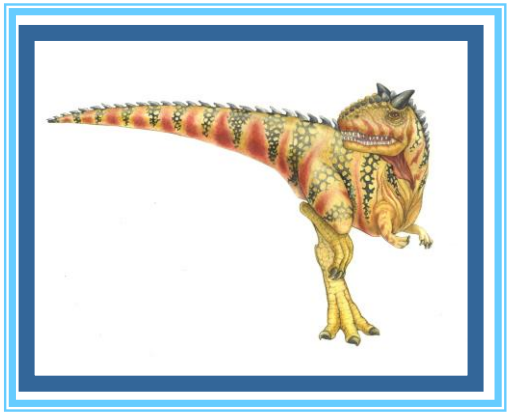
Chapter 12
Secondary-Storage Structure



Chapter 12: Secondary-Storage Structure

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

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

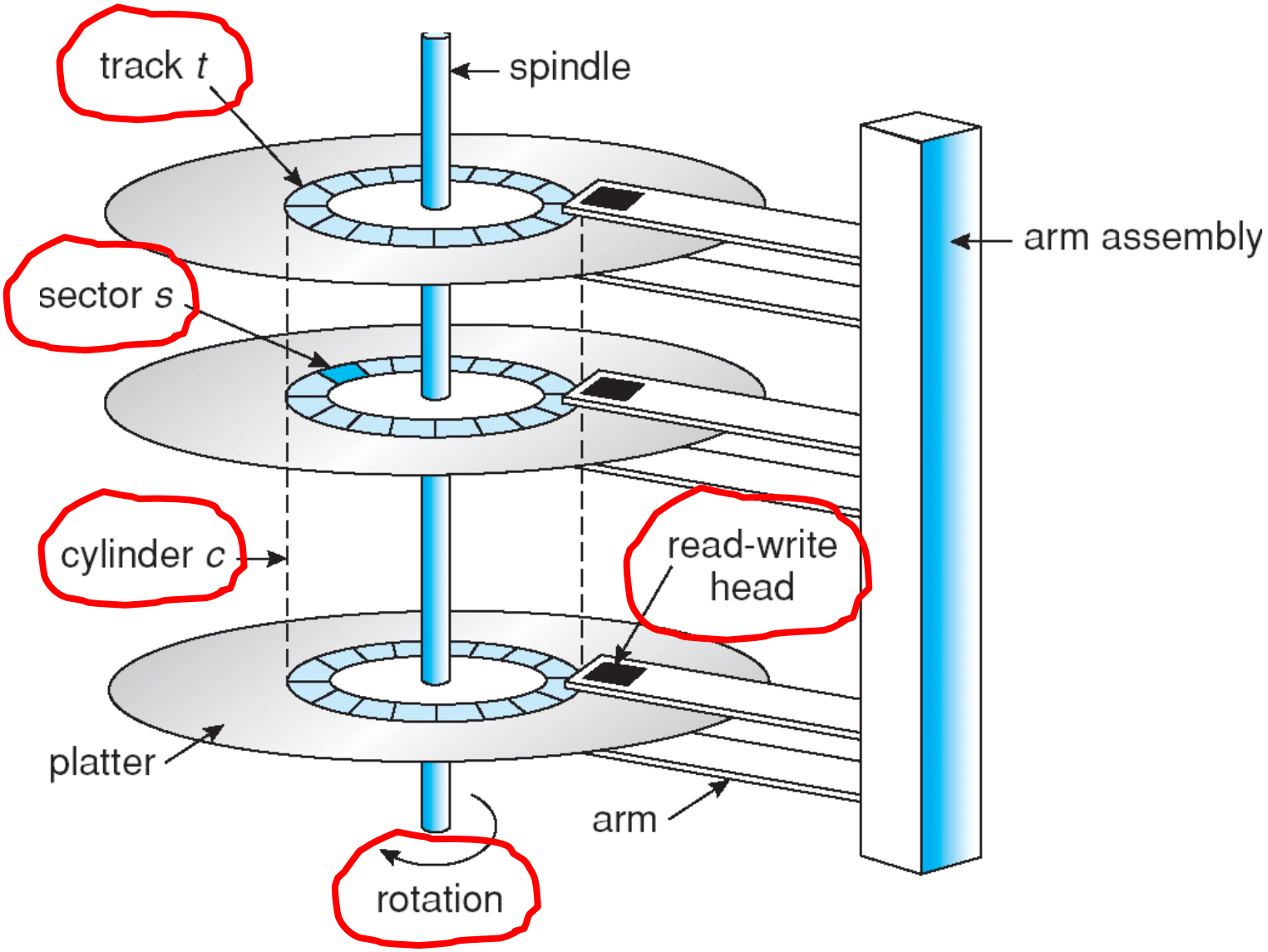
Discuss operating-system services provided for mass storage, including RAID and HSM (Hierarchical Storage Management)

12.1 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

Moving-head Disk Mechanism



Overview of Mass Storage Structure

- Disks can be removable
- Drive attached to computer via I/O bus
  - Busses vary, including EIDE, ATA, SATA, USB, Fibre Channel (FC), SCSI
  - Host controller in computer uses bus to talk to disk controller built into drive or storage array

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

12.2 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.

12.3 Disk Attachment

Host-attached storage accessed through I/O ports talking to I/O busses

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 (Fiber Channel) 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

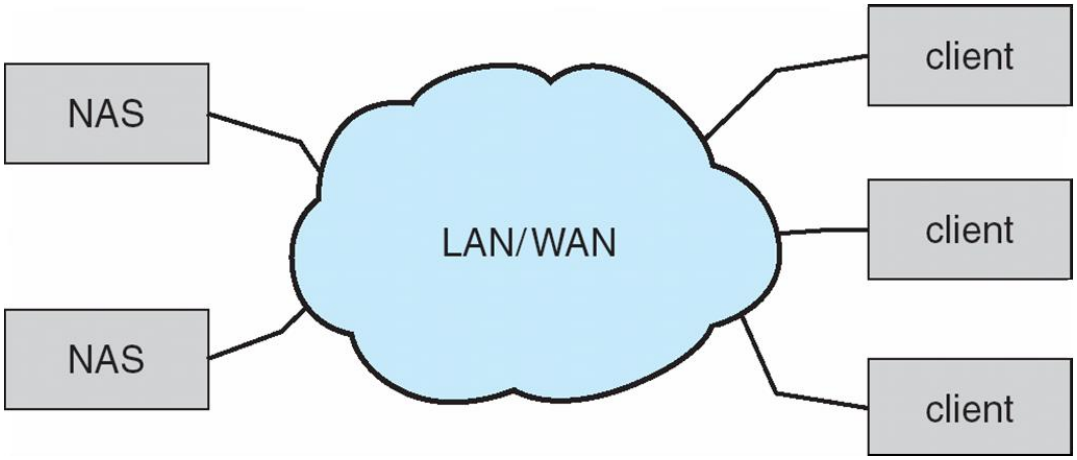
Network-Attached Storage (NAS)

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

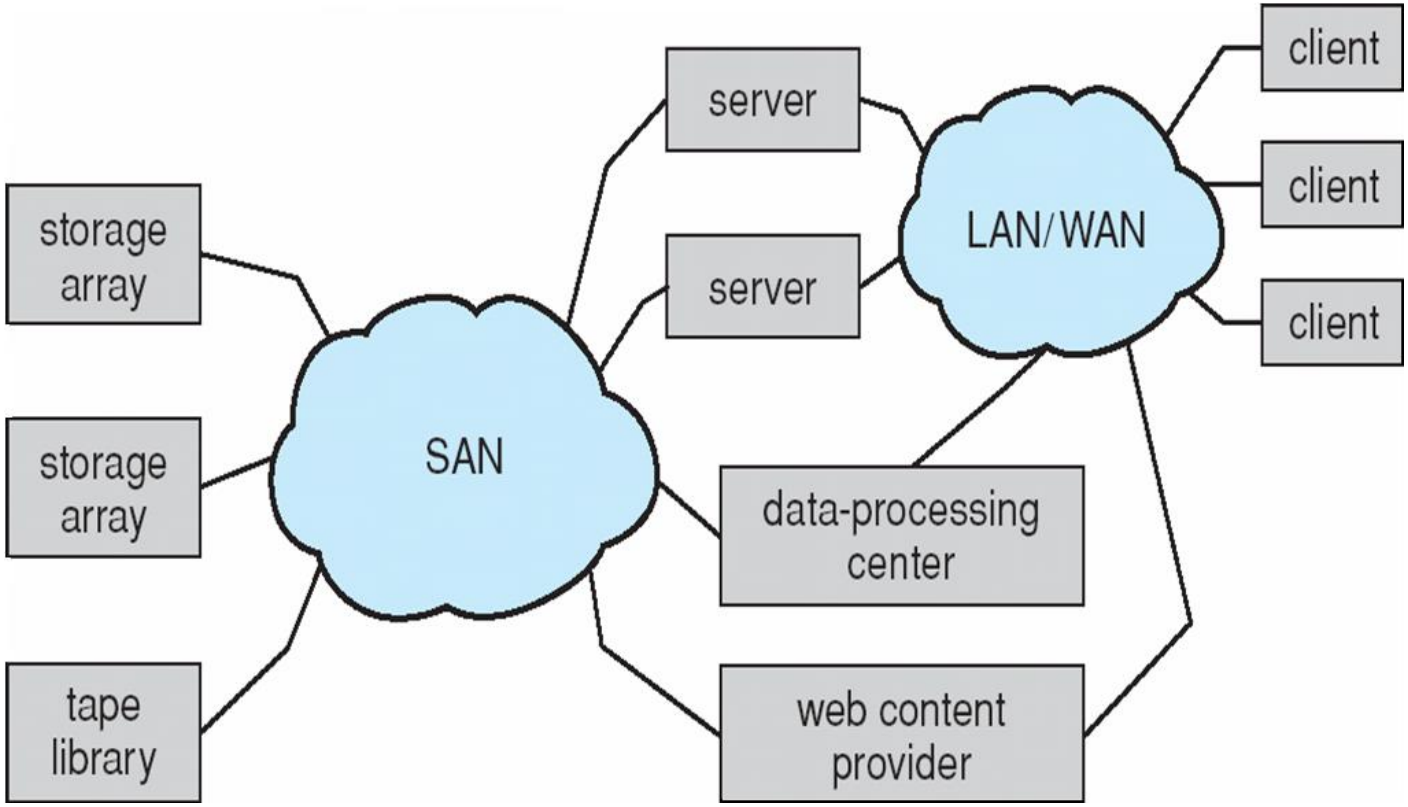
New iSCSI protocol uses IP network to carry the SCSI protocol



Storage Area Network (SAN)

Common in large storage environments (and becoming more common)

Multiple hosts attached to multiple storage arrays - flexible



12.4 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 arm 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

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.

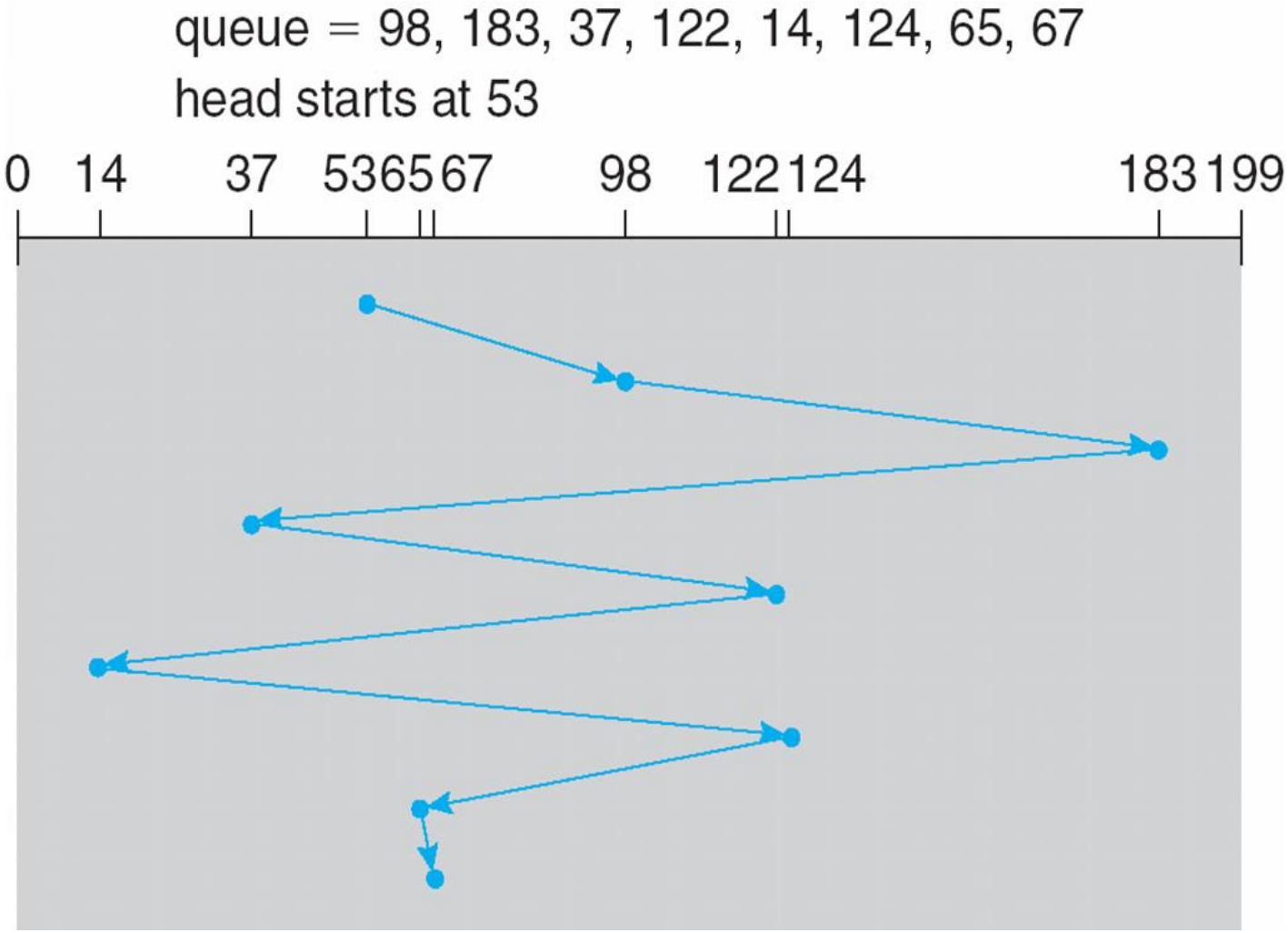
We illustrate them with a request queue (0-199 cylinders).

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

Head pointer 53

FCFS (First Come First Service)

Illustration shows total head movement of 640 cylinders.



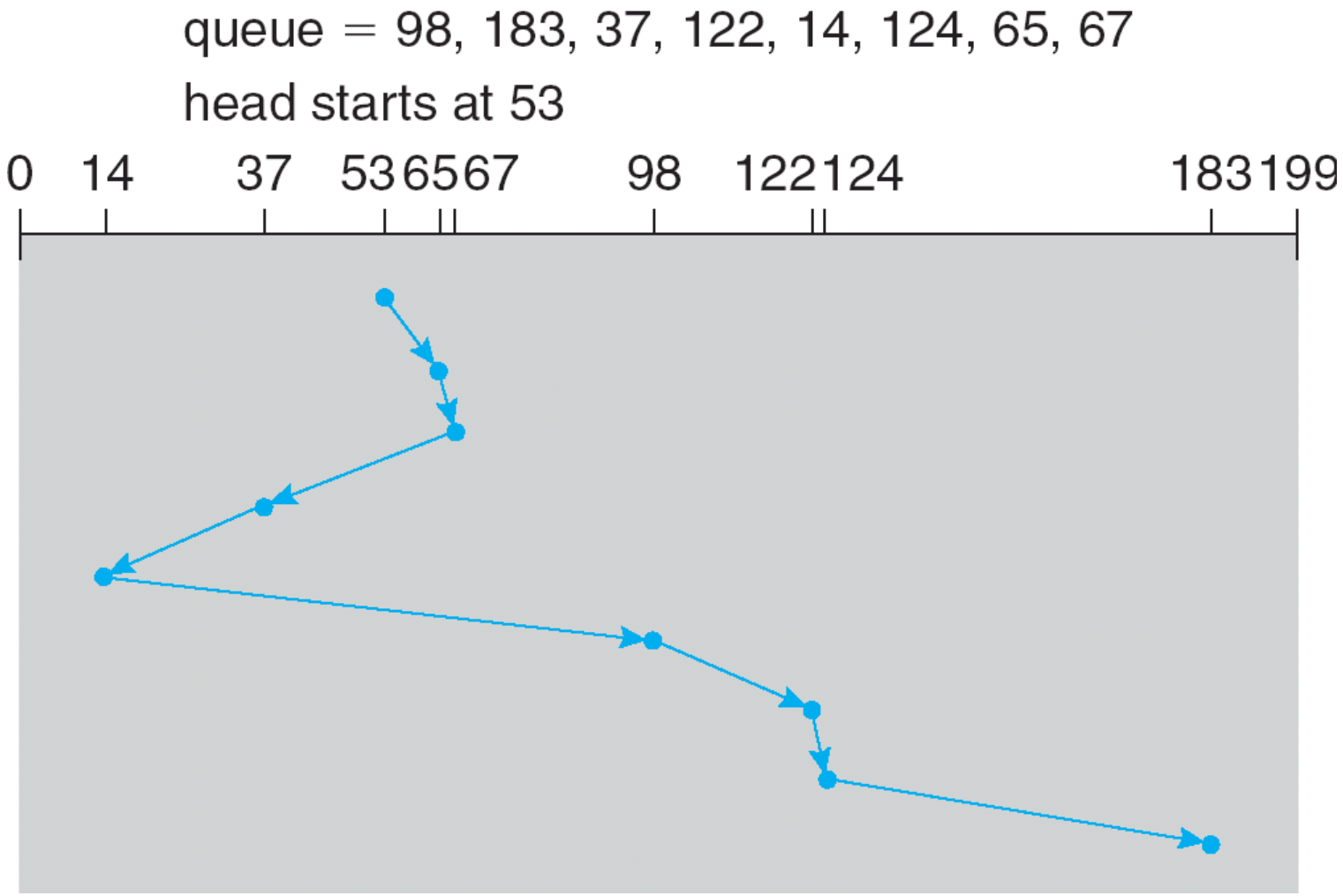
SSTF (Shortest Seek Time First)

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
0
14
37
53
65
67
98
122
124
183
199

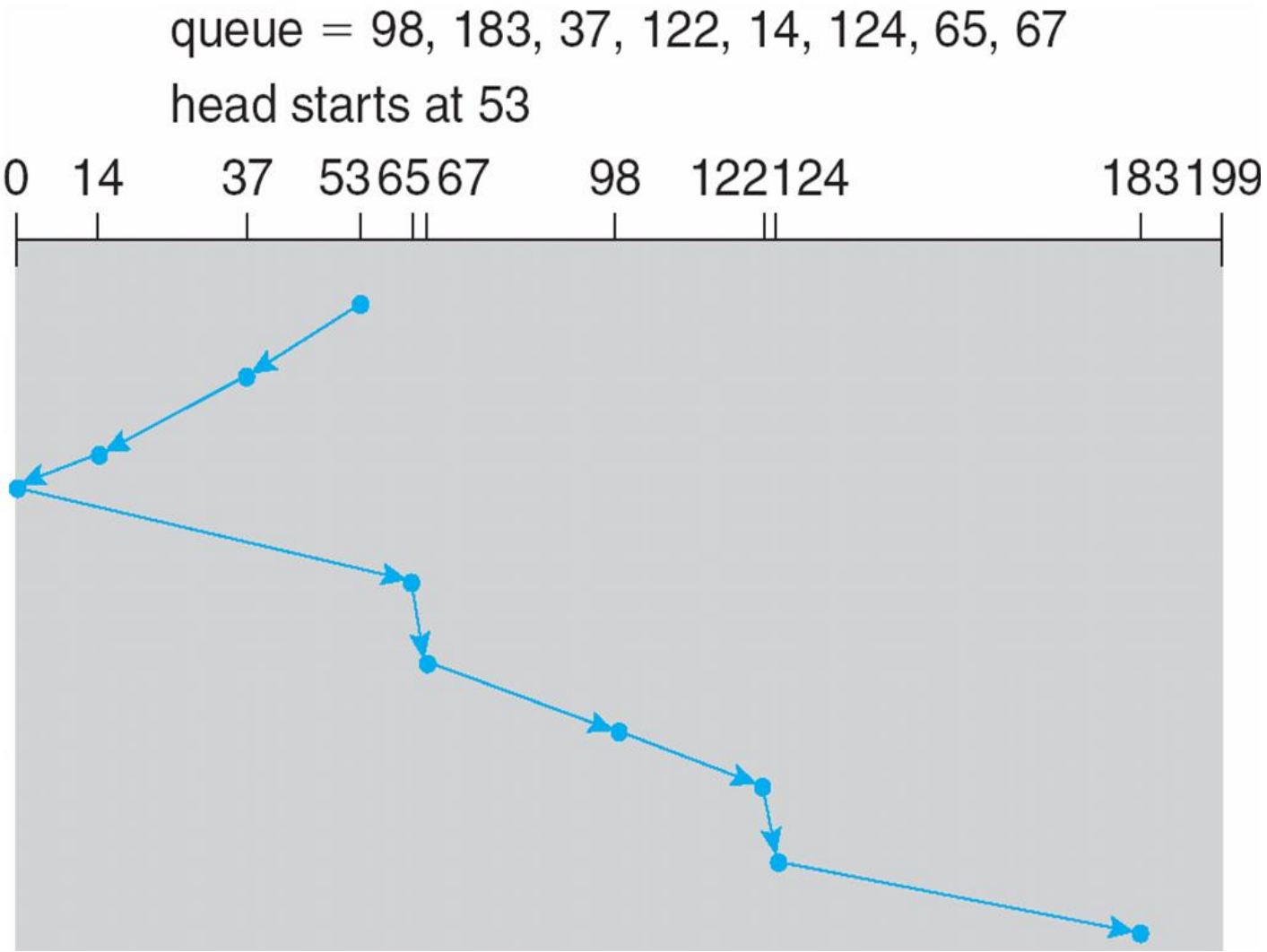
SCAN

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.)


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-SCAN

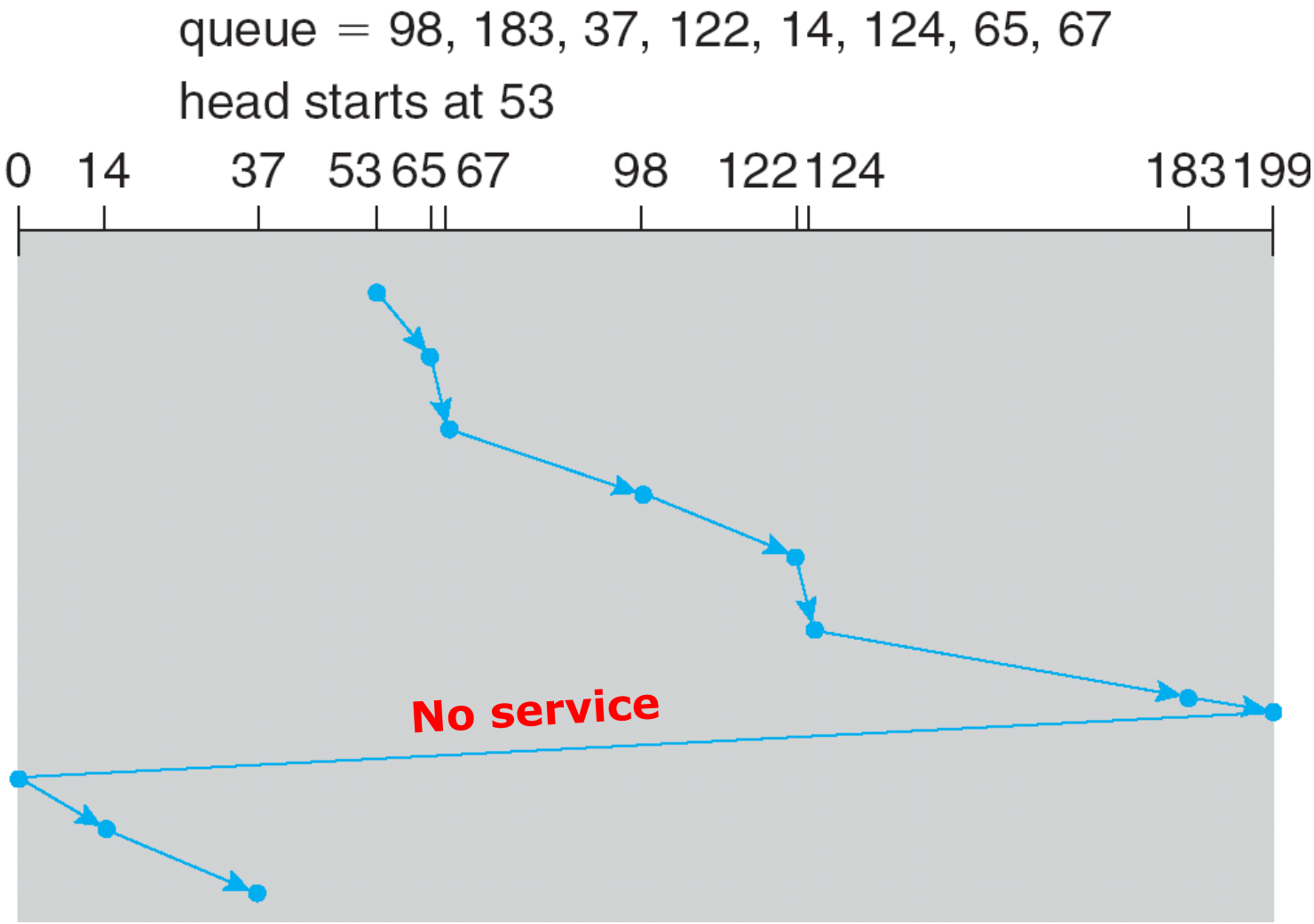
Provides a more uniform wait time than SCAN.

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
No service

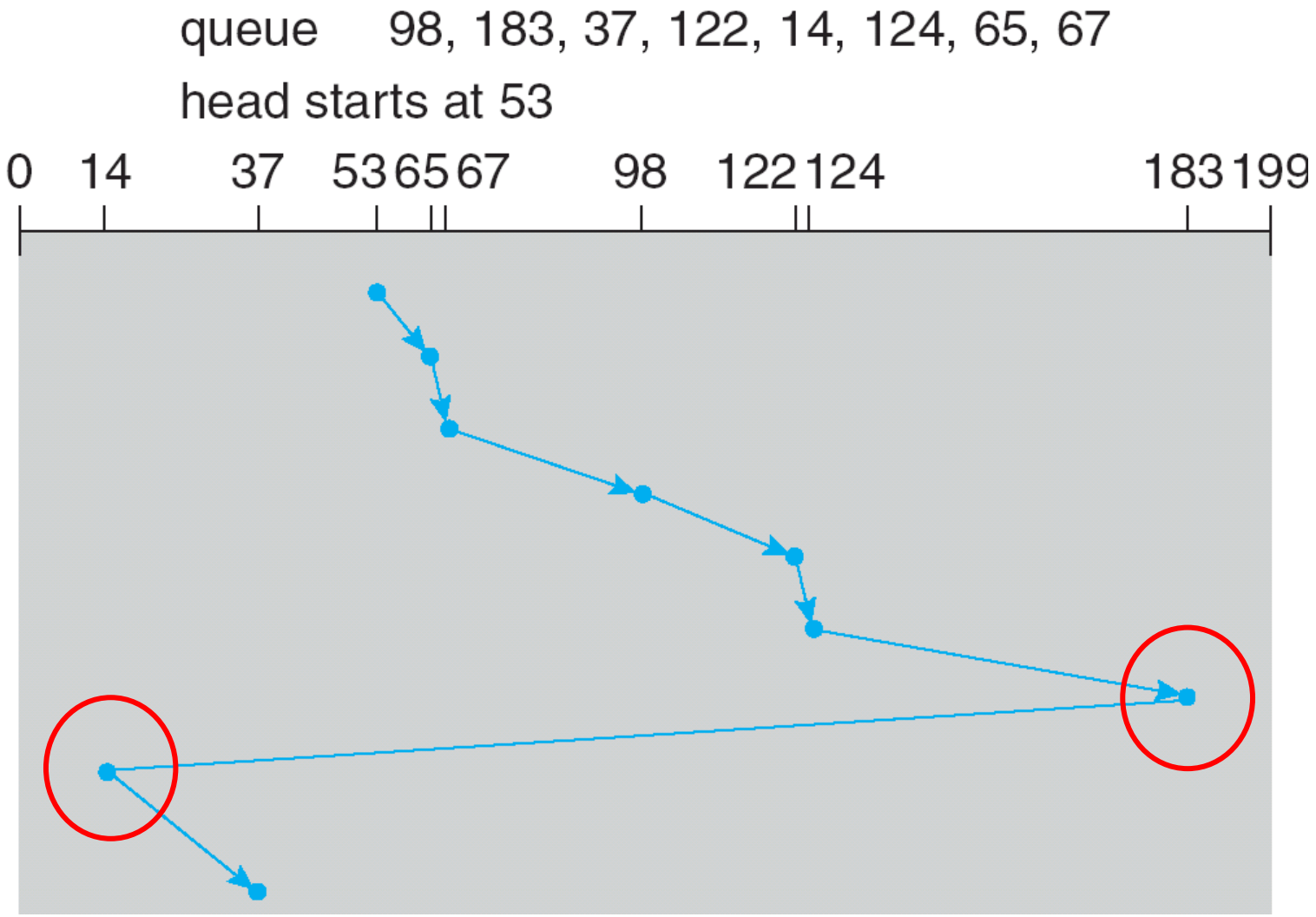
C-LOOK (or LOOK)

Versions of SACN and 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.

They look for a request before continuing to move in a given direction.

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

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

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.

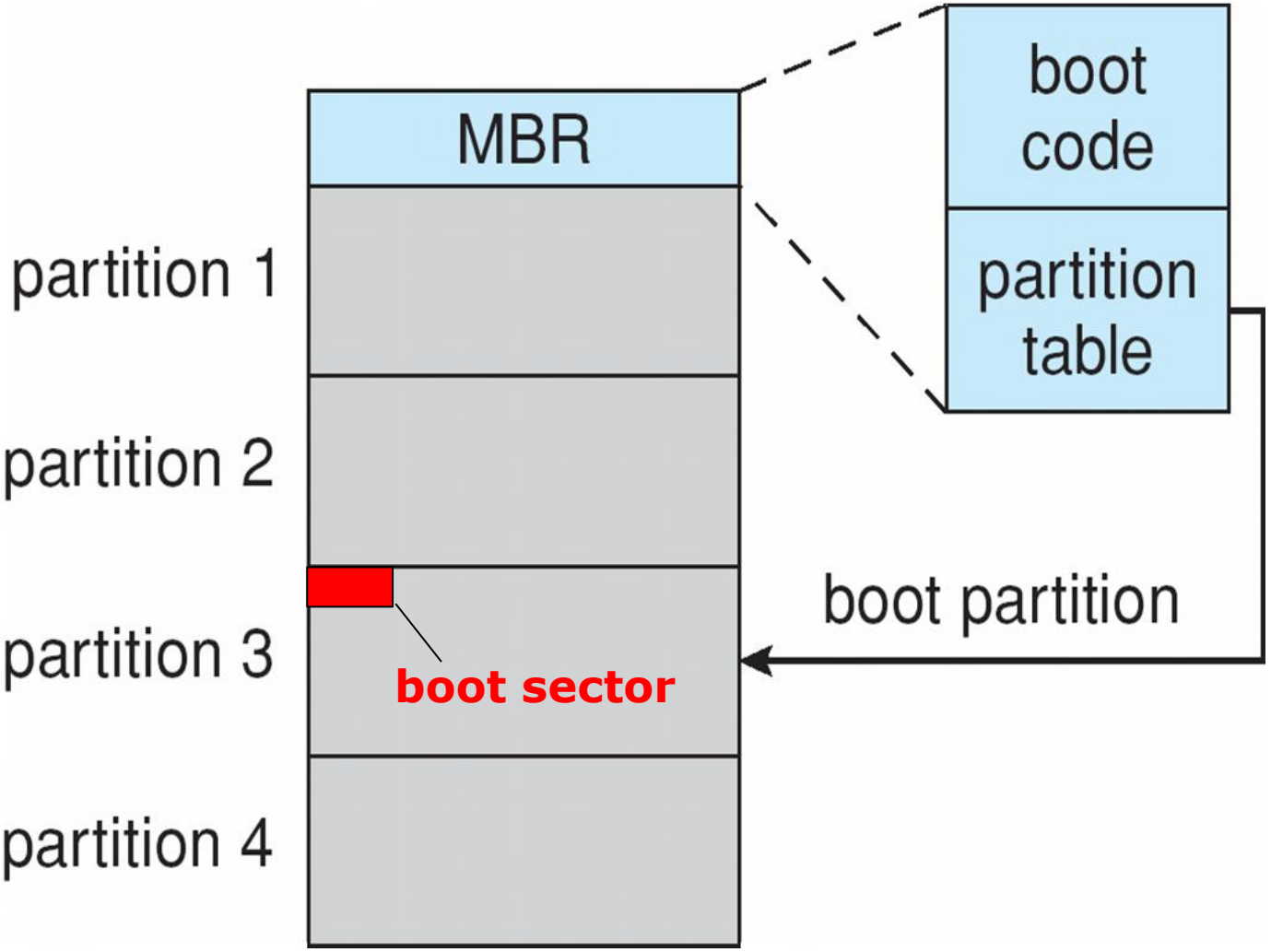
Booting from a Disk in Windows 2000

The Windows 2000 system places it boot code in the first sector on the hard disk (Master boot record, MBR).

Windows 2000 allows a hard disk to be divided into one or more partitions; one partition, identified as the boot partition, contains the OS ad device drivers.

Once the system identifies the boot partition, it reads the first sector from that partition (which is called the boot sector) and continues with the remainder of the boot process.

Booting from a Disk in Windows 2000



12.6 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.**

Swapping on Linux Systems

Linux allows one or more swap areas to be established.

A swap area may be either a swap file on a regular file system or a raw-swap-space partition.

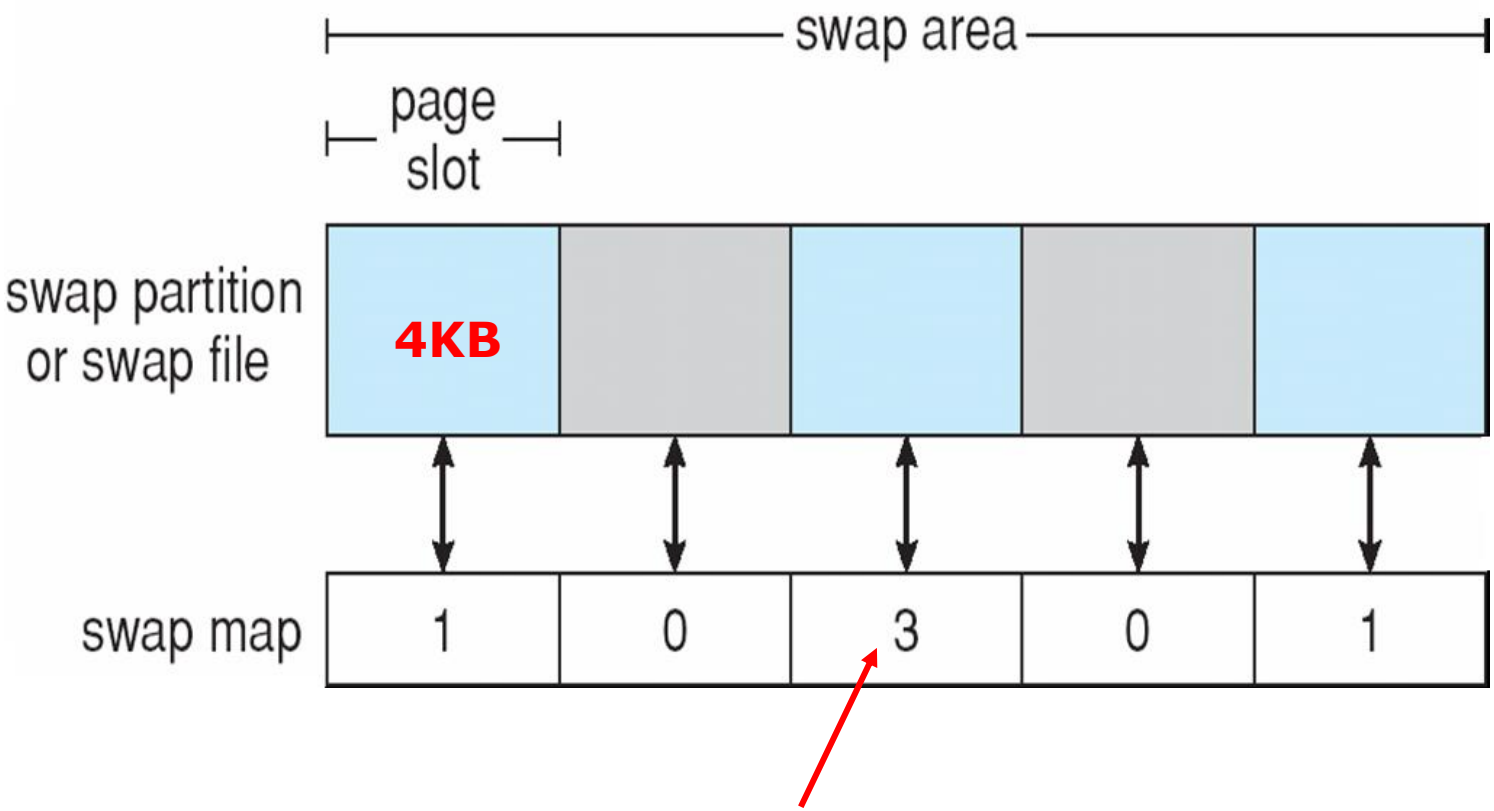
Each swap area consists of a series of 4KB page slots, which are used to hold swapped pages.

Each swap area is associated with a swap map.

Counter = 0, the corresponding page slot is available.

Counter >0, the page slot is occupied by a swapped page. The value of the counter indicates the number of mappings to the swapped page. Counter =3, the swapped paged is storing a region of memory shared by three processes.

Data Structures for Swapping on Linux Systems



The value of the counter indicates the number of mappings to the swapped page

12.7 RAID Structure

RAID (Redundant Arrays of Independent Disks) – multiple disk drives provides reliability via redundancy.

RAID is arranged into seven different levels.

With multiple disks, we can improve the transfer rate by striping data across the disks.

Data striping consists of splitting the bits of each byte across multiple disks – bit level striping

Block-level striping consists of splitting the blocks of each file across multiple disks.

RAID (cont)

Several improvements in disk-use techniques involve the use of multiple disks working cooperatively.

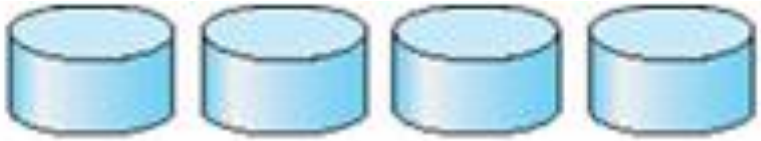
Disk striping uses a group of disks as one storage unit.

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.

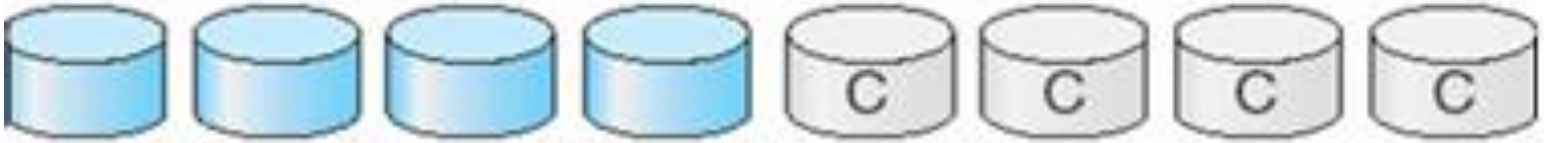
RAID (cont)

RAID level 0: RAID level 0 refers to disk arrays with striping at the level of blocks but without any redundancy.



(a) RAID 0: non-redundant striping.

RAID level 1: RAID level 1 refers to disk mirroring.



(b) RAID 1: mirrored disks.

RAID (cont)

RAID level 2: RAID level 2 also known as memory-style error-correcting-code (ECC) organization. The parity bits are used. The disks labeled P store the error-correction bits.

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

RAID (cont)

RAID level 3: RAID level 3 refers to bit-interleaved parity organization improves on level 2 by taking into account the fact that, unlike memory systems, disk controllers can detect whether a sector has been read correctly, so a single parity bit can be used for error correction and for detection.

(d) RAID 3: bit-interleaved parity.

RAID (cont)

RAID level 4: RAID level 0 refers to block-interleaved parity organization, uses block-level striping, as in RAID 0, and also keeps a parity block on a separate disk for corresponding blocks from N other disks.

(e) RAID 4: block-interleaved parity.

RAID (cont)

RAID level 5: RAID level 5 refers to block-interleaved distributed parity, differs from level 4 by spreading data and parity among all N+1 disks, rather than storing data in N disks and parity in one disk.

For each block, one of the disks stores the parity, and the others store data.

With an array of five disks, the parity for the *n*th block is stored in disk (*n* mod 5)+1; the *n*th blocks of the other four disks store actual data for the block.

A parity block cannot store parity for the blocks in the same disk.

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

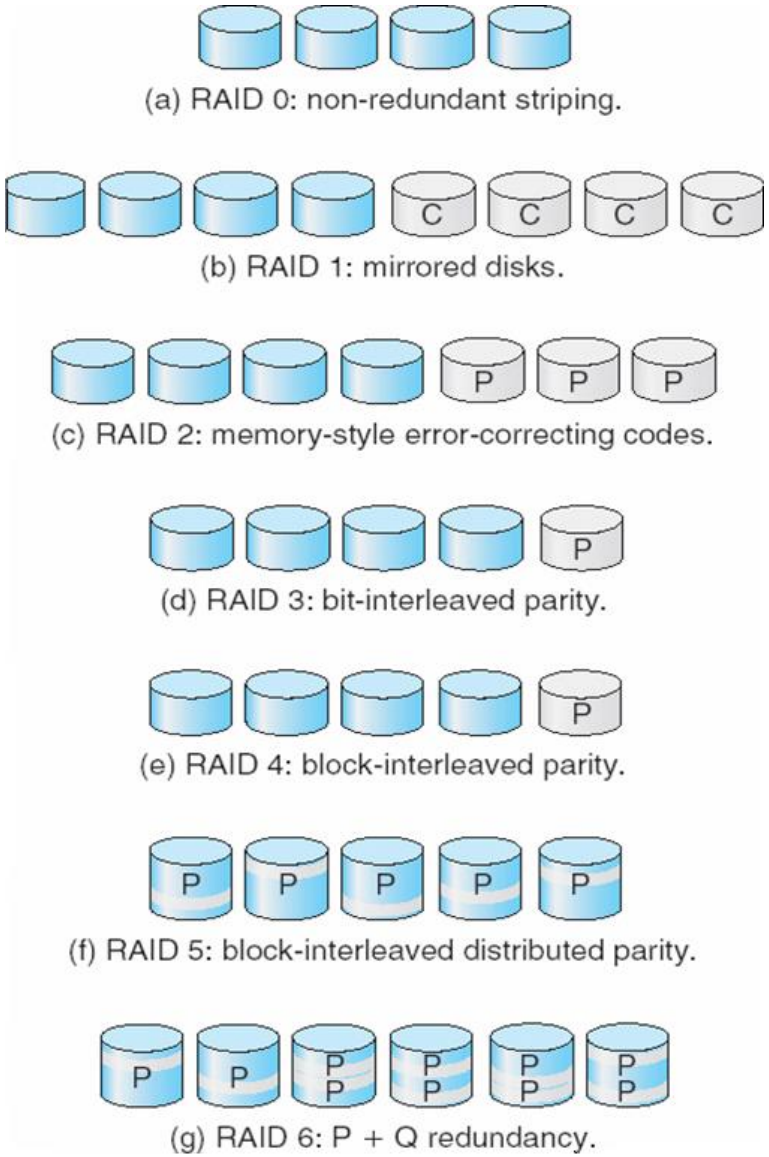
RAID (cont)

RAID level 6, P+Q redundancy scheme, is much like RAID 5 but stores extra redundant information to guard against multiple disk failures.

Instead of parity, error-correcting codes such as Reed-Solomon codes are used. 2 bits of redundant data are stored for every 4 bits of data – compared with 1 parity bit in level 5 – and the system can tolerate two disk failures.

(g) RAID 6: P + Q redundancy.

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)

RAID 0 provides the performance,

RAID 1 provides the reliability.

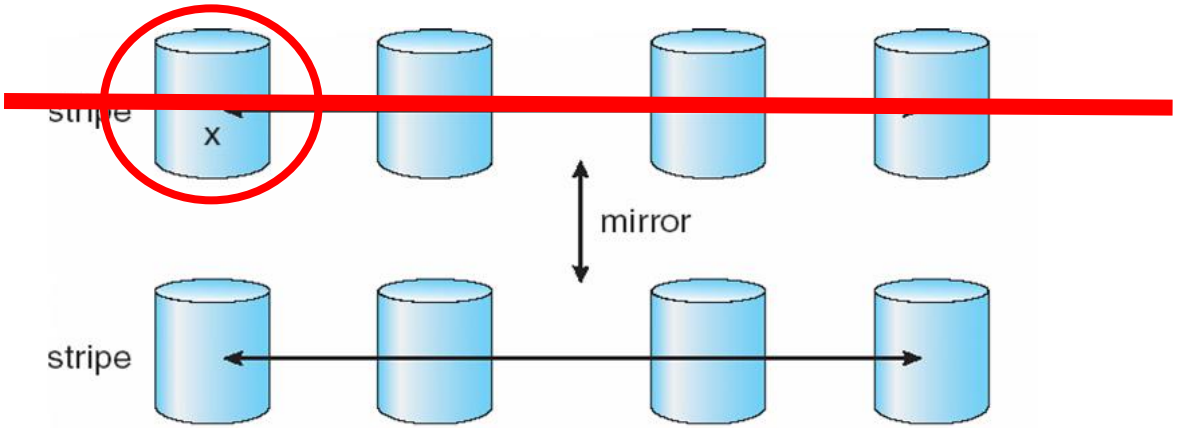
RAID 0 +1 : A set of disks are striped, and then the stripe is mirrored to another, equivalent strip

RAID 1+0: Disks are mirrored in pairs and then the resulting mirrored pairs are striped.

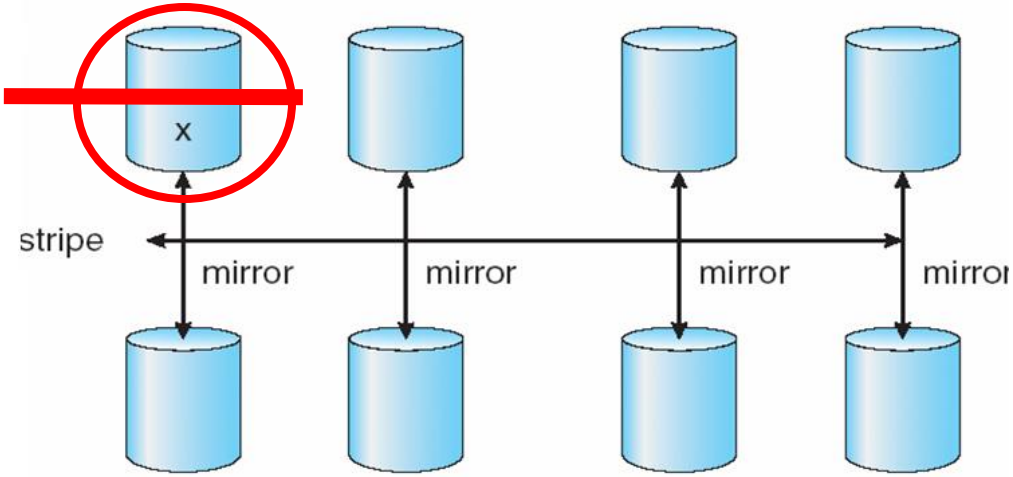
RAID 1+0 has some theoretical advantages over RAID 0+1.

For example, if a single disk fails in RAID 0+1, an entire strip is inaccessible, leaving only the other strip available. With a failure in RAID 1+0, a single disk is unavailable, but the disk that mirrors it is still available, as are all the rest of the disks.

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



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



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

12.8 Stable-Storage Implementation

Write-ahead log scheme requires stable storage.

To implement stable storage:

Replicate information on more than one nonvolatile storage media with independent failure modes.

Update information in a controlled manner to ensure that we can recover the stable data after any failure during data transfer or recovery.

12.9 Tertiary Storage Structure

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.

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 as CD-ROM and DVD, come 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.

Operating System Support

Major OS jobs are to manage physical devices and to present a virtual machine abstraction to applications

For hard disks, the OS provides two abstraction:

Raw device – an array of data blocks.

File system – the OS queues and schedules the interleaved requests from several applications.

Application Interface

Most OSs handle removable disks almost exactly like fixed disks — a new cartridge is formatted and an empty file system is generated on the disk.

Tapes are presented as a raw storage medium, i.e., and application does not open a file on the tape, it opens the whole tape drive as a raw device.

Usually the tape drive is reserved for the exclusive use of that application.

Since the OS does not provide file system services, the application must decide how to use the array of blocks.

Since every application makes up its own rules for how to organize a tape, a tape full of data can generally only be used by the program that created it.

Tape Drives

The basic operations for a tape drive differ from those of a disk drive.

locate positions the tape to a specific logical block, not an entire track (corresponds to seek).

The read position operation returns the logical block number where the tape head is.

The space operation enables relative motion.

Tape drives are "append-only" devices; updating a block in the middle of the tape also effectively erases everything beyond that block.

An EOT mark is placed after a block that is written.

File Naming

The issue of naming files on removable media is especially difficult when we want to write data on a removable cartridge on one computer, and then use the cartridge in another computer.

Contemporary OSs generally leave the name space problem unsolved for removable media, and depend on applications and users to figure out how to access and interpret the data.

Some kinds of removable media (e.g., CDs) are so well standardized that all computers use them the same way.

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

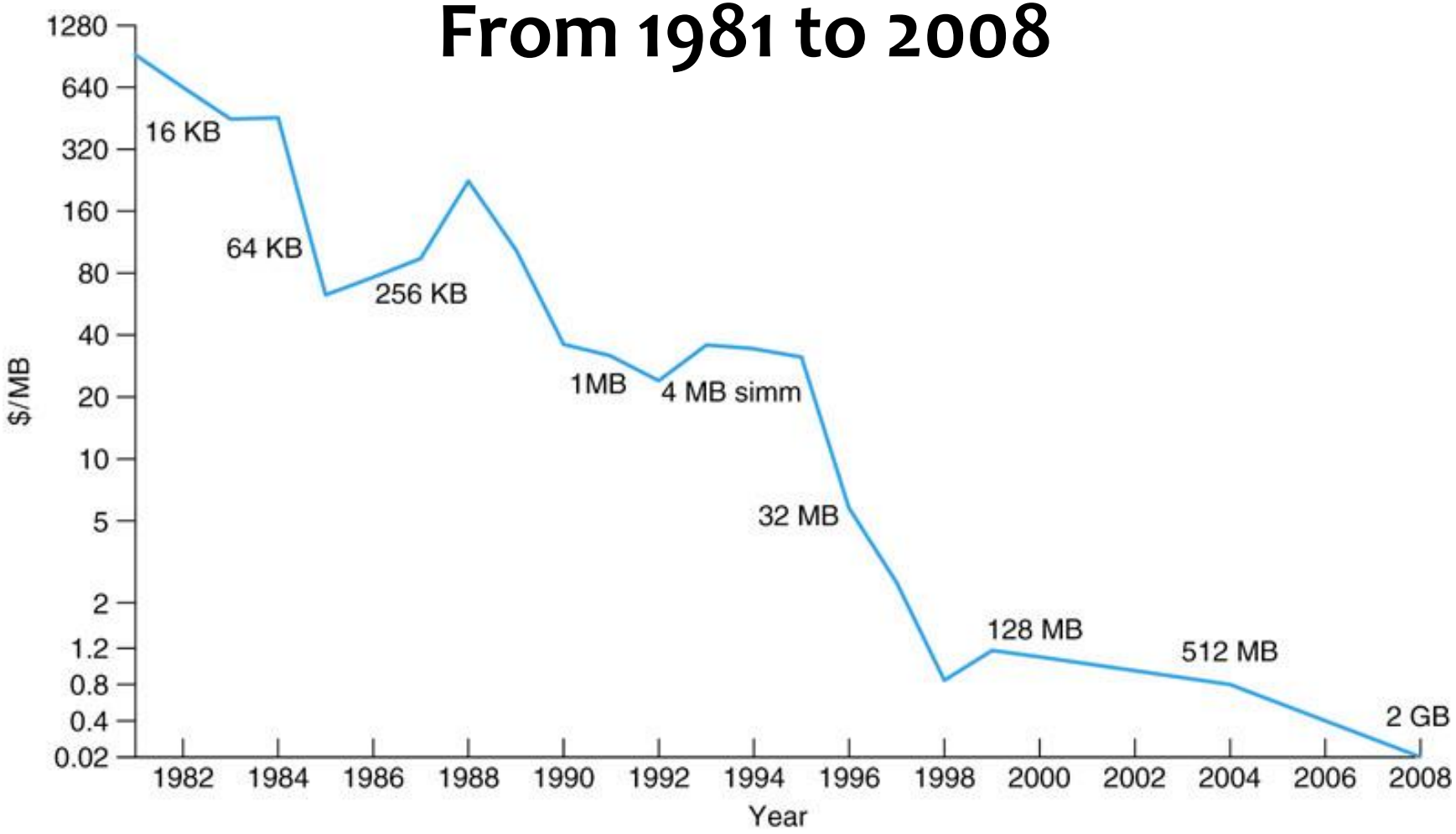
The cost per megabyte of hard disk storage is competitive with magnetic tape if only one tape is used per drive.

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.

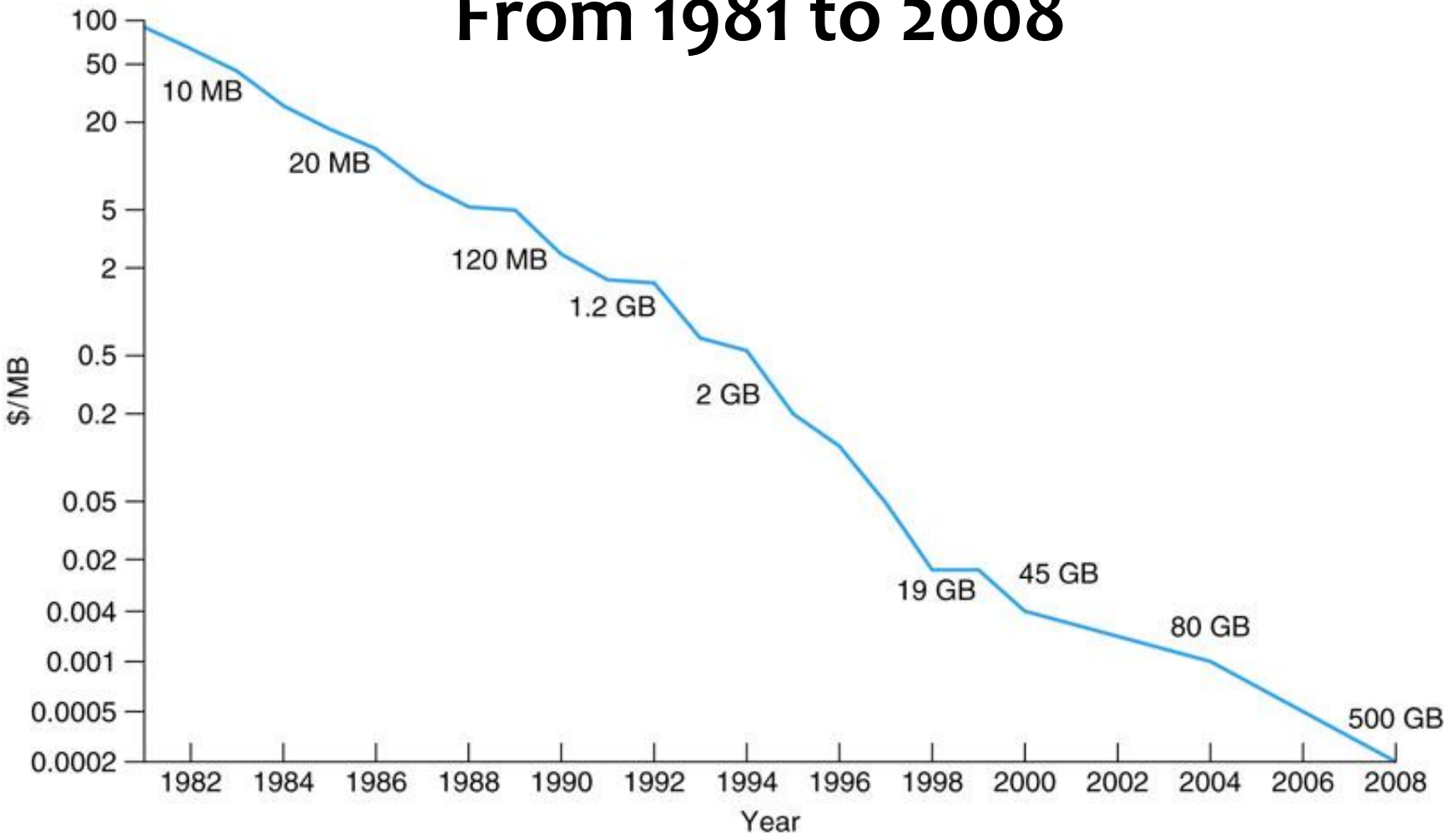
Price per Megabyte of DRAM

From 1981 to 2008



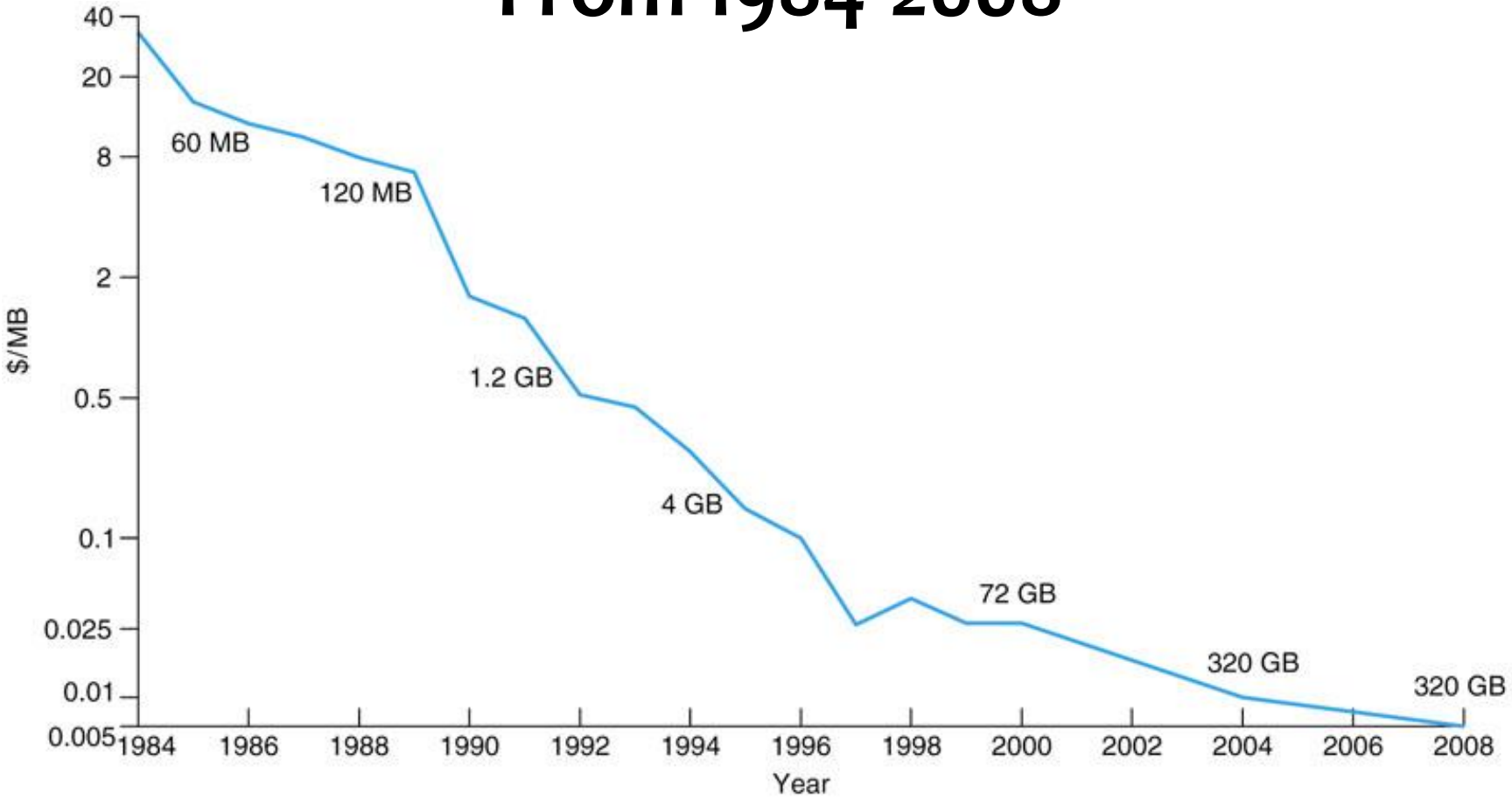
Price per Megabyte of Magnetic Hard Disk

From 1981 to 2008



Price per Megabyte of a Tape Drive

From 1984-2008



End of Chapter 12