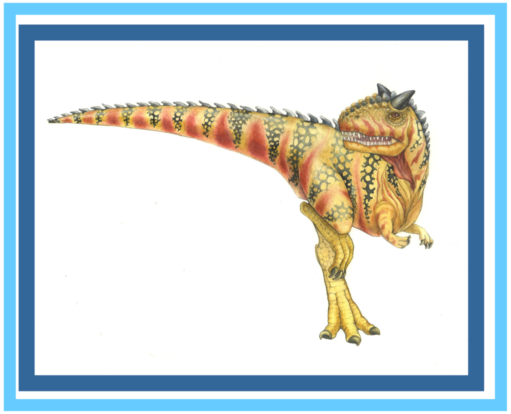
Chapter 9: Mass-Storage Systems



Chapter 9: Mass-Storage Systems

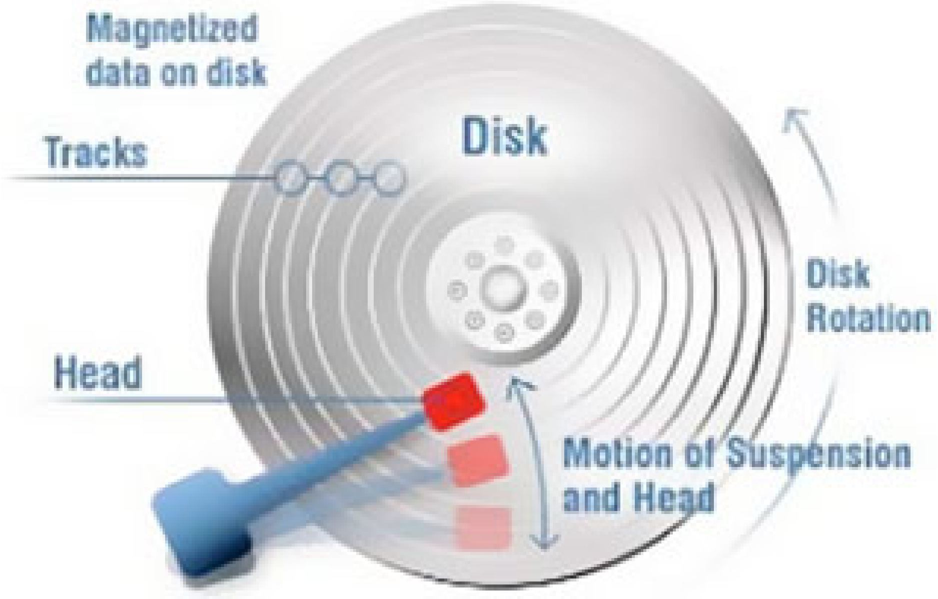
- Overview of Mass Storage Structure
- Disk Structure
- Disk Scheduling
- Disk Management
- RAID Structure
- Tertiary Storage Devices
- Performance Issues

Objectives

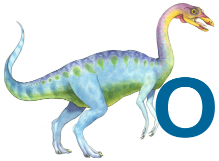
- Describe the physical structure of secondary and tertiary storage devices and the resulting effects on the uses of the devices
- Explain the performance characteristics of mass-storage devices
- Discuss operating-system services provided for mass storage, including RAID and HSM – Virtual Memory may fall into the latter category.

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
  - Illustration:



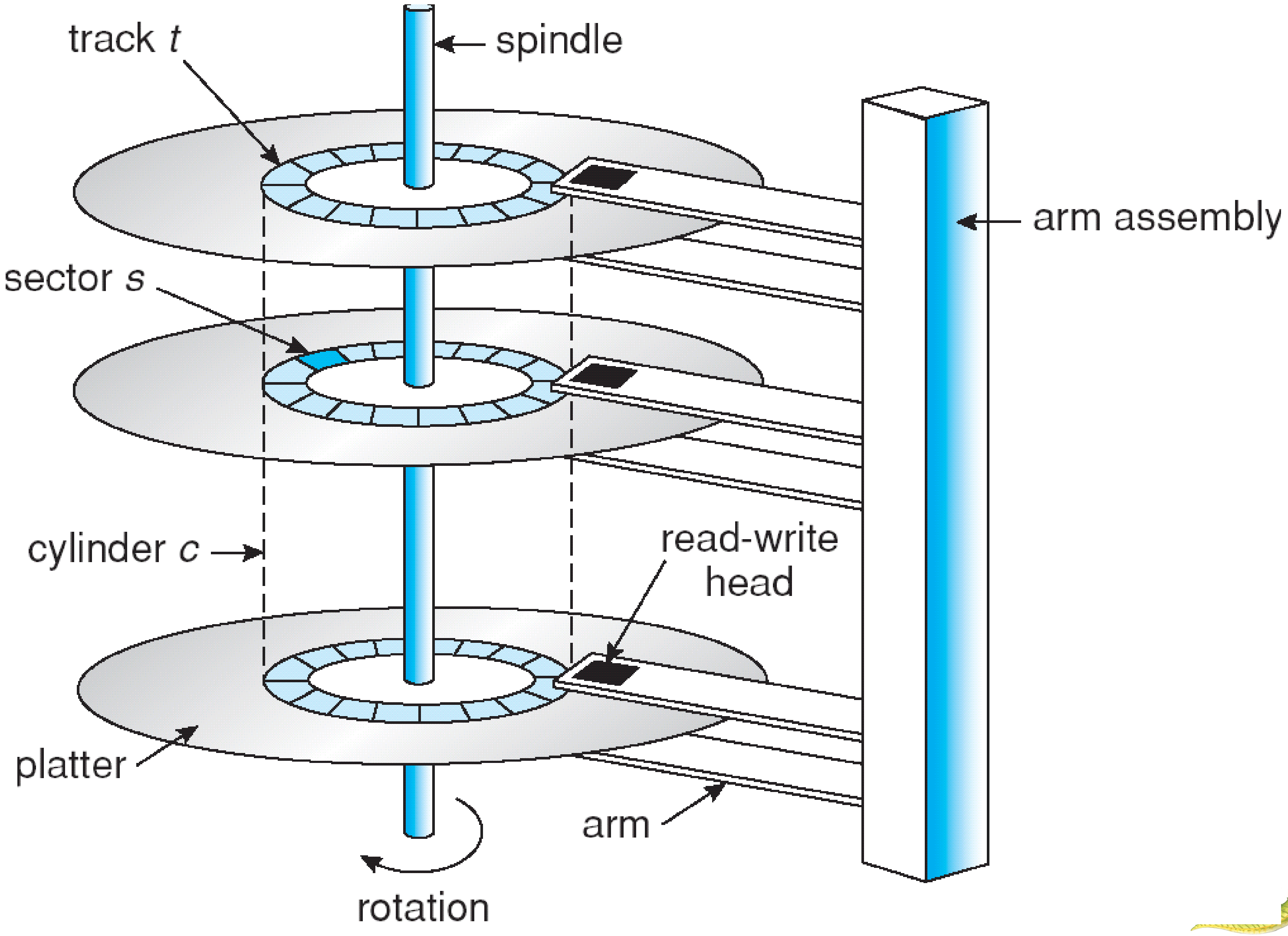
http://www.condusiv.com/images/Disk-Performance.jpg



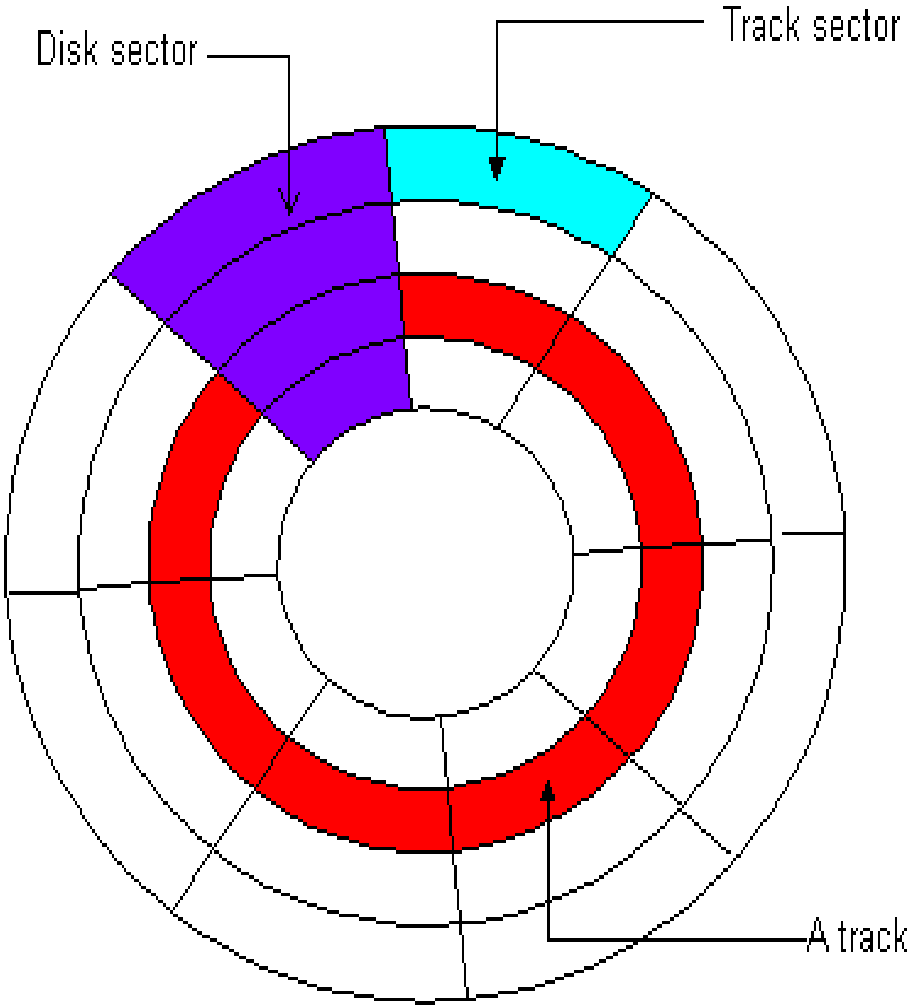
Overview of Mass Storage Structure

- Magnetic disks provide bulk of secondary storage of modern computers
  - Drives rotate at 60 to 200 times per second – 120 (7,200 rpm) being typical
  - Transfer rate is rate at which data flow between drive and computer – e.g., 1 GB/sec of sustained transfer is not uncommon
  - Positioning time (random-access time) is time to move disk arm to desired cylinder (seek time ~ 3ms – 15 ms) and time for desired sector to rotate under the disk head (rotational latency ~ 4 ms)
  - 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 (accessible through local I/O ports)
  - Buses vary; examples include EIDE, ATA, SATA, USB, Fibre Channel, SCSI
  - Host controller in computer uses bus to talk to disk controller built into drive or storage array
  - The above is usually referred to as host-attached storage
  - Another mode of disk attachment is referred to as network-attached storage (NAS)

Moving-head Disk Mechanism



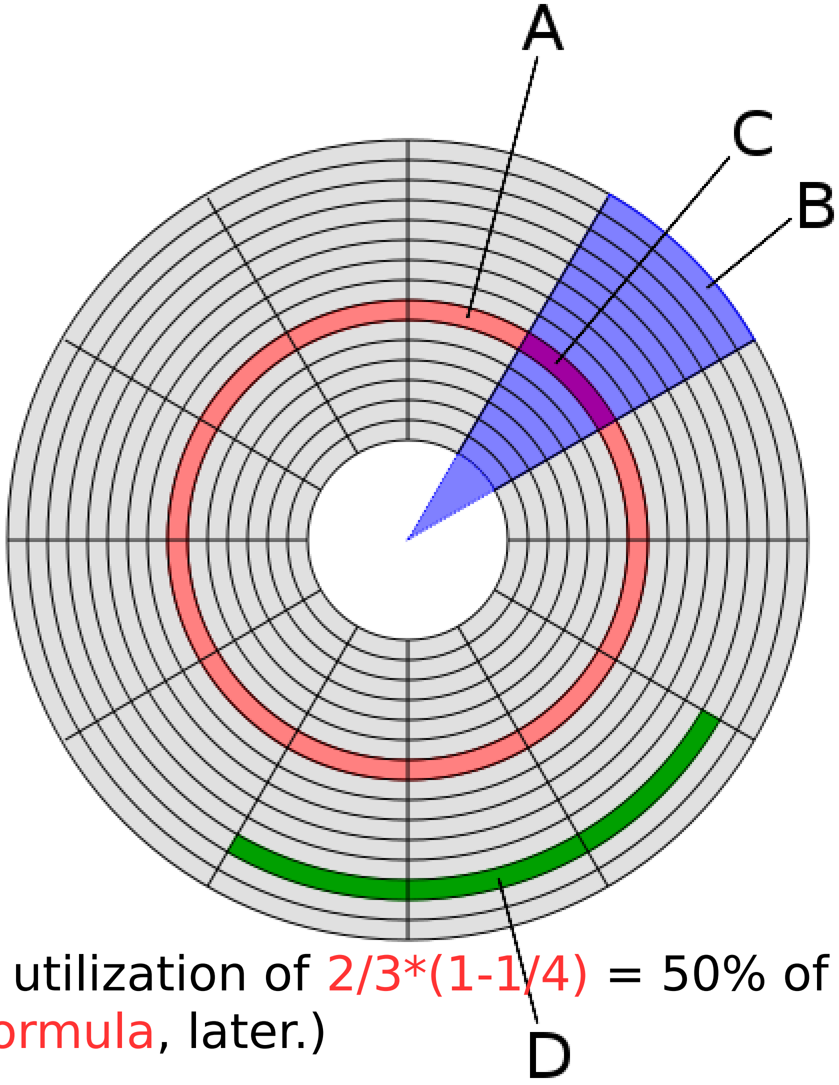
CAV Disk Addressing



Sectors and cylinders

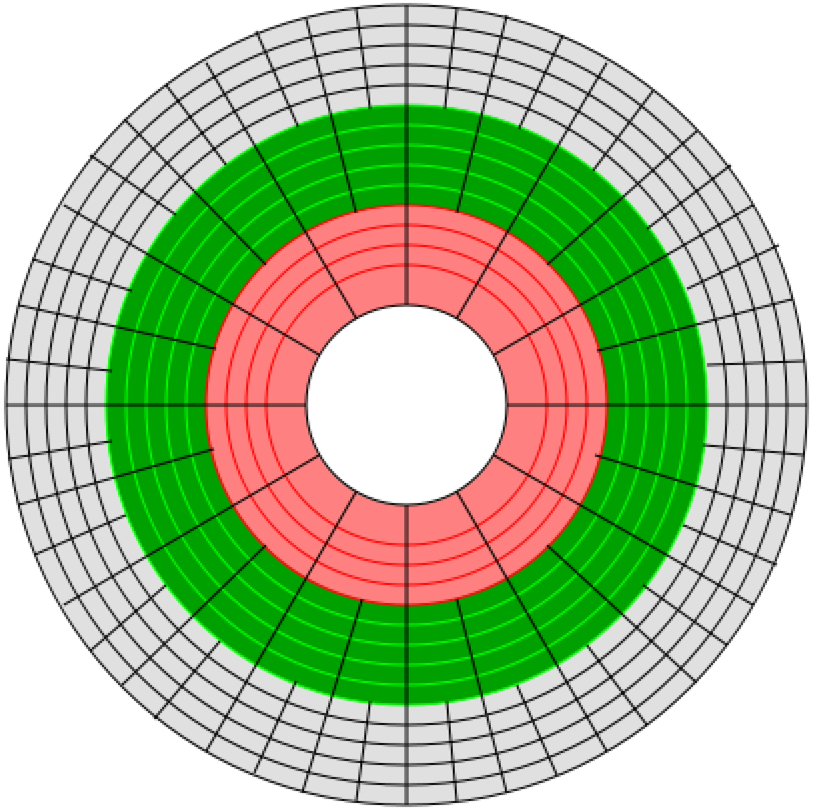
Disk usually contains several hundred sectors and may have tens of thousands of cylinders.

CAV Disk Addressing



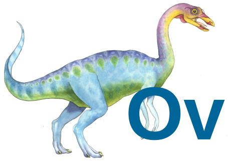
The above allows for utilization of 2/3*(1-1/4) = 50% of usable disk platter.
(We will derive this formula, later.)

Zones of Disk Recording



The above 3-zone formatting allows for utilization of 2/3*(1-1/64) ≈ 65.6% of usable platter. (We will derive this formula, later.)

Note: 2/3 ≈ 66.7% is the upper bound for multi-zoned CAV disks.



Overview of Mass Storage Structure (Cont)

- Magnetic tape
  - Was early secondary-storage medium
  - Relatively permanent (30+ yrs) and holds large quantities of data
  - Access time slow
  - Random access ~1,000 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 per reel typical storage (330 TB max.)
  - 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 ≈ seek distance between cylinders
  - For small distances, the actual seek time is considerably larger
- Rotational latency; on avg. ~ ½ of full rotation time of the disk
- 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)

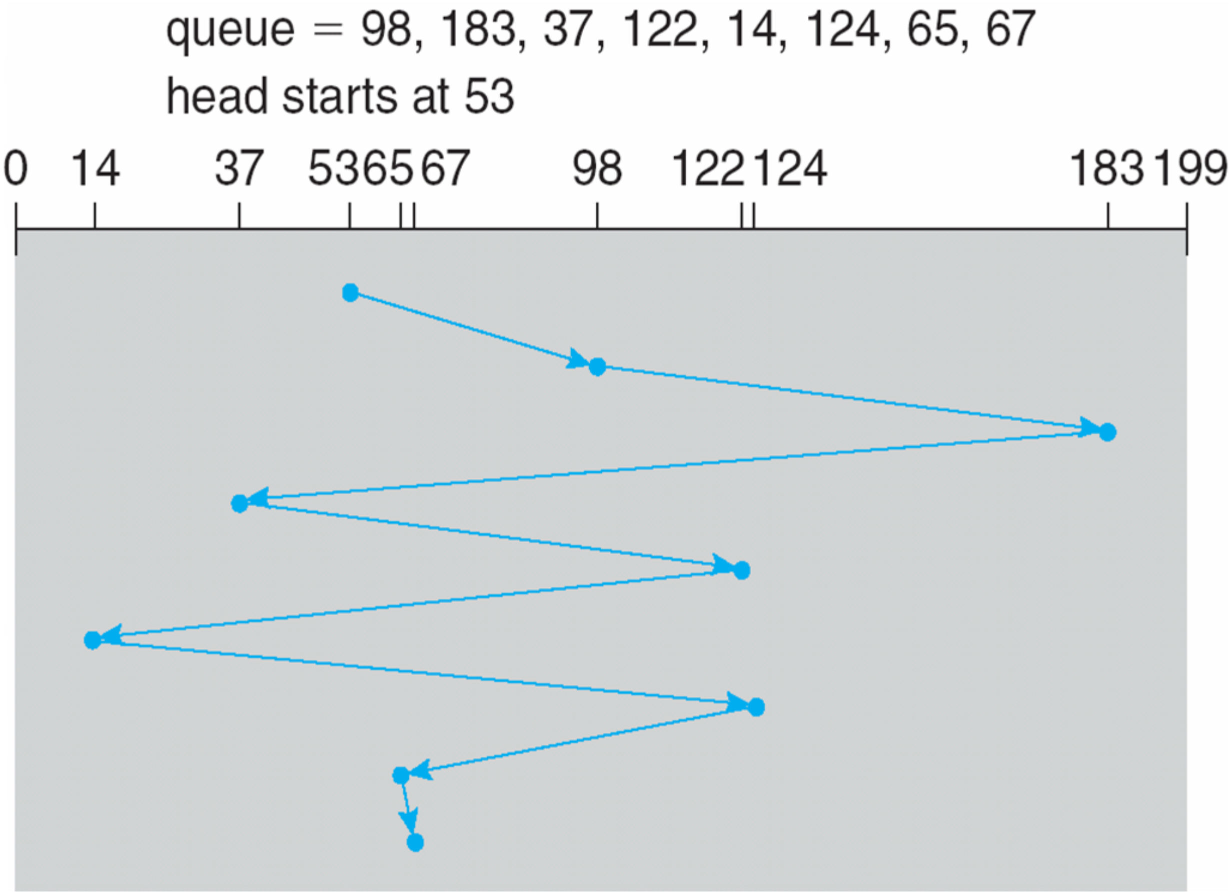
- Several algorithms exist to schedule the servicing of disk I/O requests
- We illustrate them with a request queue (cyllinders 0-199)

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

Head pointer 53

FCFS

Illustration shows total head movement of 640 cylinders



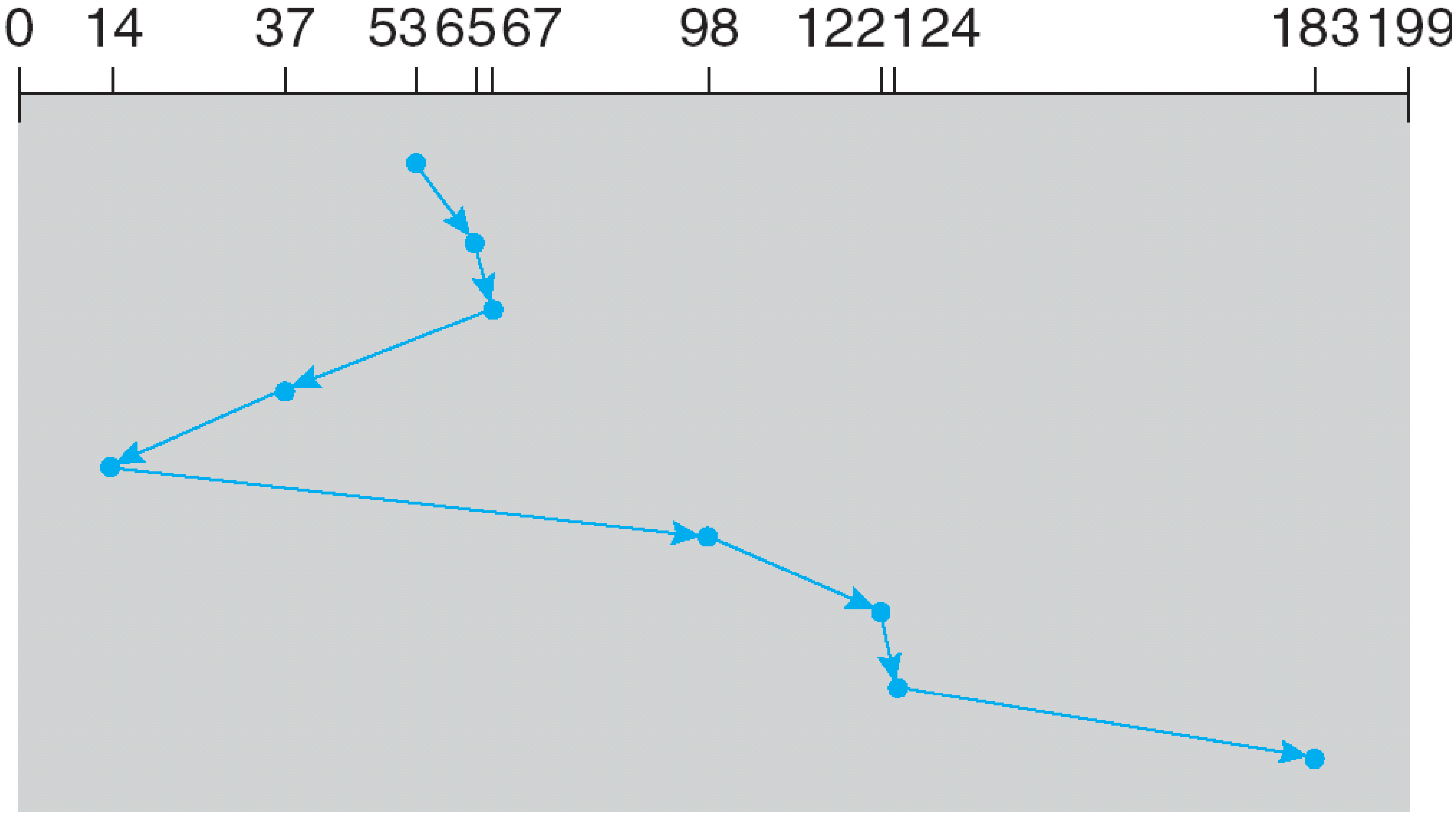
SSTF

- 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)

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

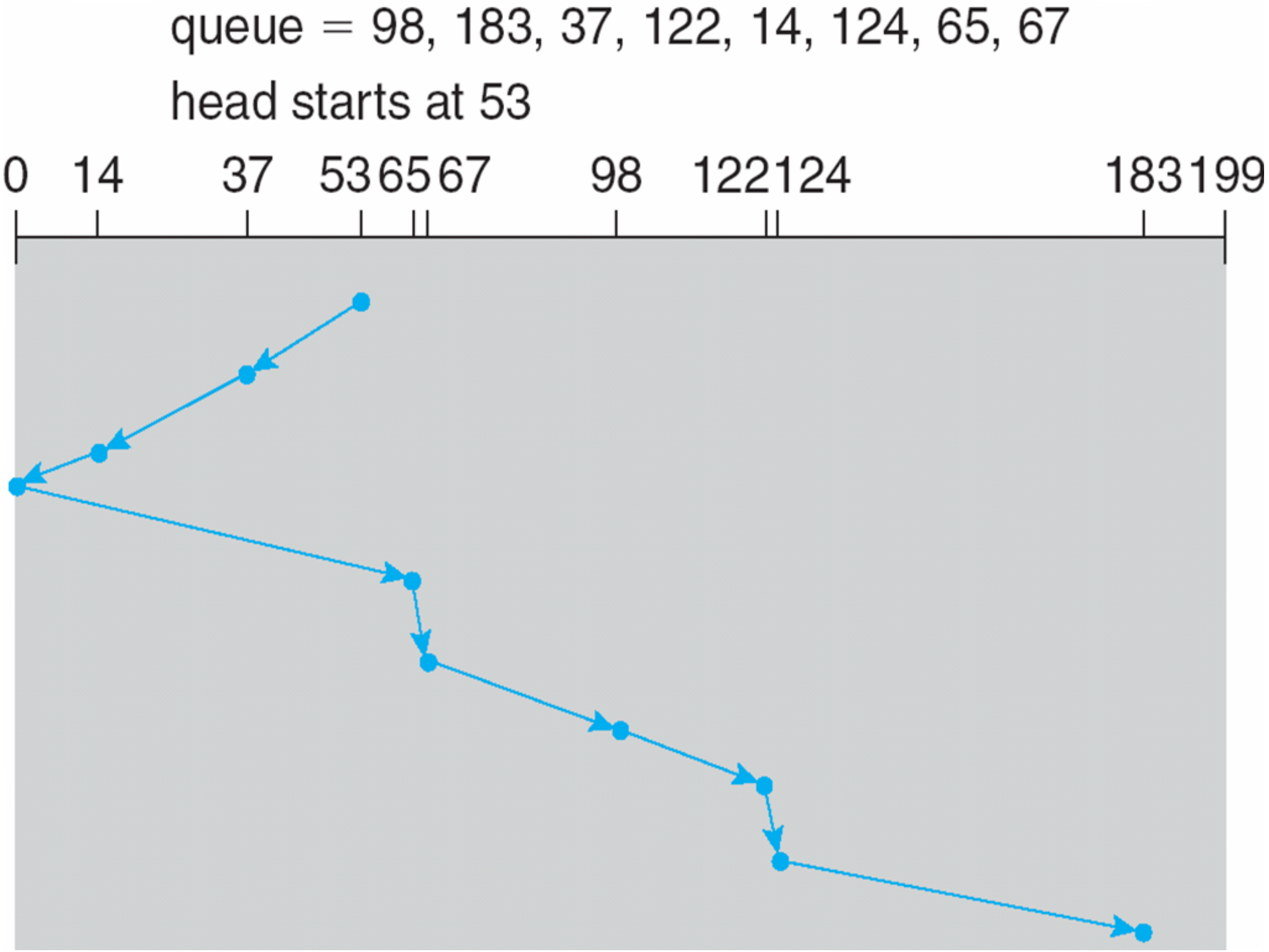
head starts at 53



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.
- SCAN algorithm Sometimes called the elevator algorithm
- Illustration shows total head movement of 208 cylinders

SCAN (Cont.)



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

head starts at 53

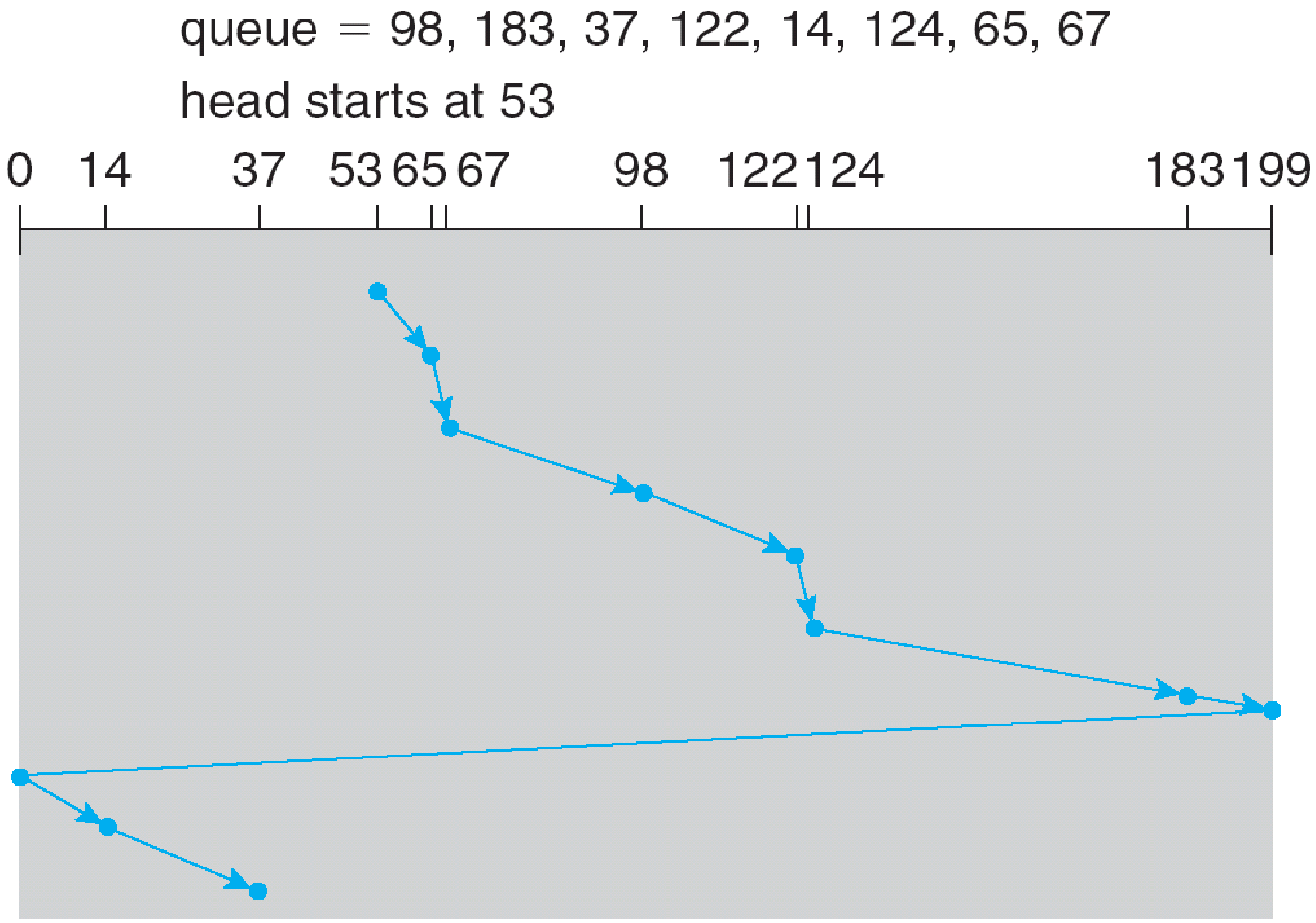
C-SCAN

- Takes advantage of the fact that seek time of moving the head over large number K of cylinders is considerably less than

  K/M * seek time moving the head over a small number M of cylinders

- Provides a shorter average 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
- Illustration shows total head movement of 183 cylinders plus the (fast) return over 199 cylinders (from cylinder 199 to cylinder 0).

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

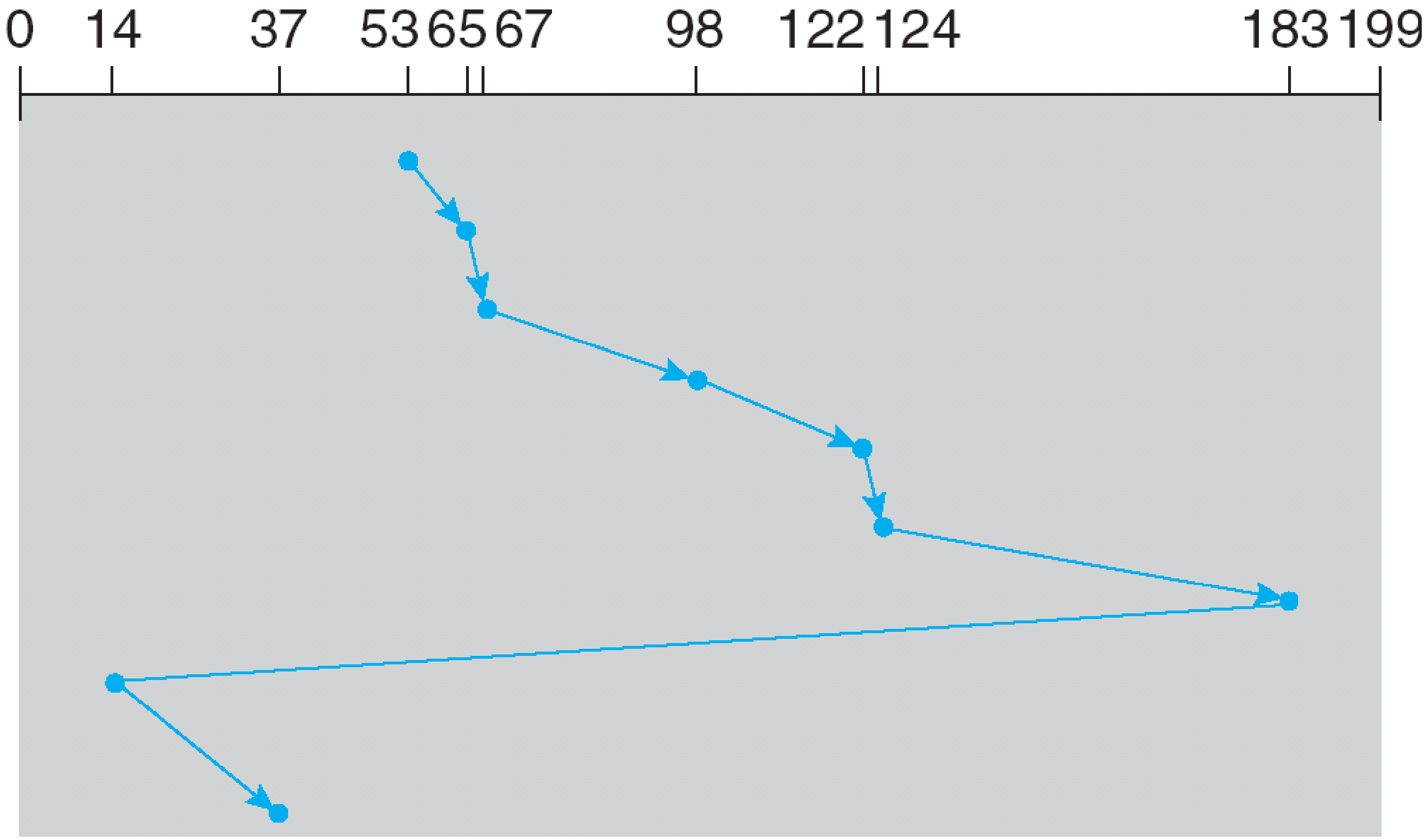
C-LOOK

- 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
- Illustration shows total head movement of 153 cylinders plus the (fast) return over 169 cylinders (from cylinder 183 to cylinder 14).

C-LOOK (Cont)

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
- 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 C-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”
  - To increase efficiency most file systems group blocks into clusters
    - Disk I/O done in blocks
    - File I/O done in clusters
- Boot block initializes system
  - The bootstrap is stored in ROM
  - Bootstrap loader program is stored on the disk in boot partition (a.k.a. boot disk or system disk)

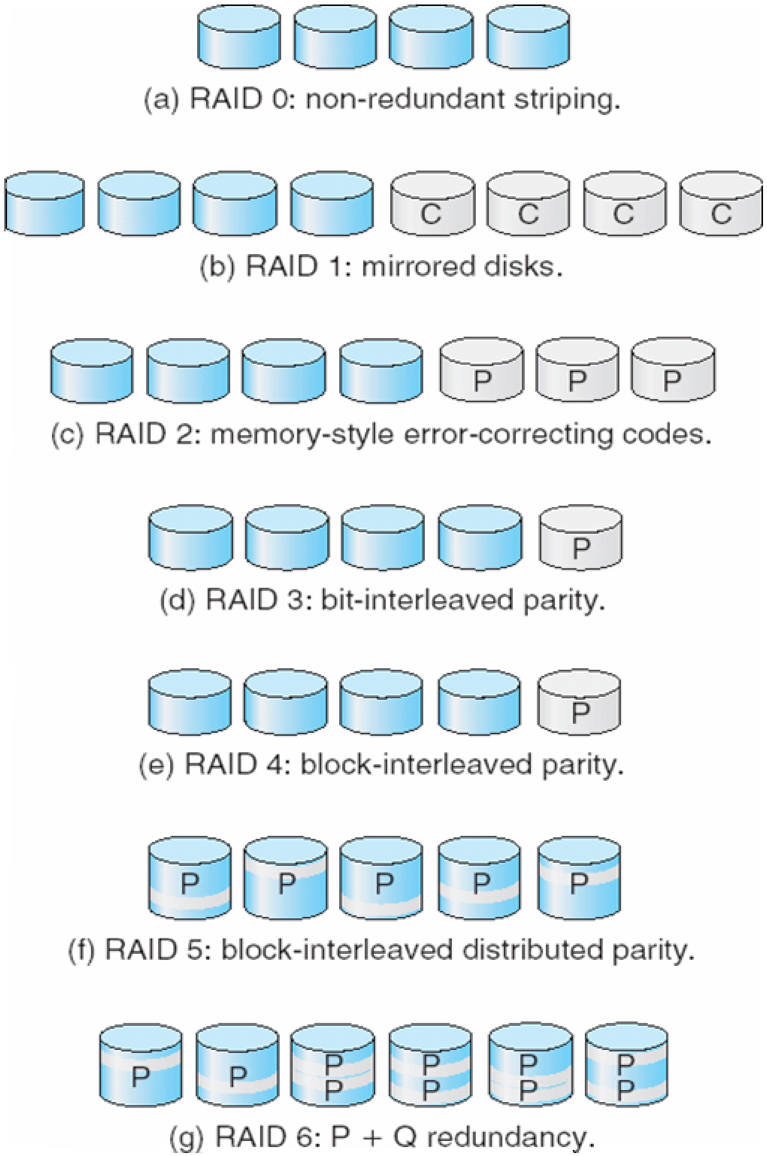
Disk Management

- Example: Windows OS
- Boot block (block 0 on the disk) initializes system
  - The bootstrap is stored in ROM
  - Bootstrap loader program is stored on the disk in master boot record MBR
  - Boot partition contains boot sector (the rest of boot code), codes OS and device drivers, and other system info.
- Methods such as sector sparing used to handle bad blocks

RAID Structure

- RAID – multiple disk drives provides reliability via redundancy
- Increases the mean time to failure
- Frequently combined with NVRAM to improve write performance
- RAID may be arranged into any of six different levels

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.

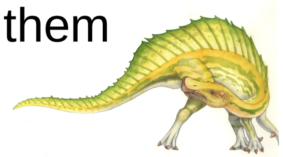
(e) RAID 4: block-interleaved parity.

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

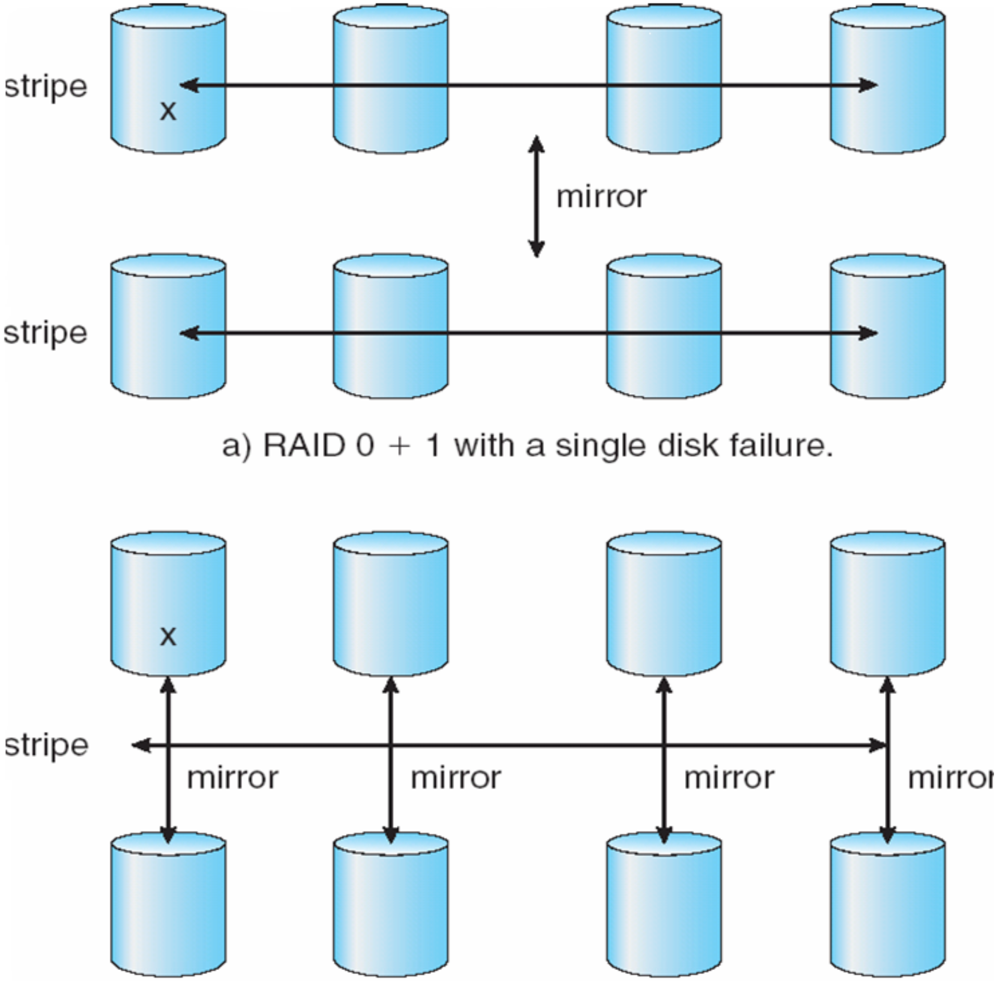
(g) RAID 6: P + Q redundancy.

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



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

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

Tertiary Storage Devices

- Low cost is the defining characteristic of tertiary storage
- Generally, tertiary storage is built using removable media
- Common examples of removable media are floppy disks and CD-ROMs; other types are available

Removable Disks

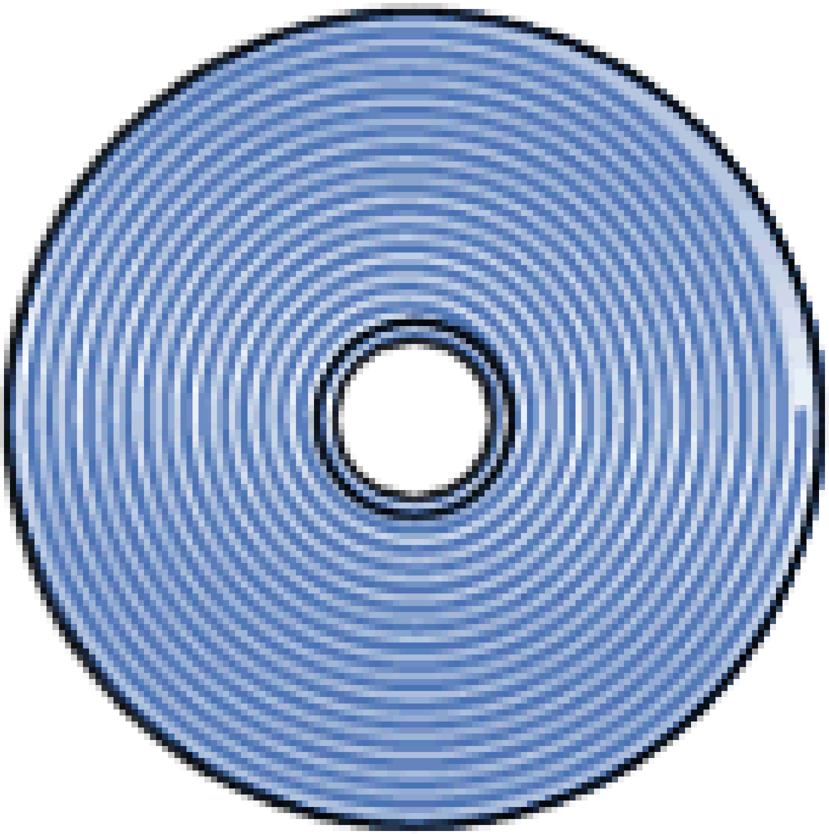
- Floppy disk — thin flexible disk coated with magnetic material, enclosed in a protective plastic case
  - Most floppies hold about 1 MB; similar technology is used for removable disks that hold more than 1 GB
  - Removable magnetic disks can be nearly as fast as hard disks, but they are at a greater risk of damage from exposure

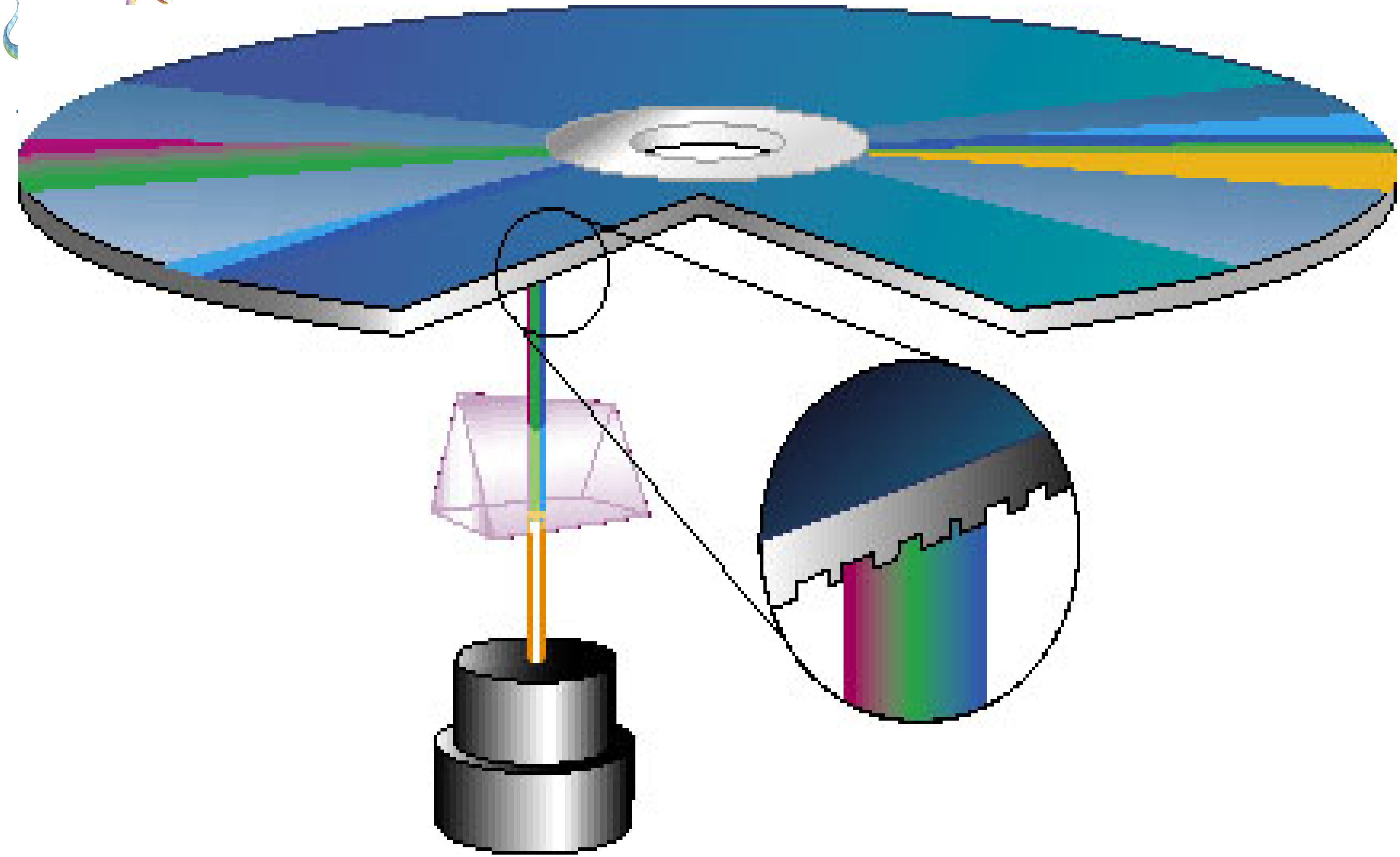
Removable Disks (Cont.)

- A magneto-optic disk records data on a rigid platter coated with magnetic material
  - Laser heat is used to amplify a large, weak magnetic field to record a bit
  - Laser light is also used to read data (Kerr effect)
  - The magneto-optic head flies much farther from the disk surface than a magnetic disk head, and the magnetic material is covered with a protective layer of plastic or glass; resistant to head crashes
- Optical disks do not use magnetism; they employ special materials that are altered by laser light

•Optical Disks

- 1 spiral track
- CLV (constant linear velocity; angular velocity varies – for a 12X CD between 2,400 – 6,000 RPM)





© 2008 Prentice-Hall, Inc.

Slide 35

WORM Disks

- The data on read-write disks can be modified over and over
- WORM (“Write Once, Read Many Times”) disks can be written only once
- Thin aluminum film sandwiched between two glass or plastic platters
- To write a bit, the drive uses a laser light to burn a small hole through the aluminum; information can be destroyed by not altered
- Very durable and reliable
- Read-only disks, such ad CD-ROM and DVD, com from the factory with the data pre-recorded

Tapes

- Compared to a disk, a tape is less expensive and holds more data, but random access is much slower
- Tape is an economical medium for purposes that do not require fast random access, e.g., backup copies of disk data, holding huge volumes of data
- Large tape installations typically use robotic tape changers that move tapes between tape drives and storage slots in a tape library
  - stacker – library that holds a few tapes
  - silo – library that holds thousands of tapes
- A disk-resident file can be archived to tape for low cost storage; the computer can stage it back into disk storage for active use

Hierarchical Storage Management (HSM)

- A hierarchical storage system extends the storage hierarchy beyond primary memory and secondary storage to incorporate tertiary storage — usually implemented as a jukebox of tapes or removable disks
- Usually incorporate tertiary storage by extending the file system
  - Small and frequently used files remain on disk
  - Large, old, inactive files are archived to the jukebox
- HSM is usually found in supercomputing centers and other large installations that have enormous volumes of data

Speed

- Two aspects of speed in tertiary storage are bandwidth and latency
- Bandwidth is measured in bytes per second
  - Sustained bandwidth – average data rate during a large transfer; # of bytes/transfer time
    Data rate when the data stream is actually flowing
  - Effective bandwidth – average over the entire I/O time, including `seek()` or `locate()`, and cartridge switching
    Drive’s overall data rate

Speed (Cont)

- Access latency – amount of time needed to locate data
  - Access time for a disk – move the arm to the selected cylinder and wait for the rotational latency; < 35 milliseconds
  - Access on tape requires winding the tape reels until the selected block reaches the tape head; tens or hundreds of seconds
  - Generally say that random access within a tape cartridge is about a thousand times slower than random access on disk
- The low cost of tertiary storage is a result of having many cheap cartridges share a few expensive drives
- A removable library is best devoted to the storage of infrequently used data, because the library can only satisfy a relatively small number of I/O requests per hour

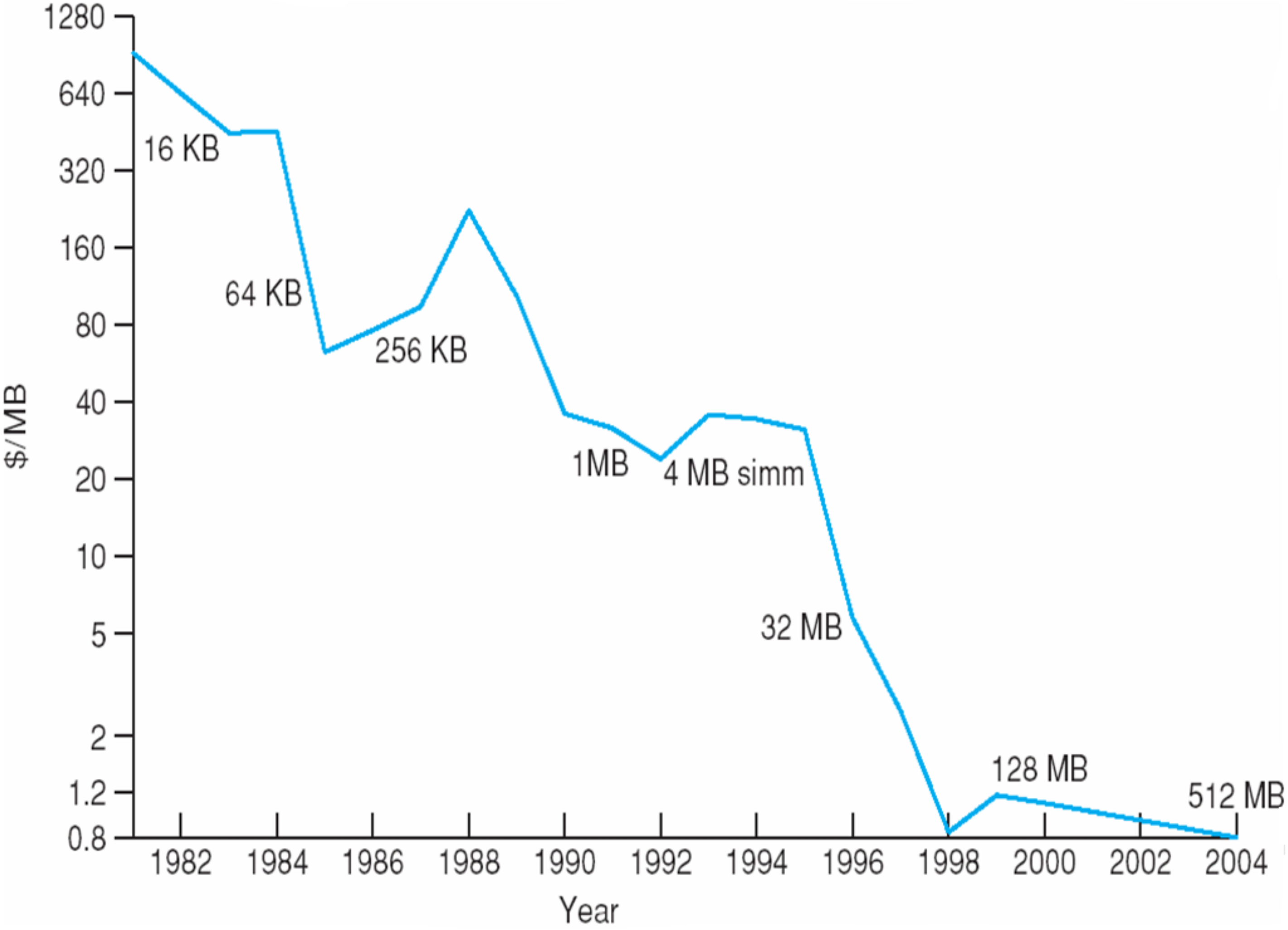
Reliability

- A fixed disk drive is likely to be more reliable than a removable disk or tape drive
- An optical cartridge is likely to be more reliable than a magnetic disk or tape
- A head crash in a fixed hard disk generally destroys the data, whereas the failure of a tape drive or optical disk drive often leaves the data cartridge unharmed

Cost

- Main memory is much more expensive than disk storage
- The cost per megabyte of hard disk storage is competitive with magnetic tape if only one tape is used per drive
- The cheapest tape drives and the cheapest disk drives have had about the same storage capacity over the years
- Tertiary storage gives a cost savings only when the number of cartridges is considerably larger than the number of drives

Price per Megabyte of DRAM, From 1981 to 2004



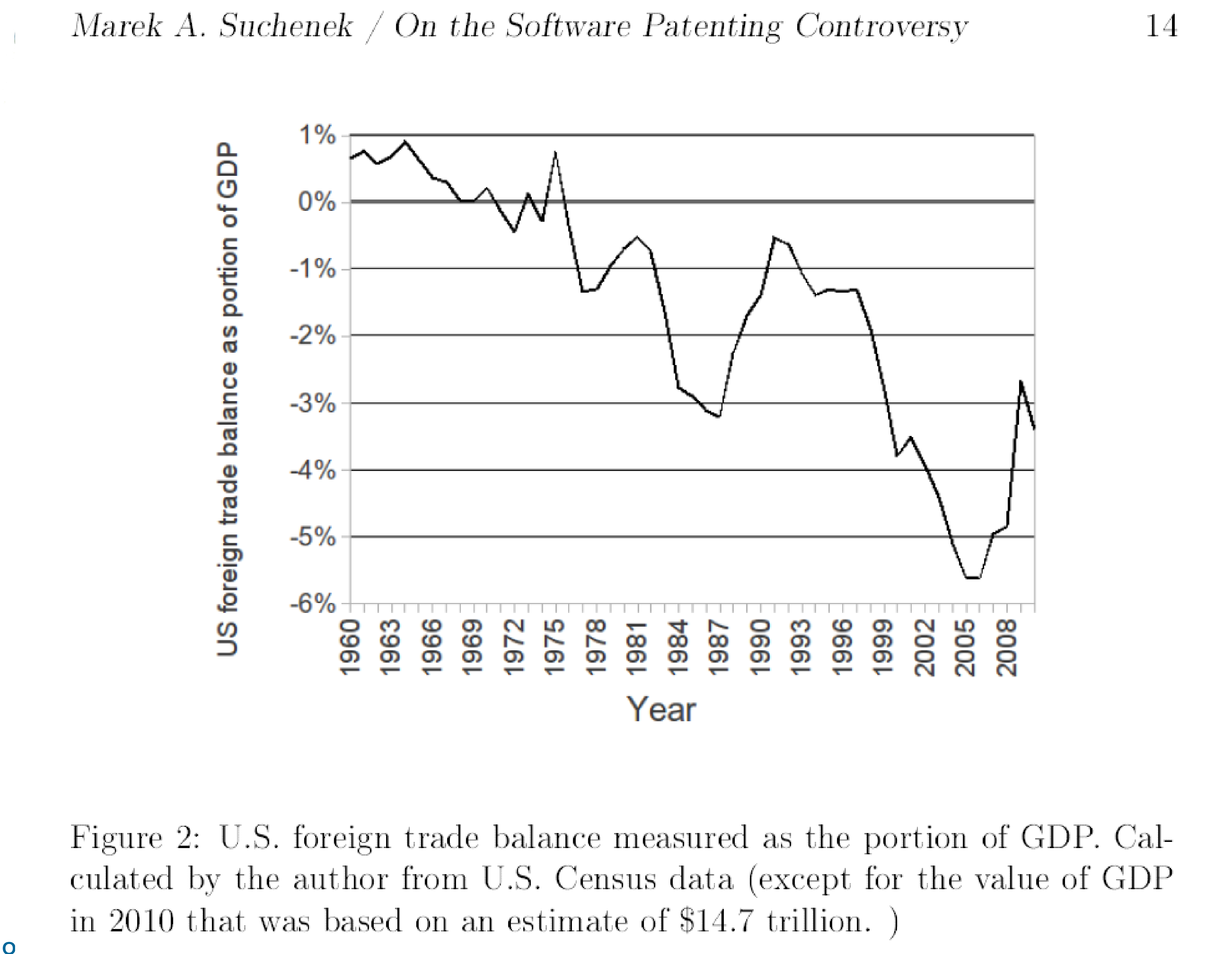
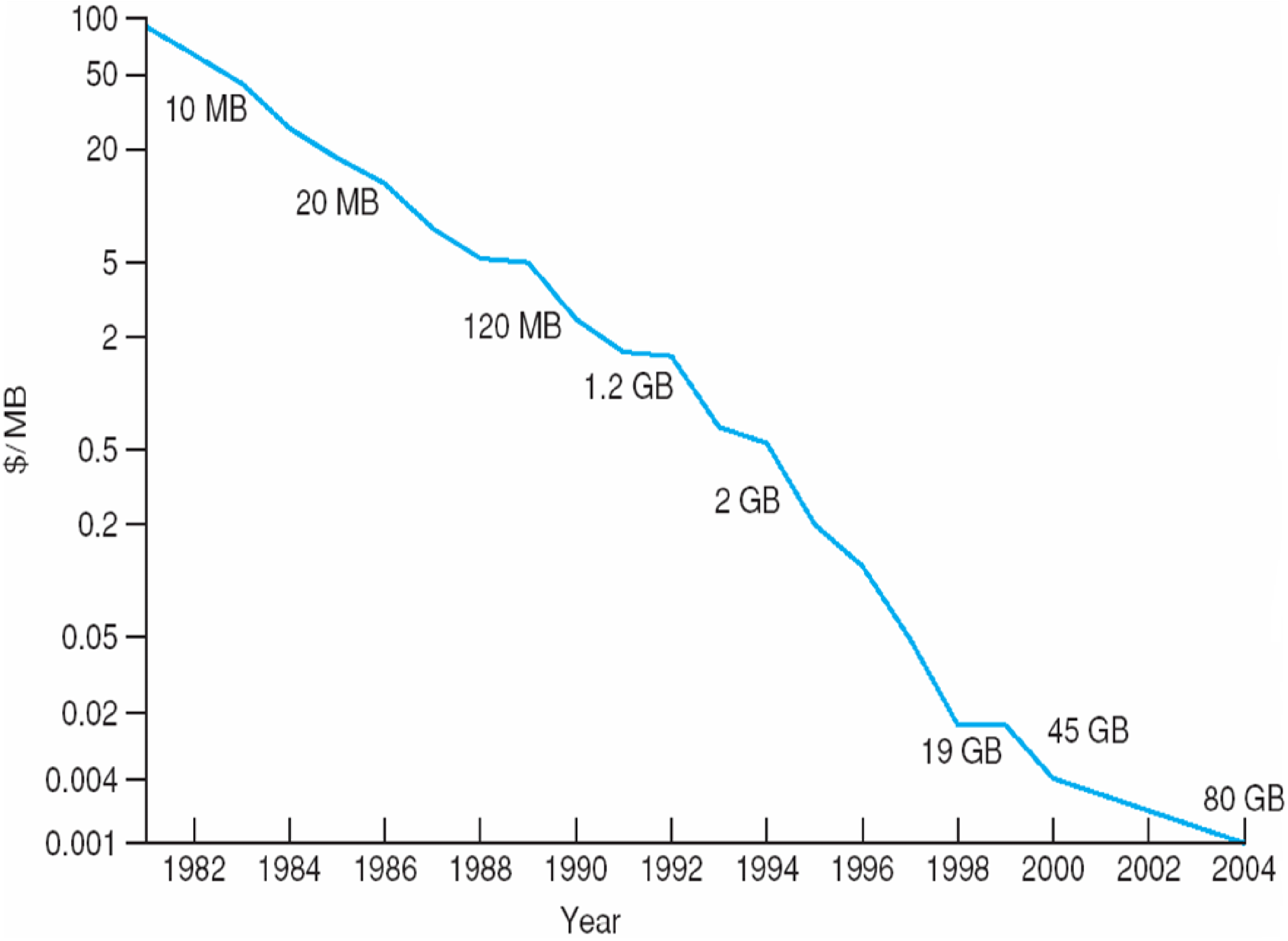
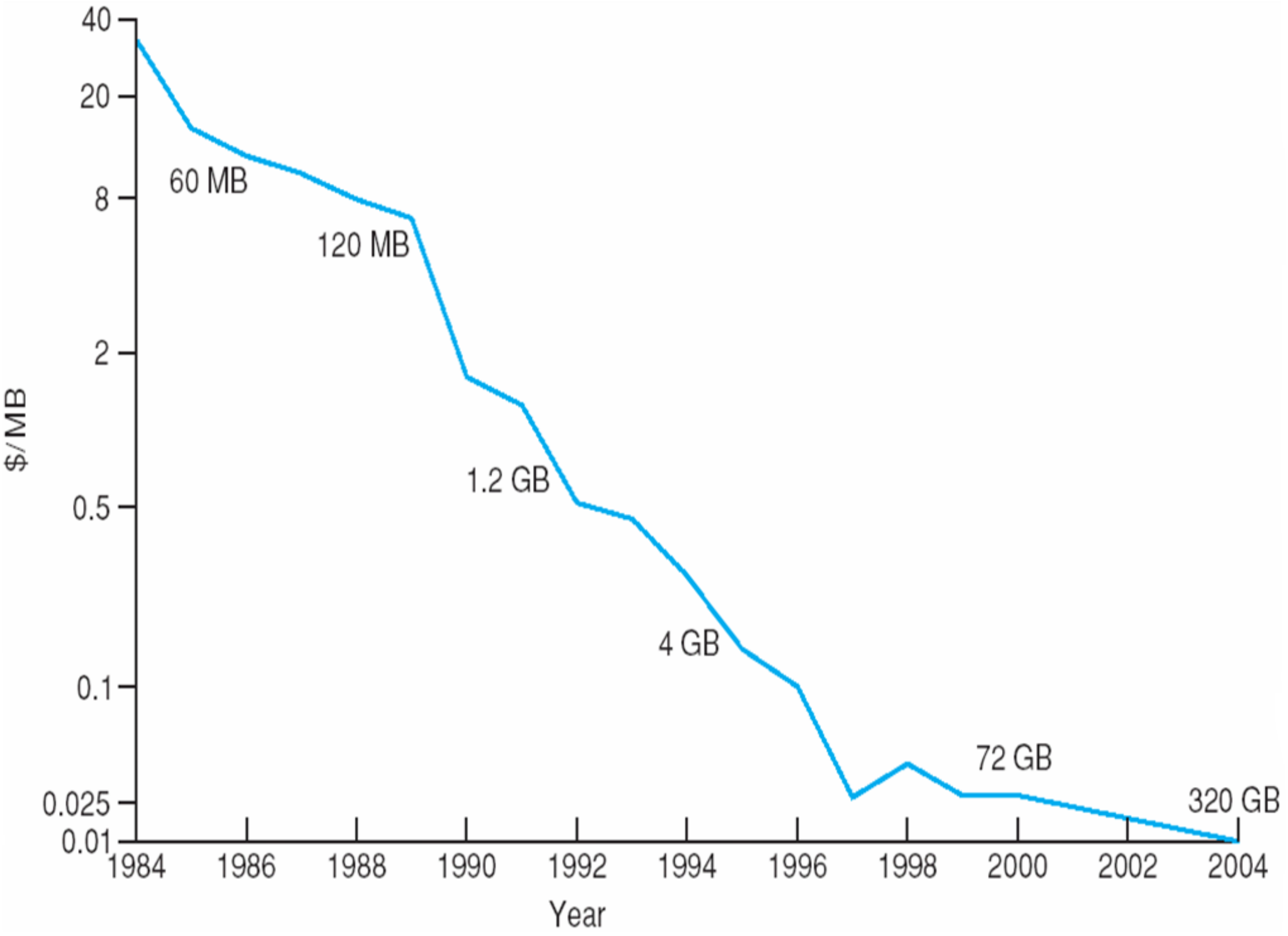


Figure 2: U.S. foreign trade balance measured as the portion of GDP. Calculated by the author from U.S. Census data (except for the value of GDP in 2010 that was based on an estimate of $14.7 trillion.)

Price per Megabyte of Magnetic Hard Disk, From 1981 to 2004



Price per Megabyte of a Tape Drive, From 1984-2000



End of Chapter 9