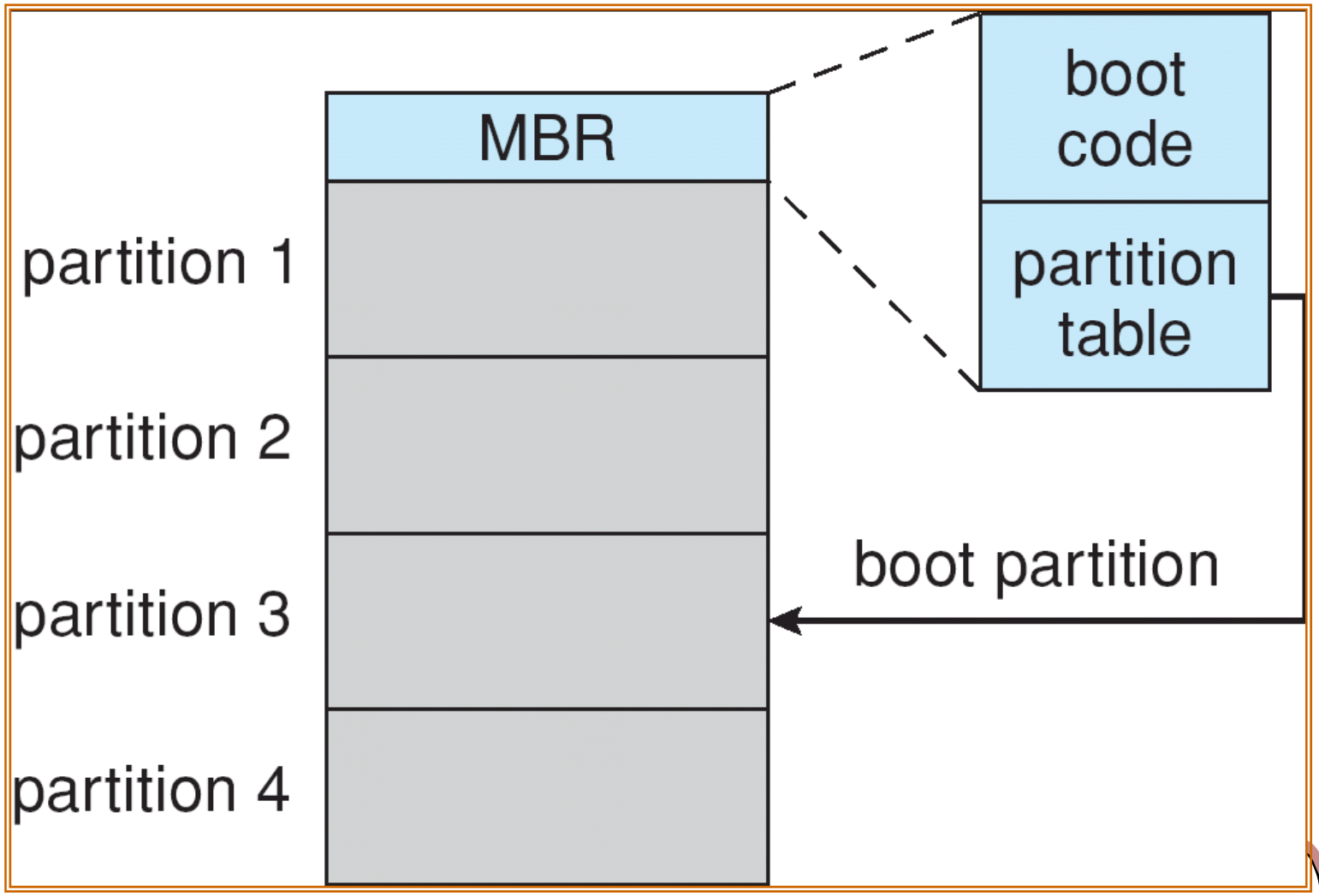
Error Handling

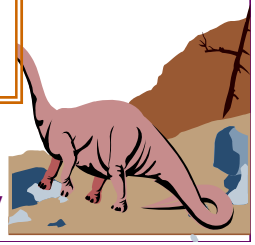


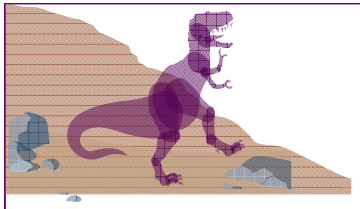
- A disk track with a bad sector
- Substituting a spare for the bad sector
- Shifting all the sectors to bypass the bad one



Booting from a Disk in Windows 2000





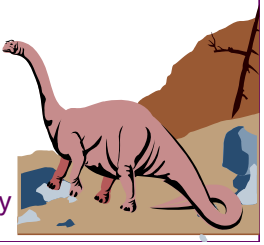


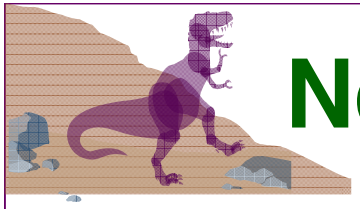
Disk Attachment

- Disks may be attached either of two ways:

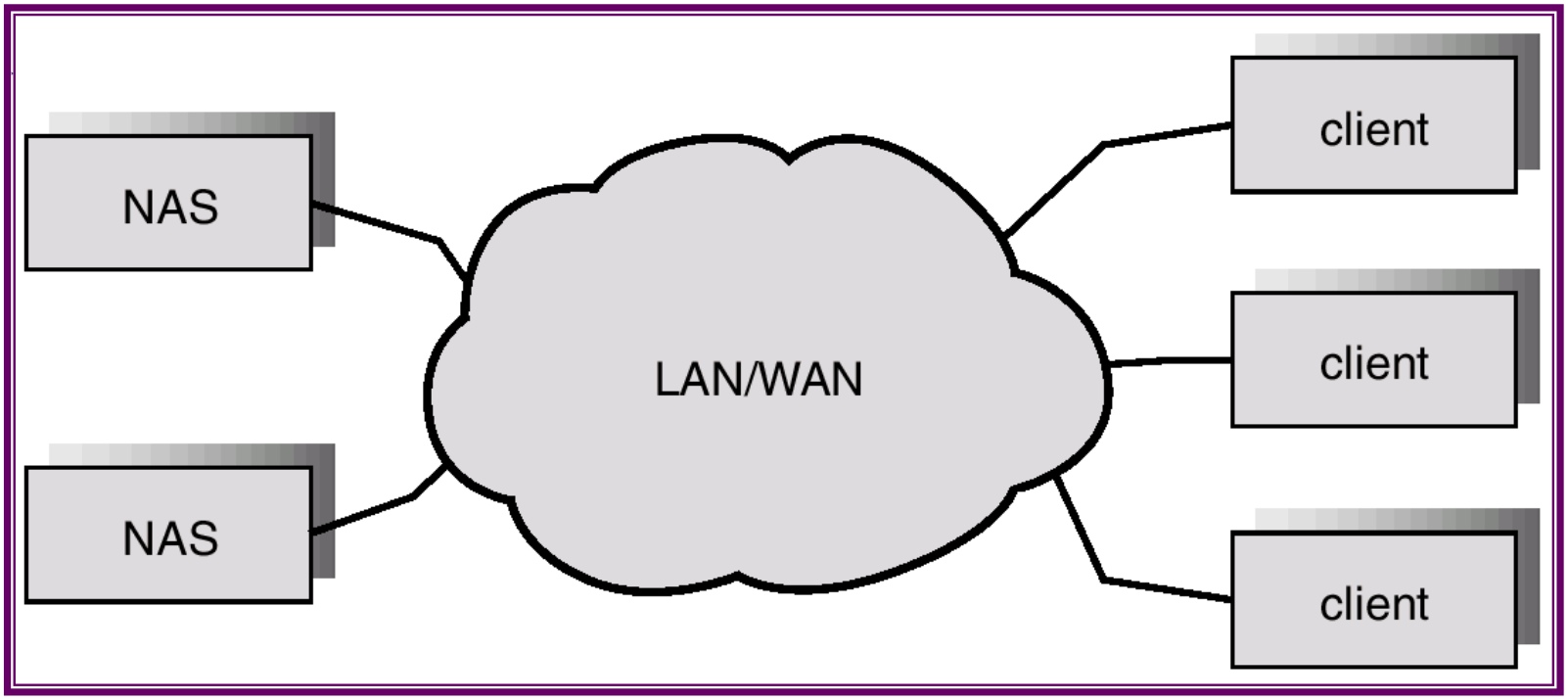
1. **Host attached** via an I/O port

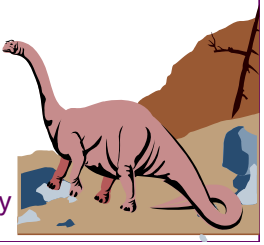
2. **Network attached** via a network connection

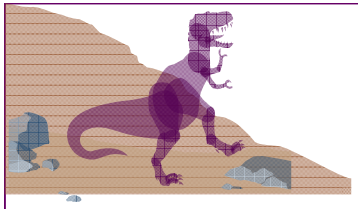




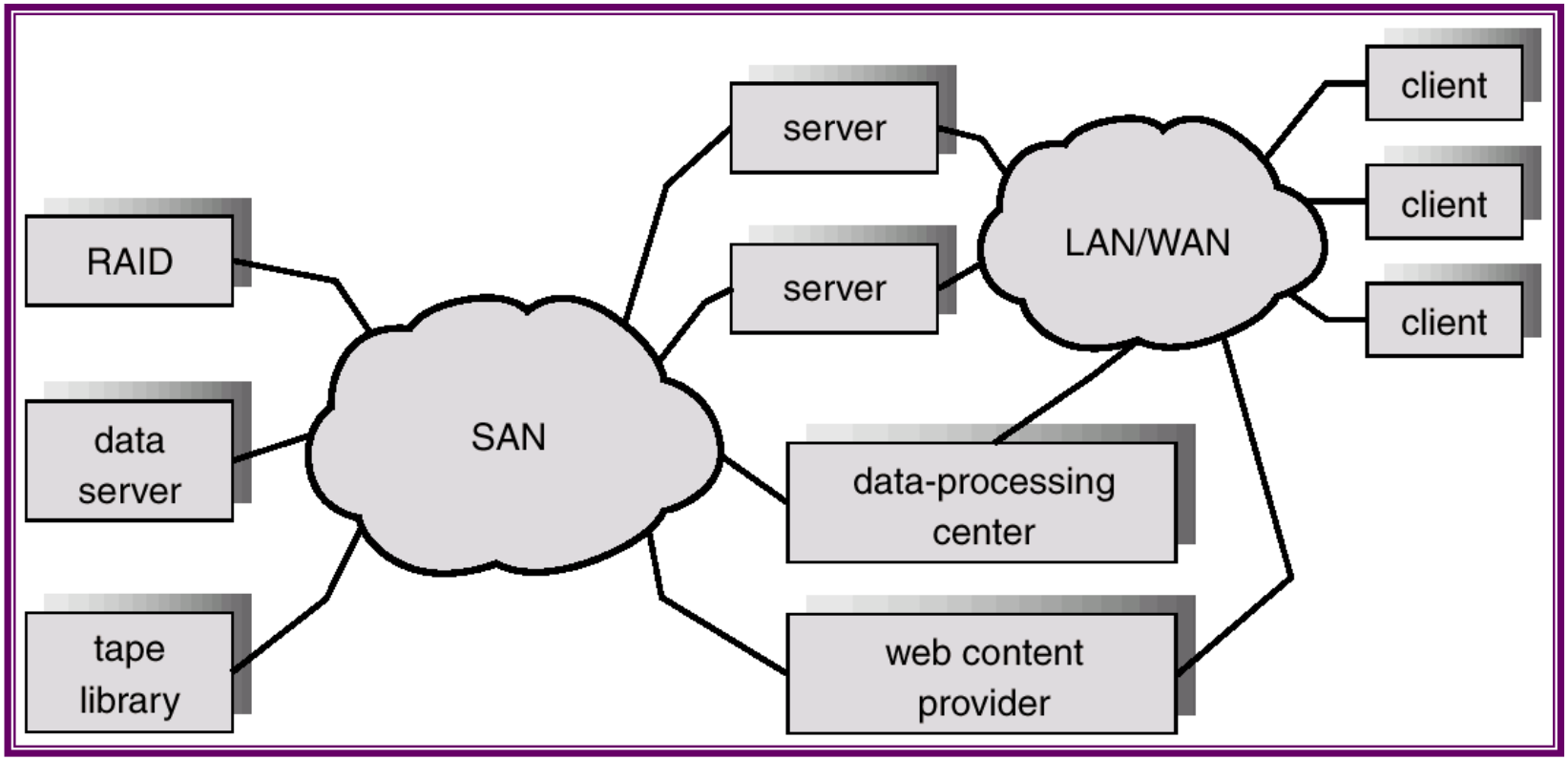
Network-Attached Storage (NAS)

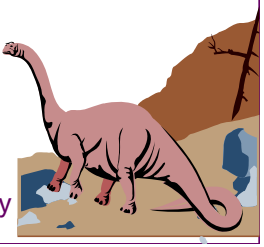


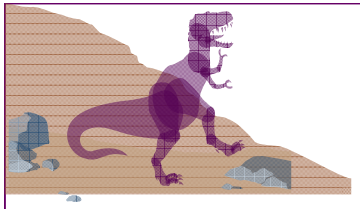




Storage-Area Network

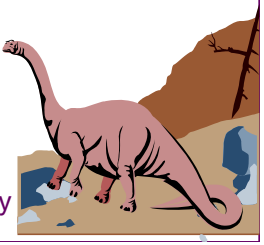


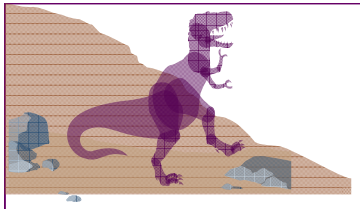




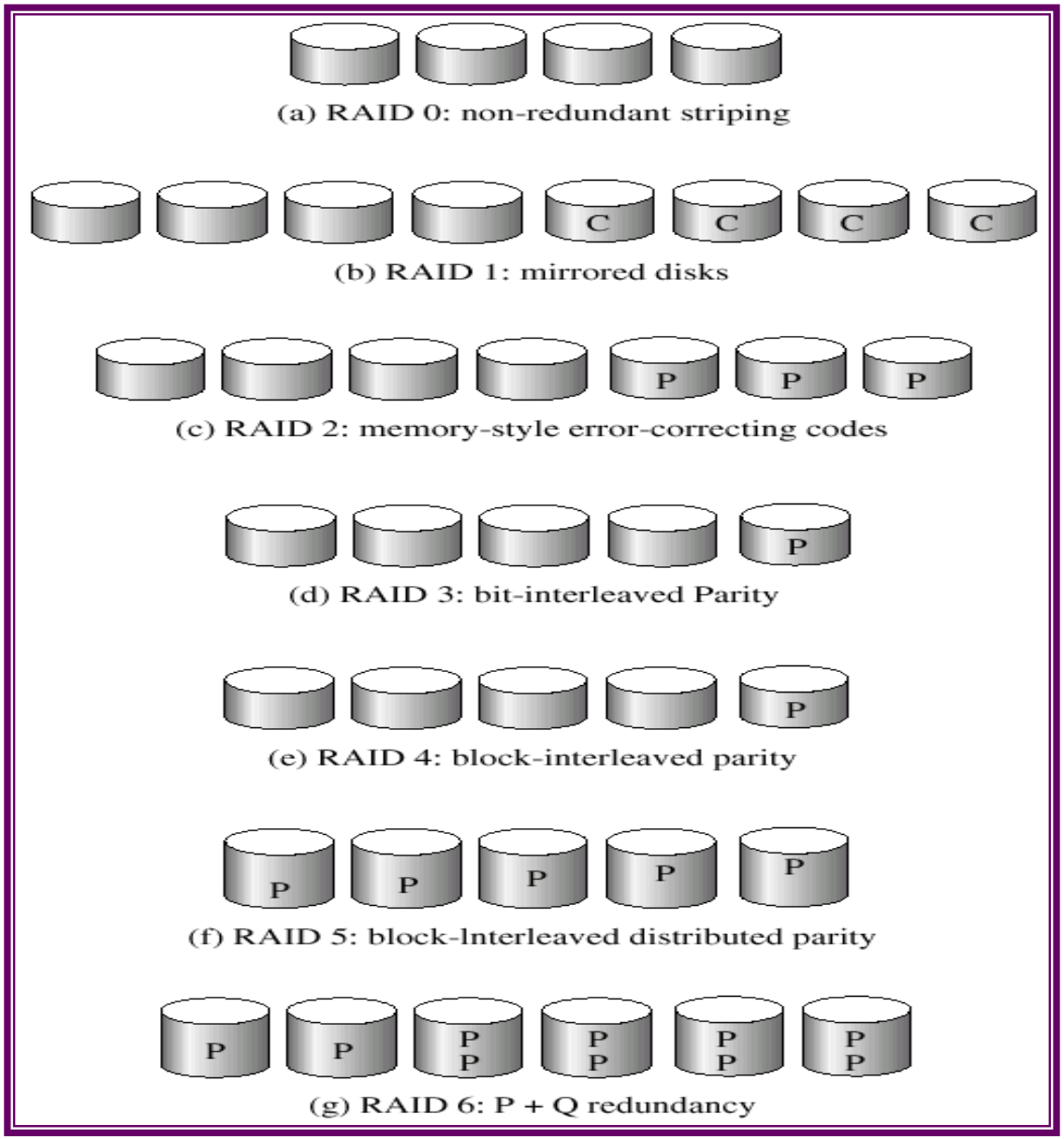
RAID Structure

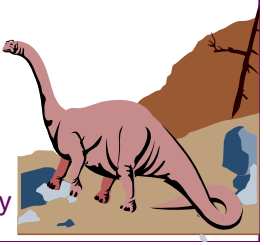
- **RAID** – multiple disk drives provides **reliability** via **redundancy**.
- RAID is arranged into six different levels.

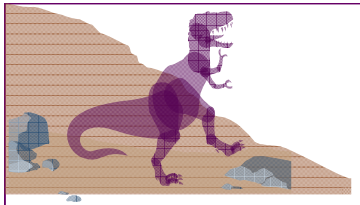




RAID Levels

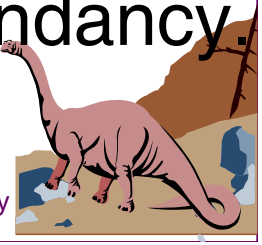

C
P
P
(a) RAID 0: non-redundant striping
(b) RAID 1: mirrored disks
(c) RAID 2: memory-style error-correcting codes
(d) RAID 3: bit-interleaved Parity
(e) RAID 4: block-interleaved parity
(f) RAID 5: block-Interleaved distributed parity
(g) RAID 6: P + Q redundancy

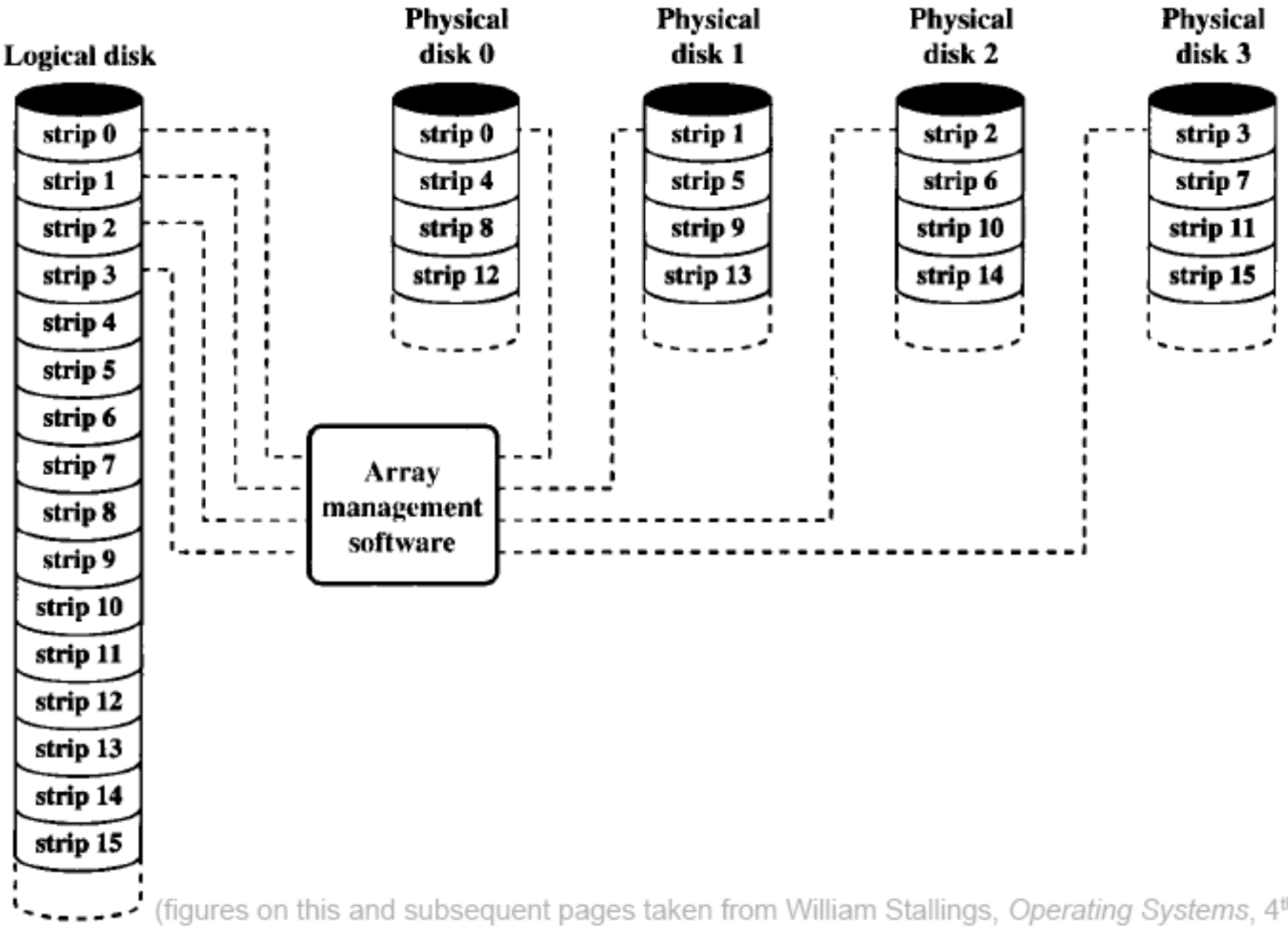




RAID (cont.)

- 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* keeps duplicate of each disk.
  - *Block interleaved parity* uses much less redundancy.





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