Chapter 12
Mass-Storage Systems

Contents

- 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

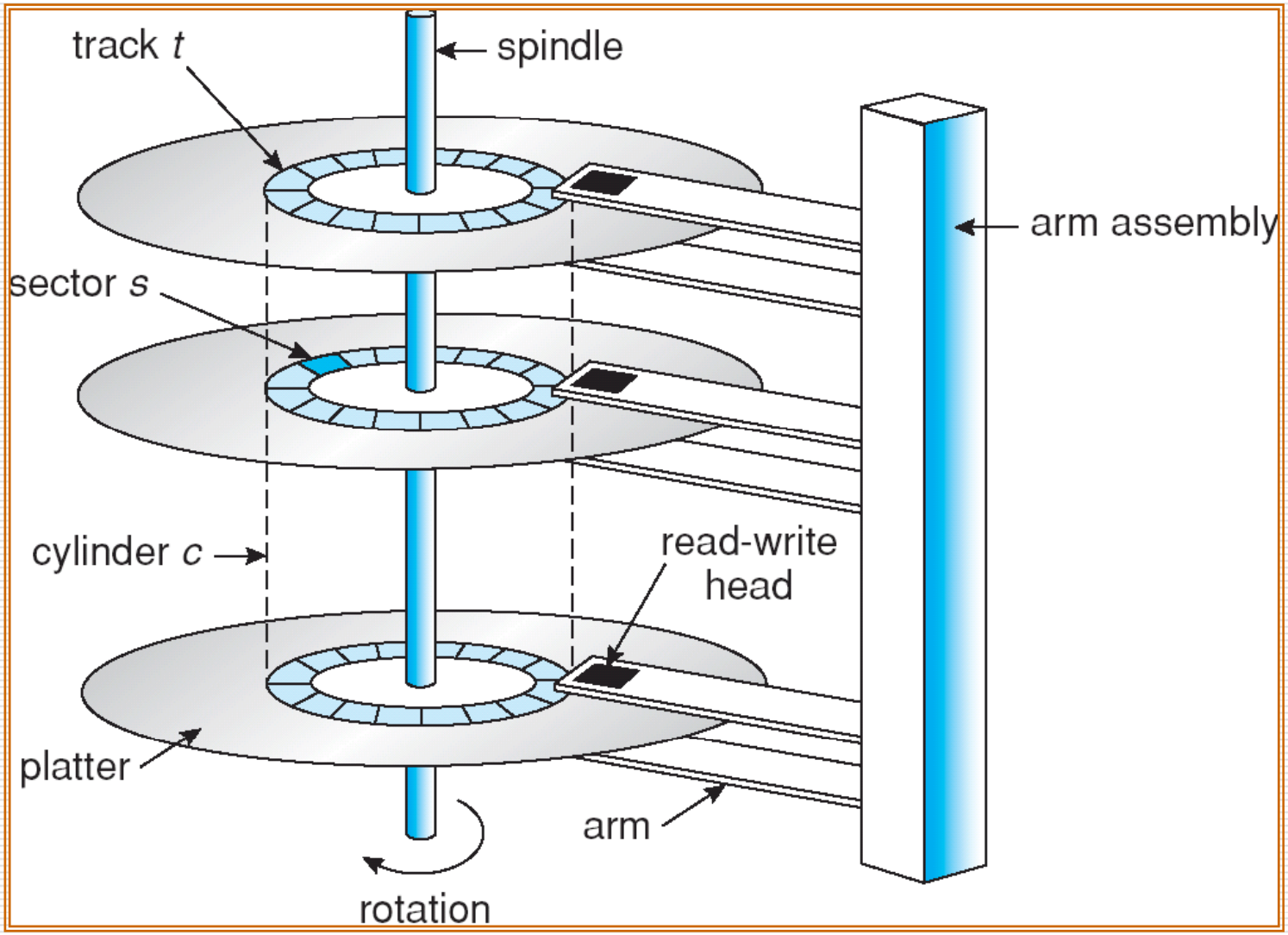
Objectives

- 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, Fiber Channel, SCSI**
  - **Host controller** in computer uses bus to talk to **disk controller** built into drive or storage array

Moving-head Disk Machanism



Overview of Mass Storage Structure (Cont.)

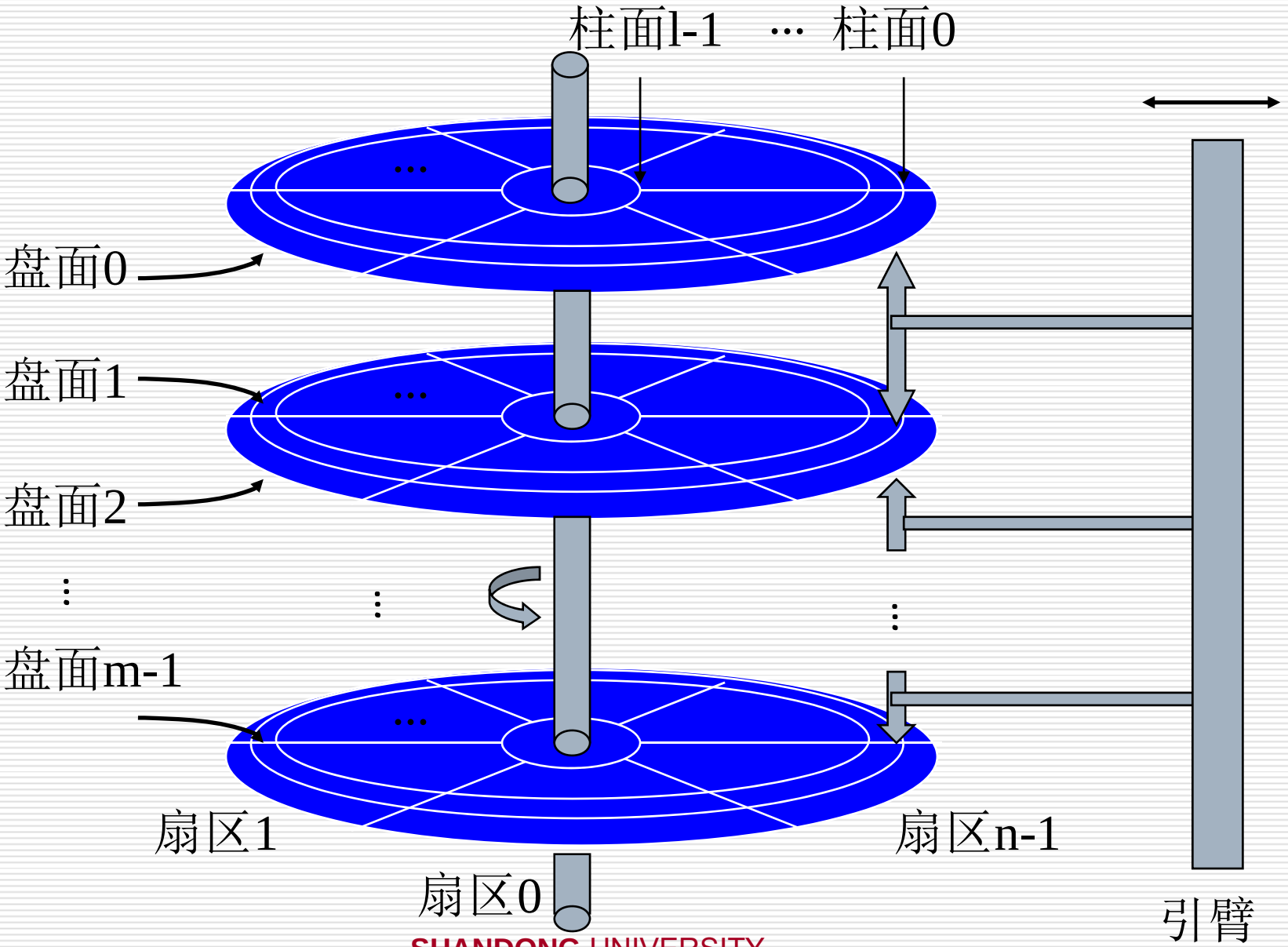
- Magnetic tape
  - Was early secondary-storage medium
  - Relatively permanent and holds large quantities of data
  - Access time slow
  - Random access ~1000 times slower than disk
  - Mainly used for backup, storage of infrequently-used data, transfer medium between systems
  - Kept in spool and wound 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.

磁盘组的物理特性



柱面号i
盘面号j
扇区号k
（三维地址） ⟺ 块号b（一维地址）

编址方法：使相邻块物理上最近

例子：柱面数l=2; 盘面数m=3; 扇区数n=3

柱面号：	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
盘面号：	0	0	0	1	1	1	2	2	2	0	0	0	1	1	1	2	2	2
扇区号：	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
块　号：	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

Disk Structure

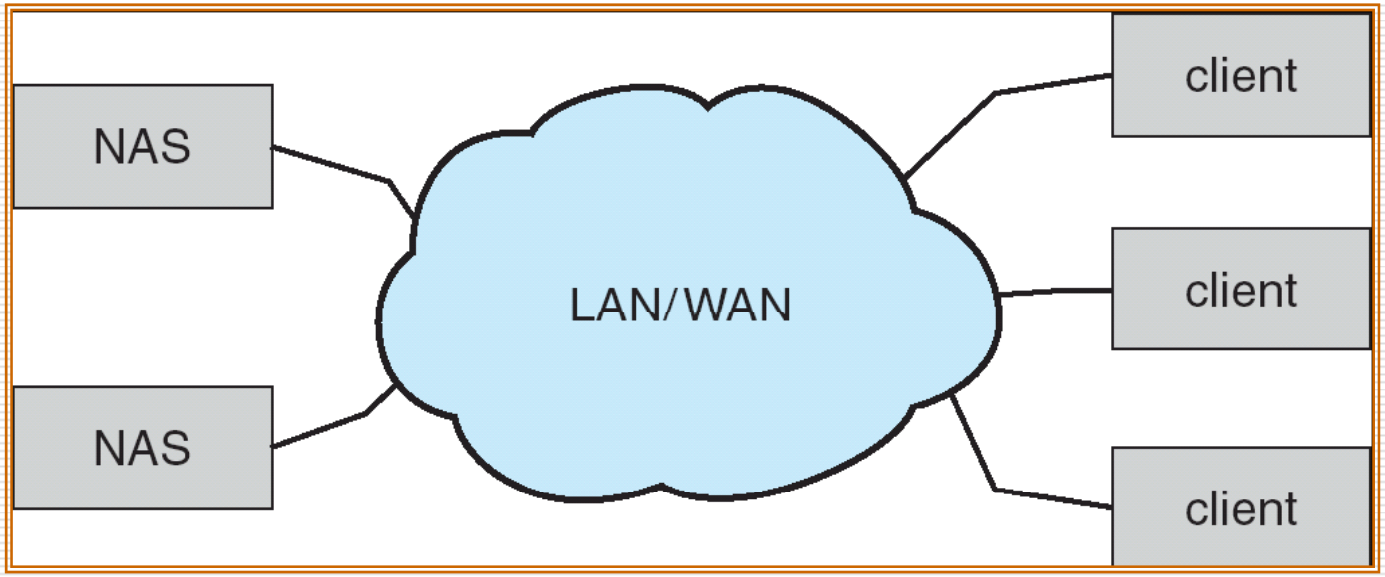
- It is not easy to perform such translation:
  - Most disks have some defective sectors.
  - The number of sectors per track is not a constant on some disks.
    - constant linear velocity, CLV
    - constant angular velocity, CAV

12.3 Disk Attachment

- Computers access disk storage in two ways
  - Via I/O ports—host-attached storage
  - Via a remote host in a distributed file system—network-attached storage.
- 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 is high-speed serial architecture
  - Can be switched fabric with 24-bit address space – the basis of **storage area networks** (**SAN**s) in which many hosts attach to many storage units
  - Can be **arbitrated loop** (**FC-AL**) of 126 devices

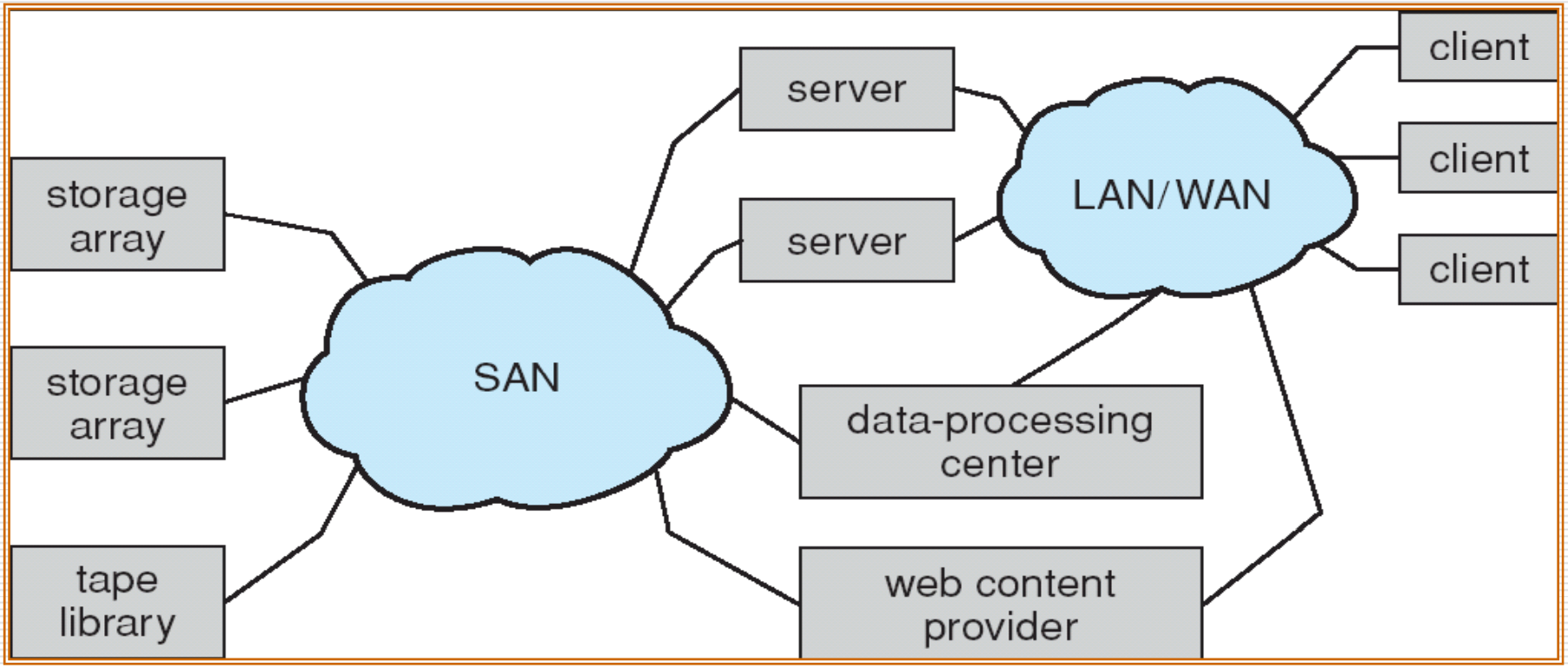
Network-Attached Storage

- 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

- 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 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
- 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).

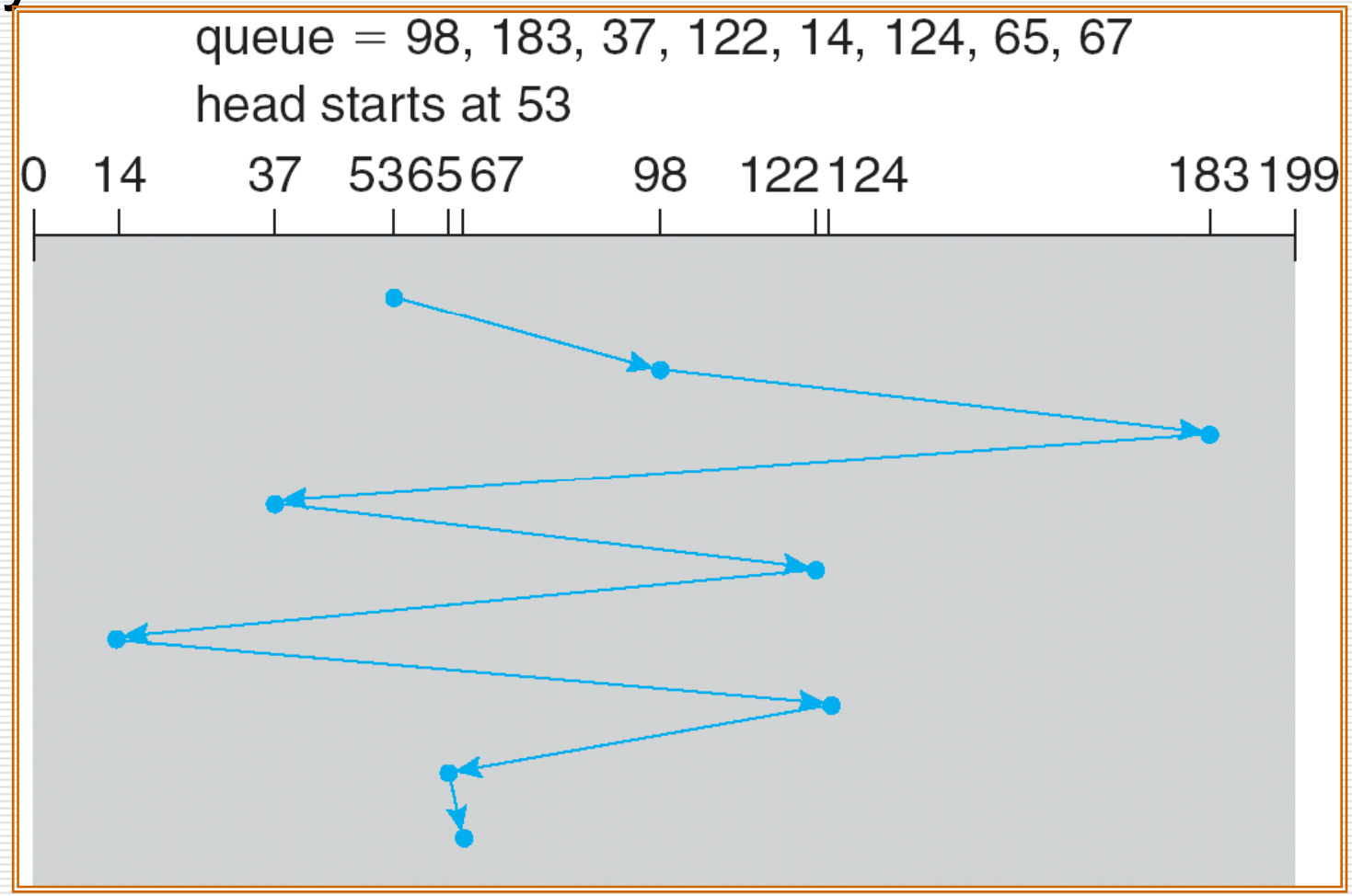
  98, 183, 37, 122, 14, 124, 65, 67

  Head pointer 53

FCFS

磁头总移动距离为640

- Illustration shows total head movement of 640 cylinders.

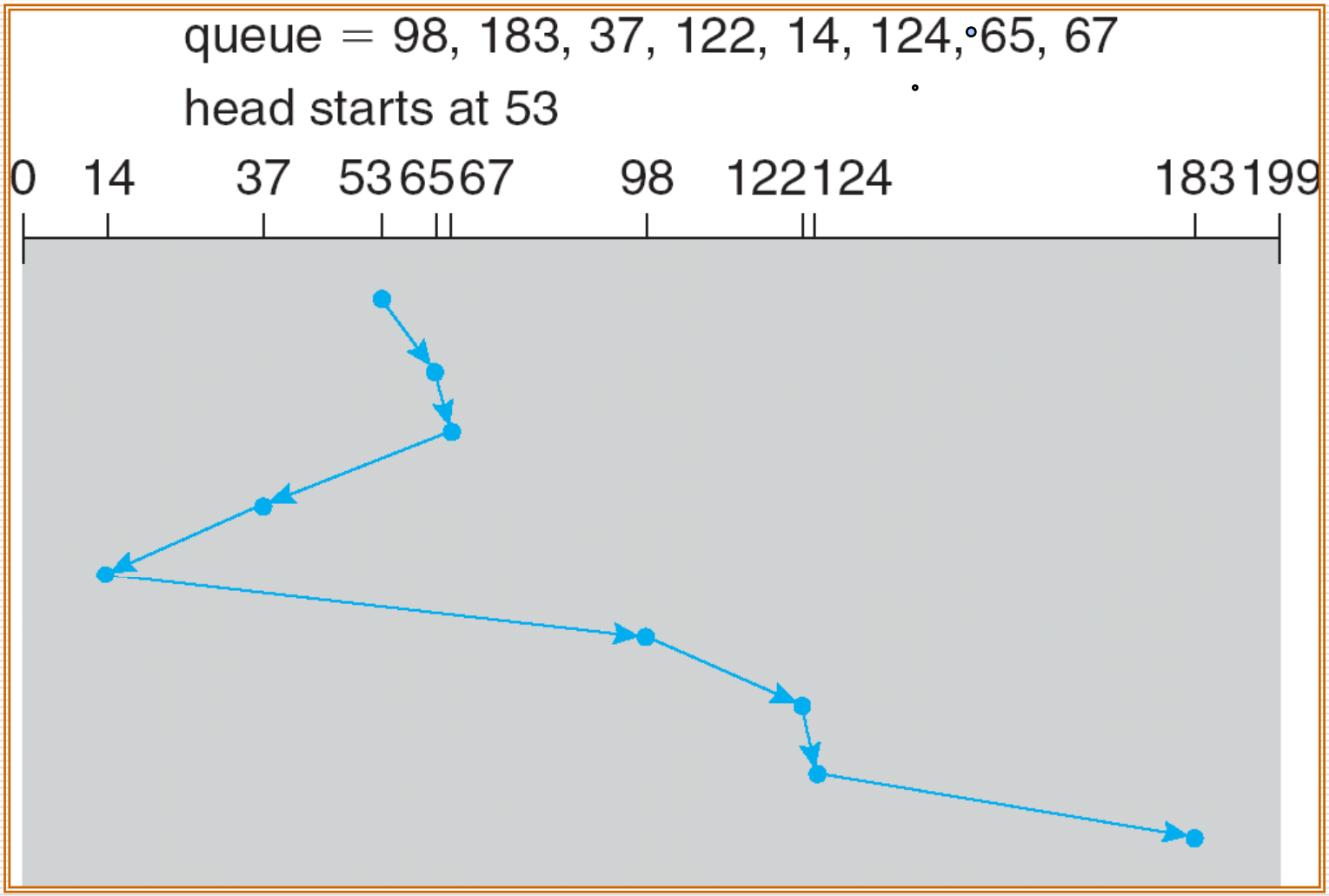


SSTF—shortest-seek-time-first

- 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.)

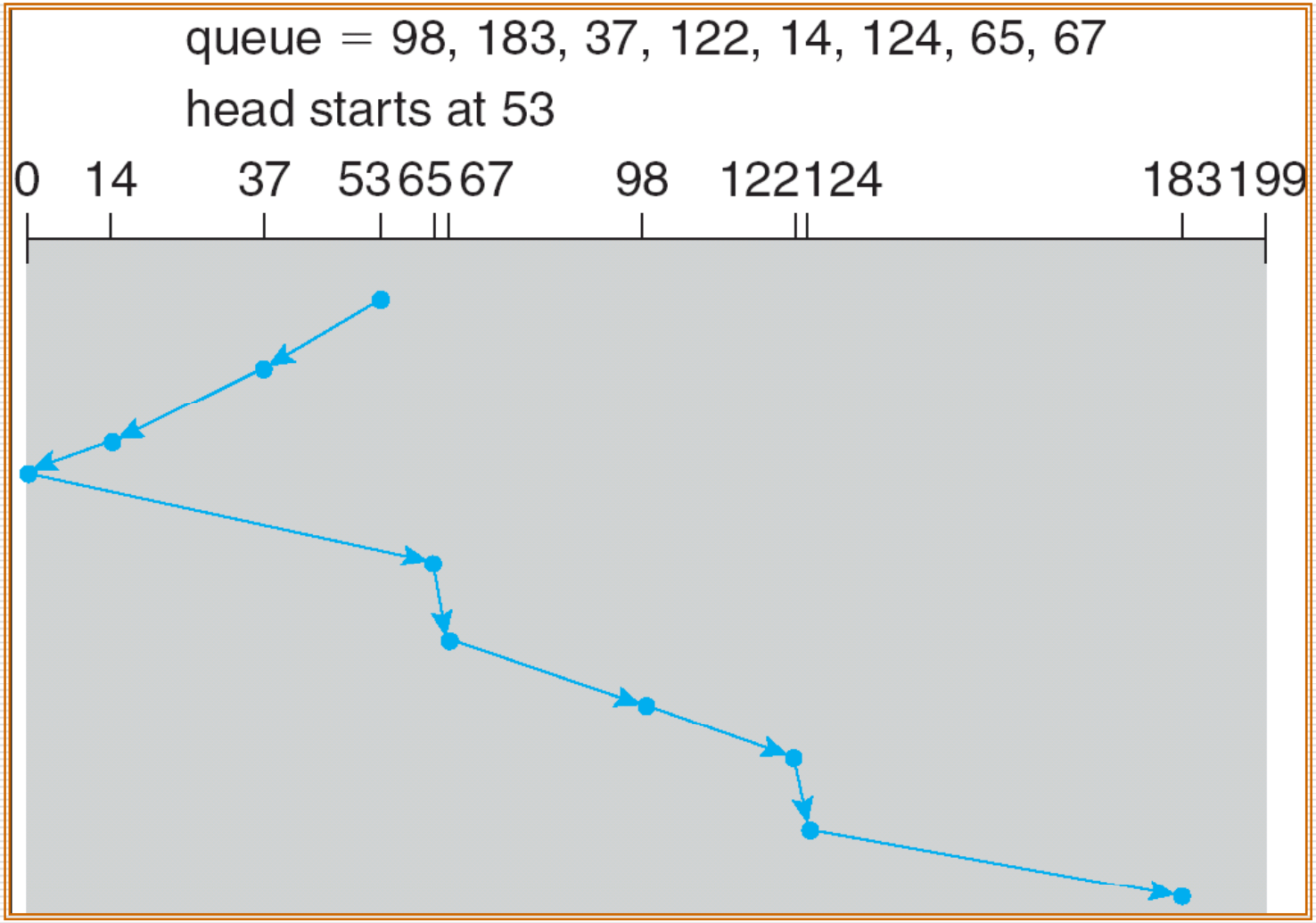
磁头总移动距离为236



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.

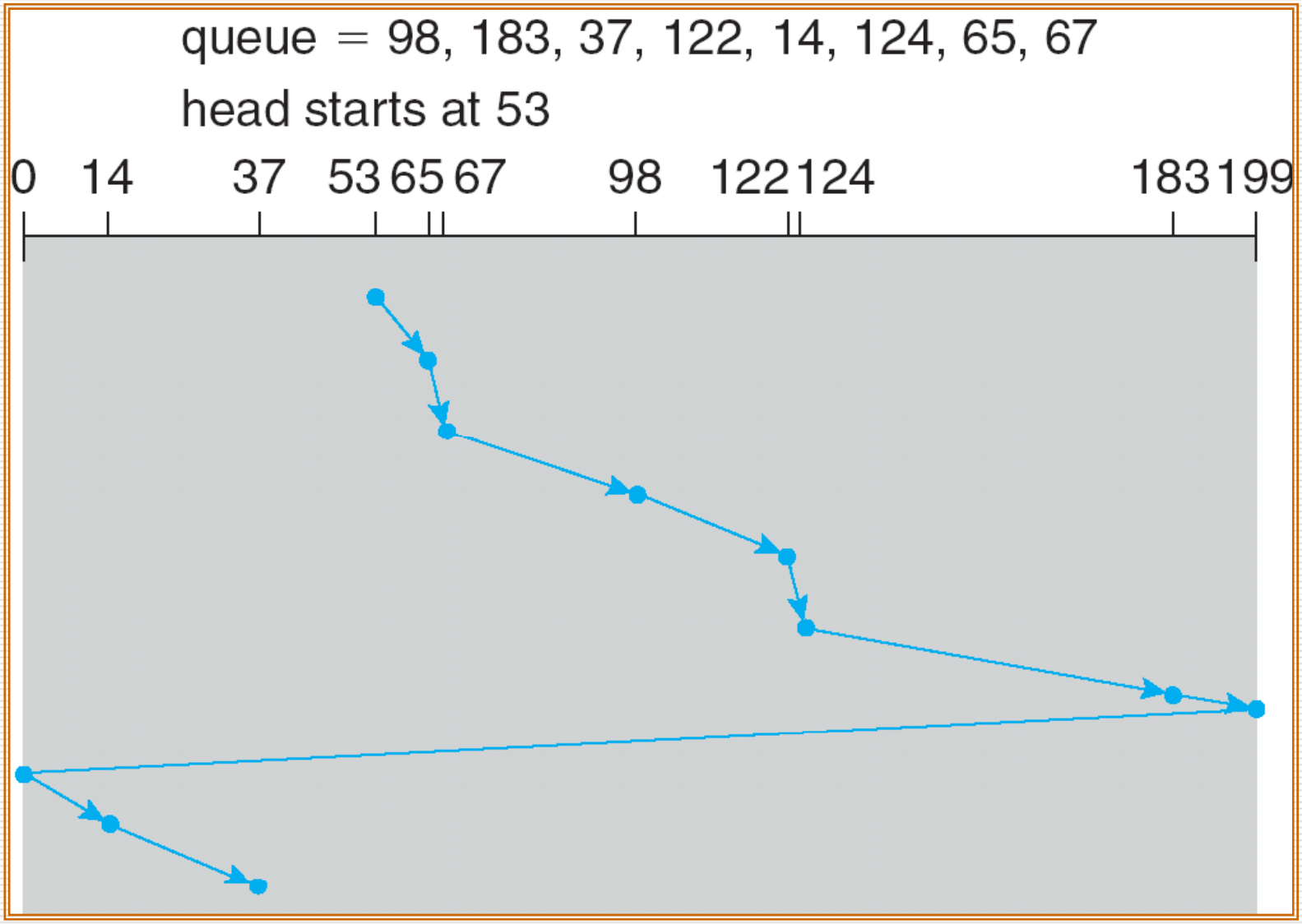
SCAN (Cont.)


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.

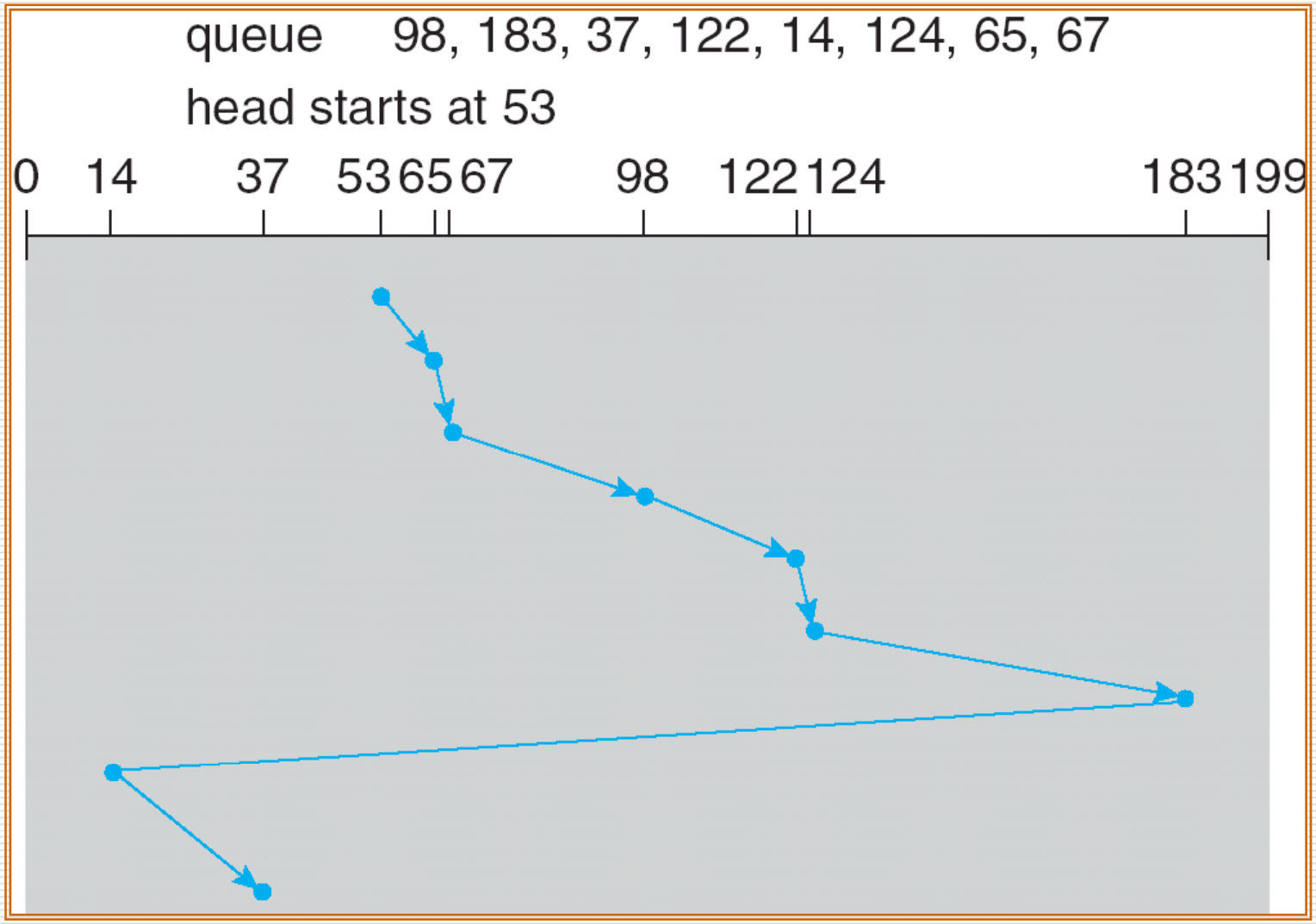
C-SCAN (Cont.)


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

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

Other algorithms

- Priority algorithm
- LIFO
- N-step-Scan
  - 把磁盘I/O请求队列分成长度为N的段，每次使用扫描算法处理这N个请求。当N=1时，该算法退化为FIFO算法
- **FSCAN**
  - 把磁盘I/O请求分成两个队列，交替使用扫描算法处理一个队列，新生成的磁盘I/O请求放入另一队列中

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

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

Disk Formating

- Low-level formating or physical formating
- Fills the disk with a special data structure for each. The Structure consists of a header, a data area, and a trailer.
- The header and tailer contain information used by the disk controller, such as a sector number and an error-correcting code(ECC).

多项式编码（CRC—循环冗余校验码）

把位串看成时多项式的系数

110001，表示成多项式 $x^5 + x^4 + 1$

生成多项式G(x)

生成多项式的高位和低位必须为1

生成多项式必须比信息串对应的多项式短。

CRC码基本思想

校验和（checksum）加在尾部，使带校验和的多项式能被G(x)除尽；检查时，用G(x)去除它，若有余数，则出错。

CRC

校验和计算算法

设G(x)为 r 阶，在位串的末尾加 r 个0，使其为m + r位，相应多项式为$x^rM(x)$；

按模2除法用对应于G(x)的位串去除对应于$x^rM(x)$的位串，可以得到一个余数；

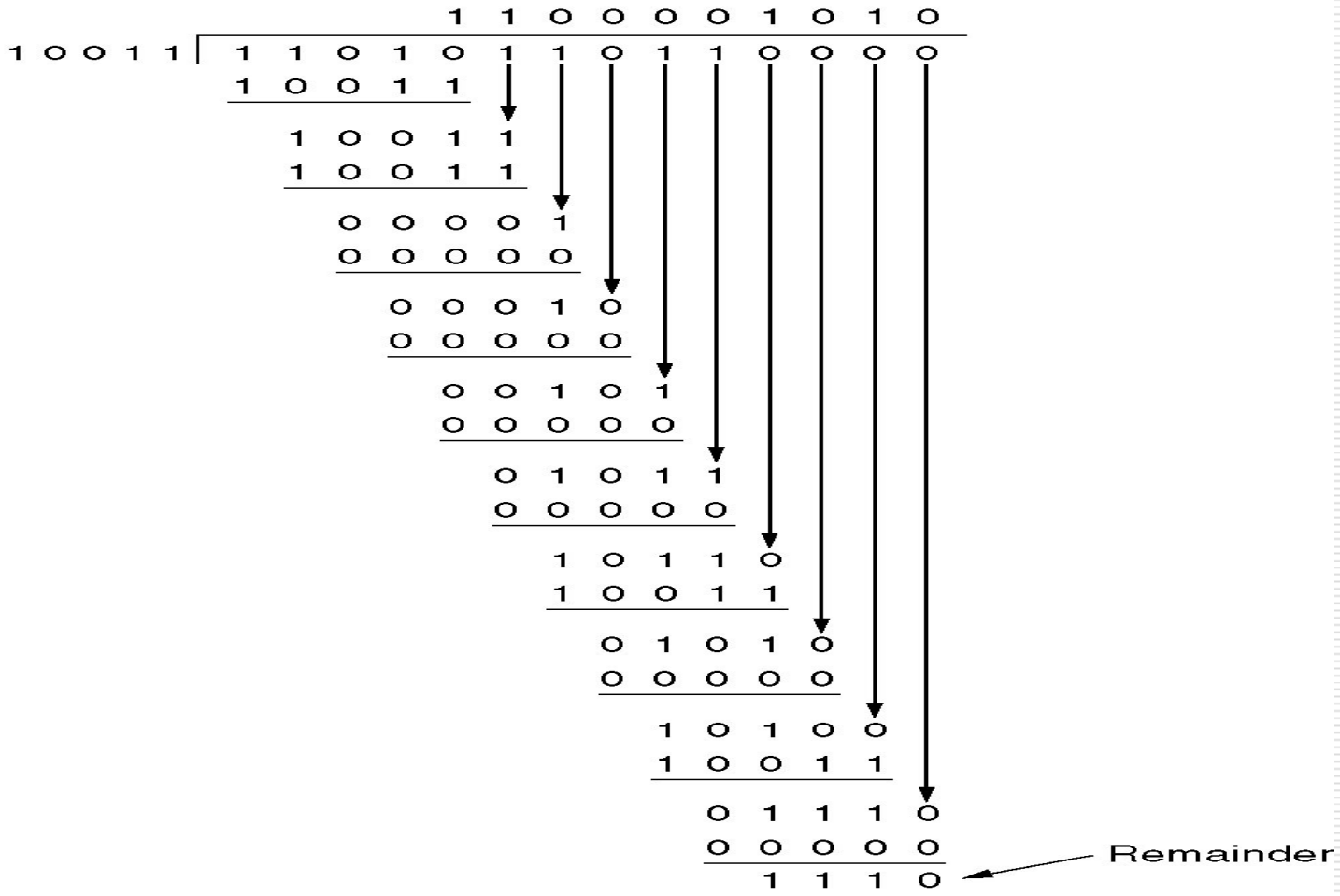
该余数可作为校验和

CRC

Frame : 1 1 0 1 0 1 1 0 1 1

Generator: 1 0 0 1 1

Message after 4 zero bits are appended: 1 1 0 1 0 1 1 0 1 1 0 0 0 0

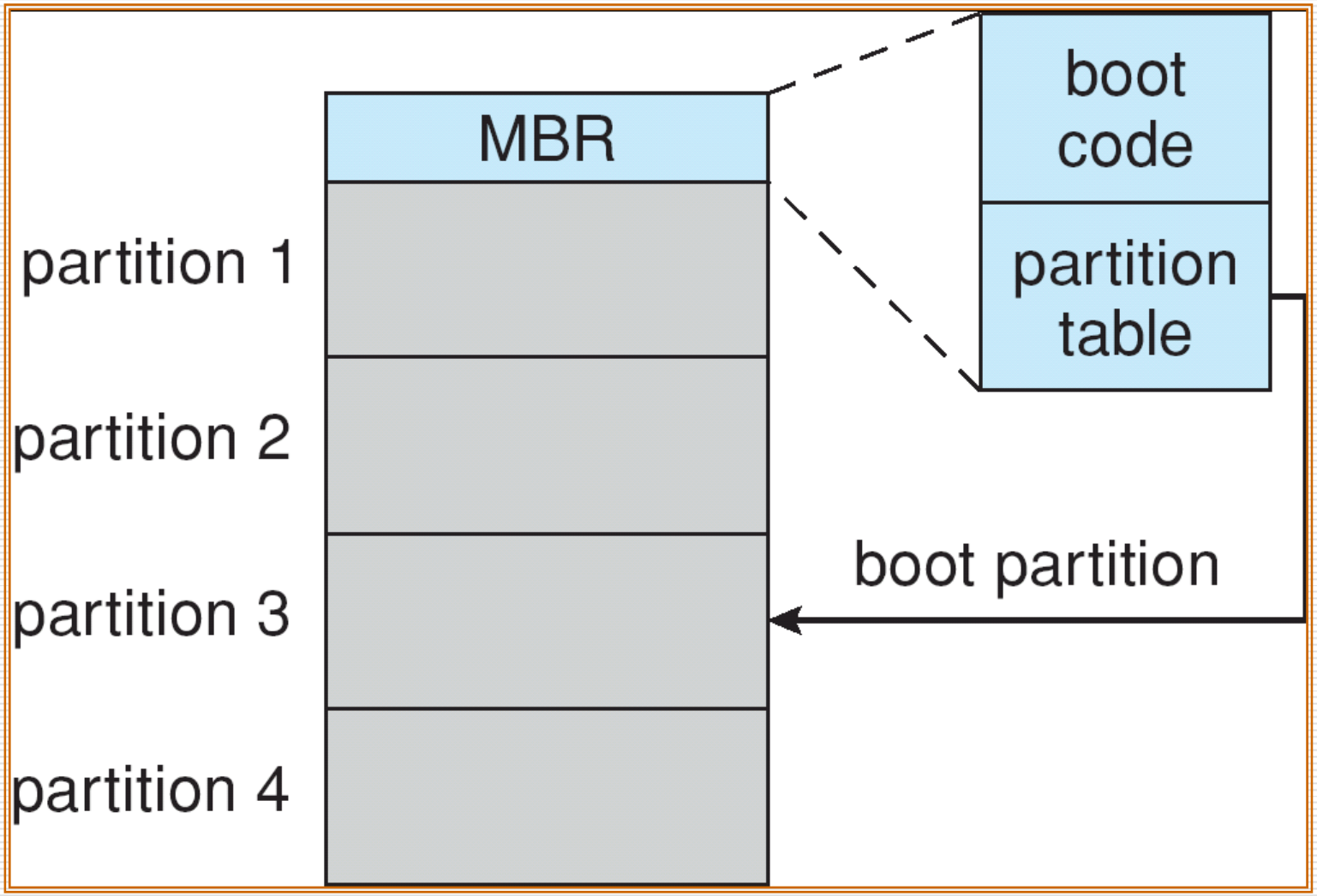


Transmitted frame: 1 1 0 1 0 1 1 0 1 1 1 1 1 0

Others

- OS still needs to record its own data structures on the disk:
  - Partition
  - Logical formating

Booting from a Disk in Windows 2000



Bad Blocks

- Most disks even come from the factory with bad blocks
  - On simple disks, bad blocks are handled manually
    - Format
    - Chkdsk
  - On sophisticated disks, are smarter about bad-block
    - Sector sparing
    - forwarding

A typical bad-sector translation

- The OS tries to read logical block 87
- The controllers calculates the ECC and finds that the sector is bad. It reports this to OS
- When rebooted again, the system tells the controller to replace the bad sector with a spare
- After that, the replacement sector's address will be used when block 87 is requested.

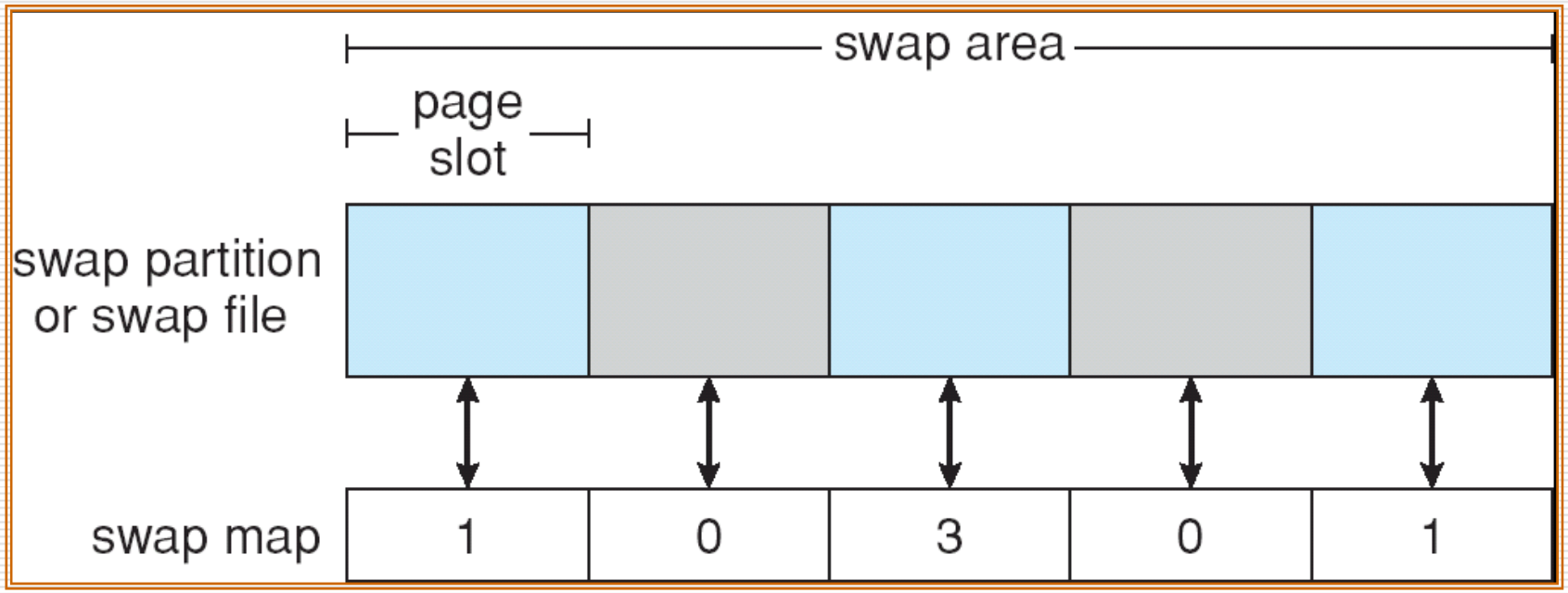
12.6 Swap-Space Management

- Swap-space — Virtual memory uses disk space as an extension of main memory.
- Swap-space use
  - To hold an entire process image
  - To store pages that have been pushed out of main memory
- Swap-space location
  - 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

- 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

Data Structures for Swapping on Linux Systems



12.7 RAID Structure

- **Redundant Array of Independent (inexpensive) Disks.**
- **RAID** – multiple disk drives provides **reliability** via **redundancy**.
- Reliability
  - Mean time to failure of a single disk is α, then ,the mean time to failure of n disks is α/n。
- Mirroring
  - The mean time to lose data depends on two factors
    - The mean time to failure of an individual disk
    - The mean time to repair
    - 100,000 hours (mean time to failure)
    - 10 hours (mean time to repair)
    - 100000^2/(2*10)→about 57000 years

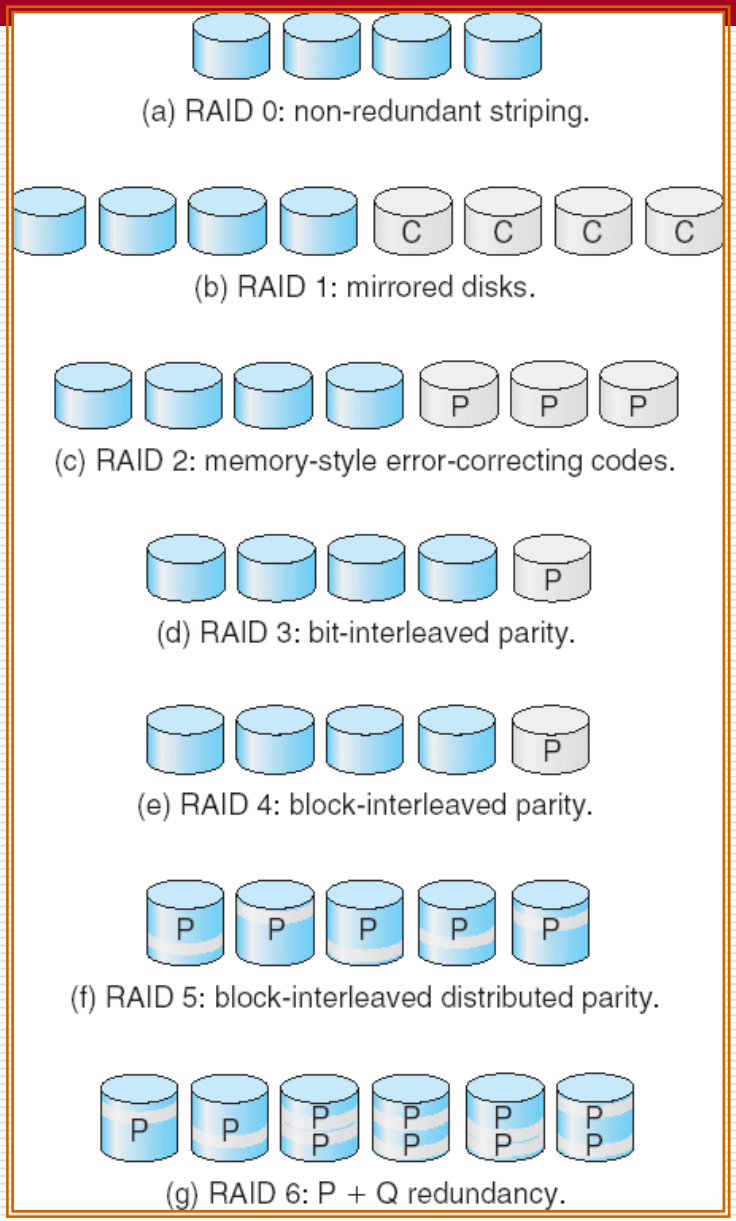
RAID Structure

- Improve performance via parallelism
  - Strip data across the disks。
    - Bit-level striping
    - Block-level striping
- RAID is arranged into six different levels.

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 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.

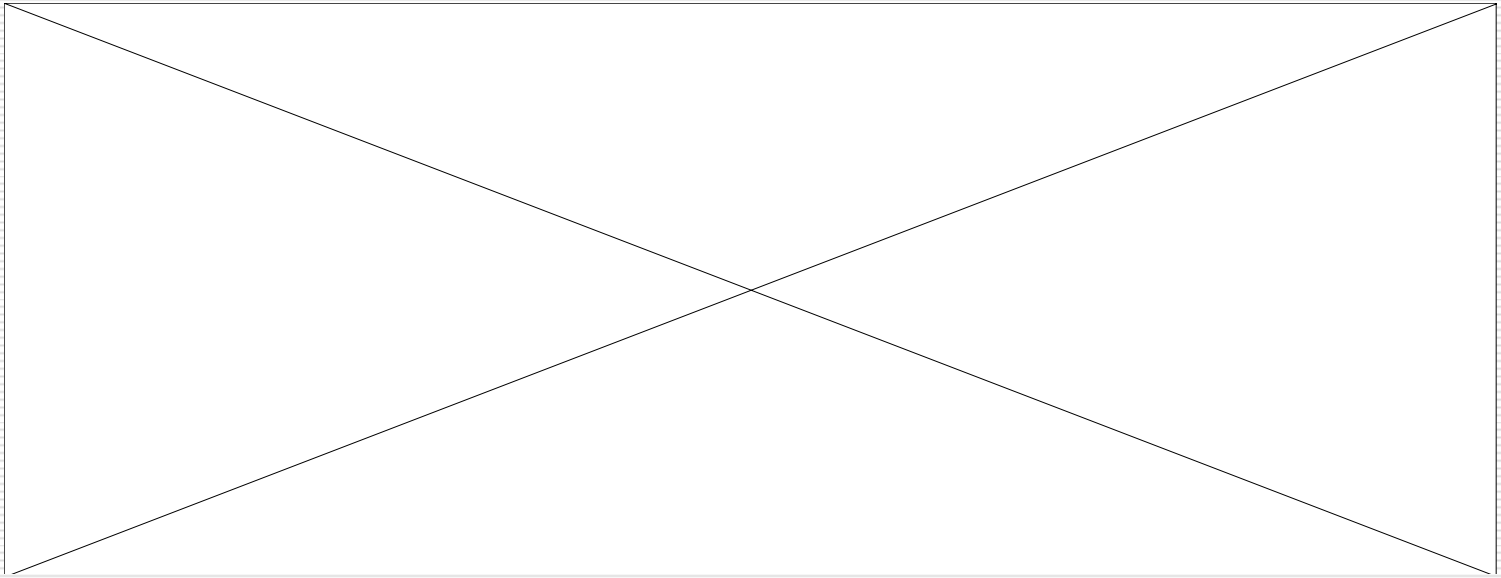
(e) RAID 4: block-interleaved parity.

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

(g) RAID 6: P + Q redundancy.

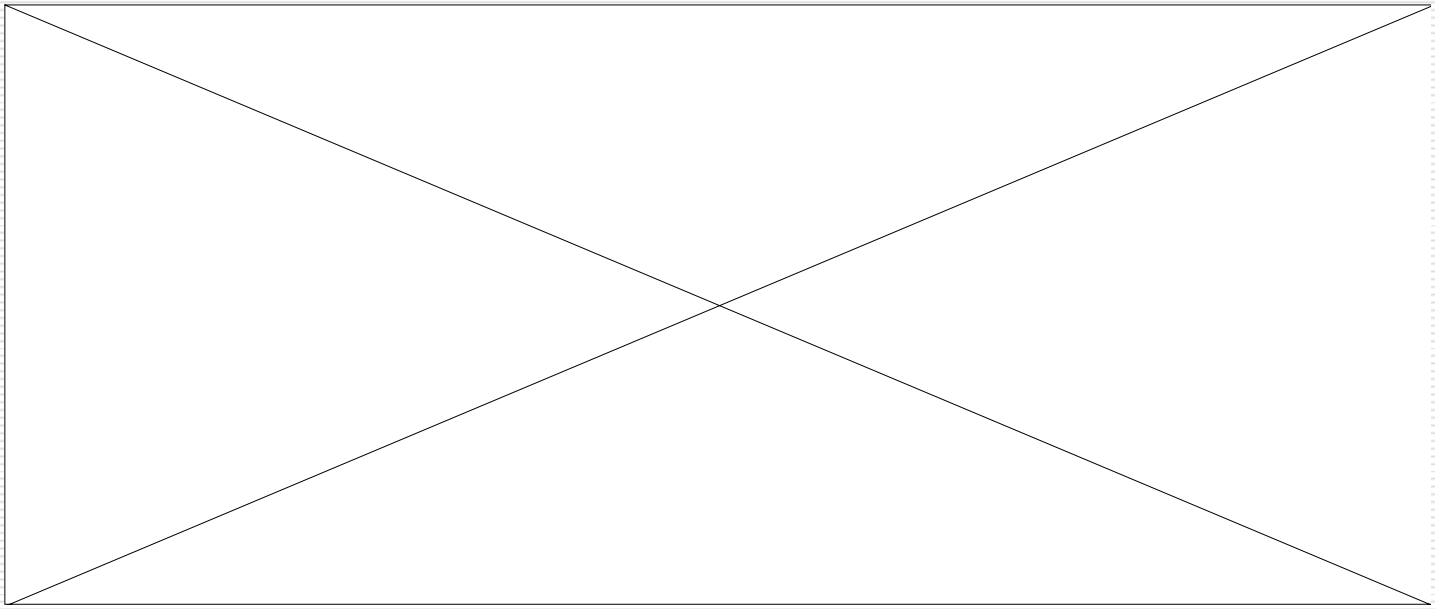
RAID Level 0

- Stripping at the level of blocks



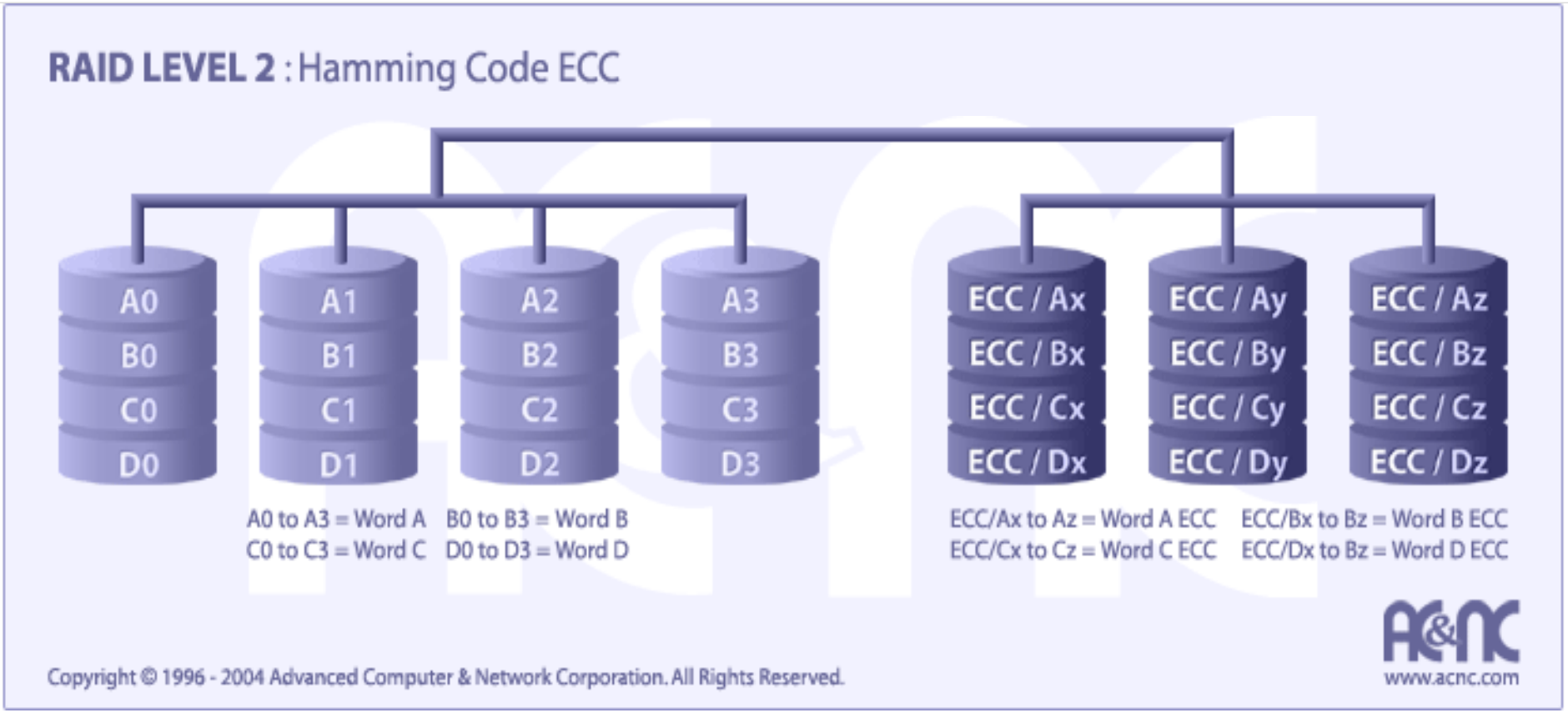
RAID Level 1

- Mirror(mirroring)
  - completely duplicated
  - high recoverability
  - expensive



RAID Level 2

- 内存方式的差错纠正代码结构



纠错码—海明码

- R.Hamming在1950年提出
- 是一种可以纠正一比特错的编码

基本思想：

- 在k比特信息上附加r比特冗余信息（也称校验比特），构成n=k+r比特的码字，其中每个校验比特和某几个特定的信息比特构成偶校验的关系。
- 接收端对这r个奇偶关系进行校验，即将每个校验比特和与它关联的信息比特进行相加（异或），相加的结果称为校正因子。
- 如果没有错误的话，这r个校正因子都为0
- 如果有一个错误，则校正因子不会全为0
- 根据校正因子的不同取值，可以知道错误发生在码字的哪一个位置上
- 校验比特数r必须满足以下条件：$2^r \geq n+1$，即$2^r \geq k+r+1$。

4位或7位的信息位，需要几位校验位？

Hamming Code

构造方法：

•以k=4为例，要满足上述不等式，则有r≥3，如取r=3,于是n=k+r=7。

I_4	I_3	I_2	r_2	I_1	r_1	r_0
7	6	5	4	3	2	1

•2^k位置上是校验比特，其余位置上是信息比特I_i。

•将每个信息比特的位置写成2的幂次之和的形式，即：

$$7 = 2^2 + 2^1 + 2^0$$

$$6 = 2^2 + 2^1$$

$$5 = 2^2 + 2^0$$

$$3 = 2^1 + 2^0$$

Hamming Code

构造方法（续）：

·计算校验比特的公式：

$$r_2 = I_4 + I_3 + I_2$$

$$r_1 = I_4 + I_3 + I_1$$

$$r_0 = I_4 + I_2 + I_1$$

I4	I3	I2	r2	I1	r1	r0
7	6	5	4	3	2	1
1	0	0		0		

相加实际是进行异或运算

Hamming Code

构造方法（续）：

·计算校验比特的公式：

$$r_2 = I_4 + I_3 + I_2$$

$$r_1 = I_4 + I_3 + I_1$$

$$r_0 = I_4 + I_2 + I_1$$

·校正因子：

$$S_2 = r_2 + I_4 + I_3 + I_2$$

$$S_1 = r_1 + I_4 + I_3 + I_1$$

$$S_0 = r_0 + I_4 + I_2 + I_1$$

·如果没有错误的话，这三个校正因子都为0

·如果校正因子不全为0，则有错误发生，错误的位置就在$S = S_2 S_1 S_0$处

·将相应位取反，即得到正确的数据

Hamming Code

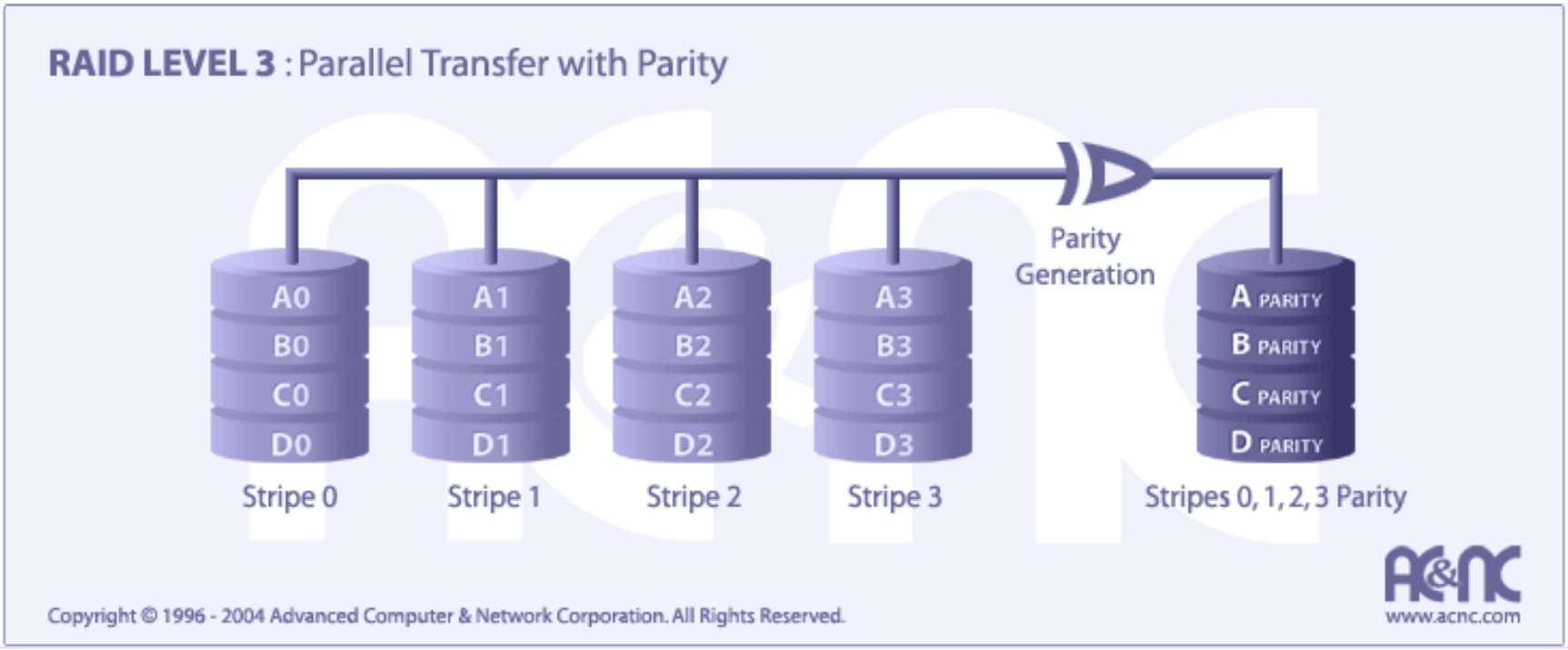
- 对于一段信息1000，则
- r_2= 1+0+0=1，
- r_1 =1+0+0=1，
- r_0 =1+0+0=1，
- 于是码字为1001011

$$S_2 = r_2 + I_4 + I_3 + I_2$$
$$S_1 = r_1 + I_4 + I_3 + I_1$$
$$S_0 = r_0 + I_4 + I_2 + I_1$$

- 假如在接收端收到码字1011011 ，则
- S_2=1+1+0+1=1，S_1=1+1+0+0=0，S_0=1+1+1+0=1
- 说明码字有错，错误位置为$S = S_2 S_1 S_0 = 101 = 5$
- 将比特5上的1变为0， 1011011—〉 1001011，即可纠正错误

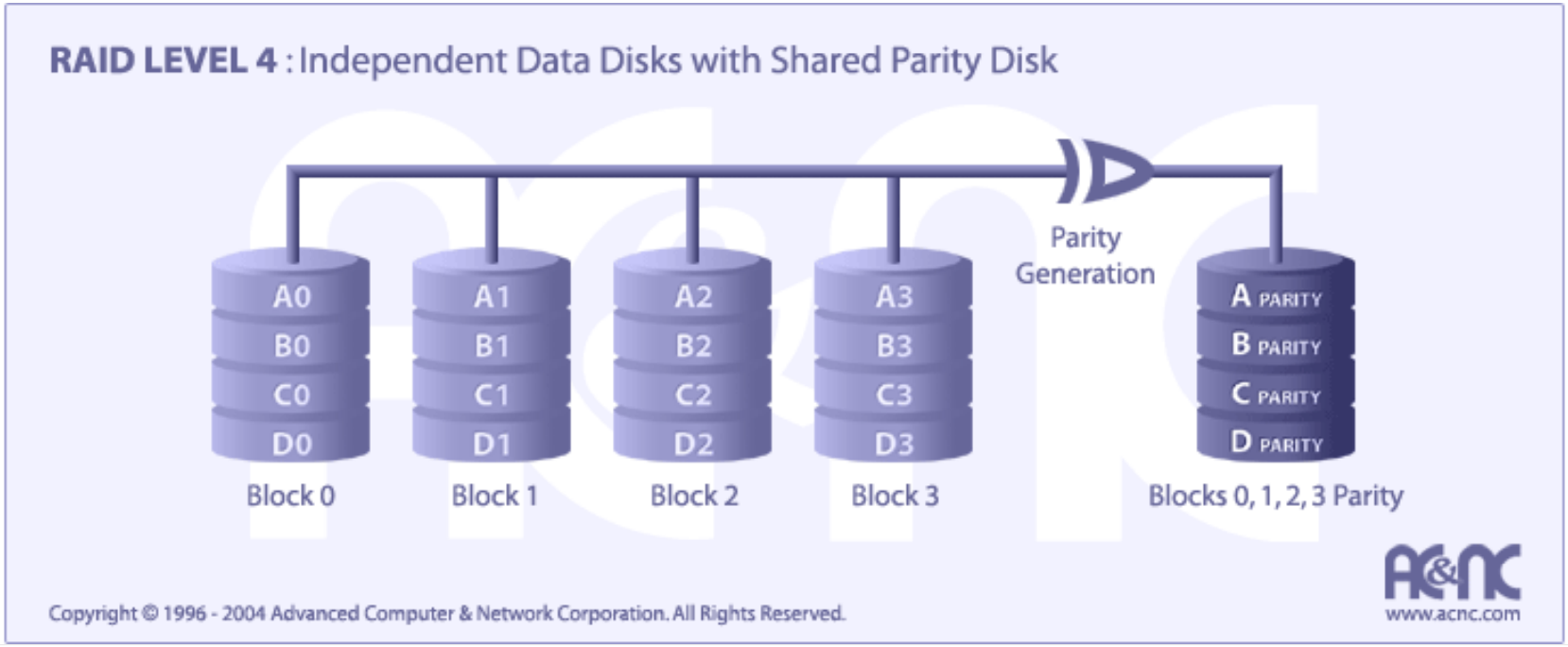
RAID Level 3

- 位交织奇偶结构
  - Less disks are required
  - Fast reading and writing speed



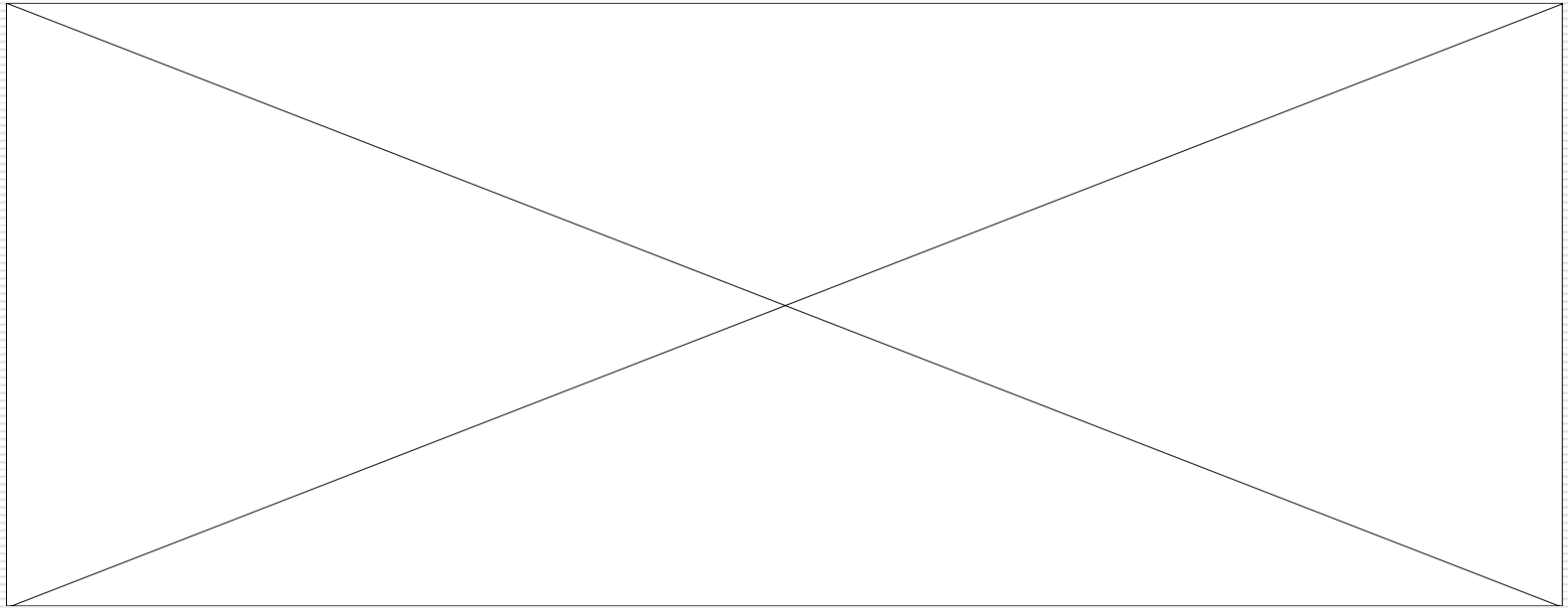
RAID Level 4

- 块交织奇偶结构



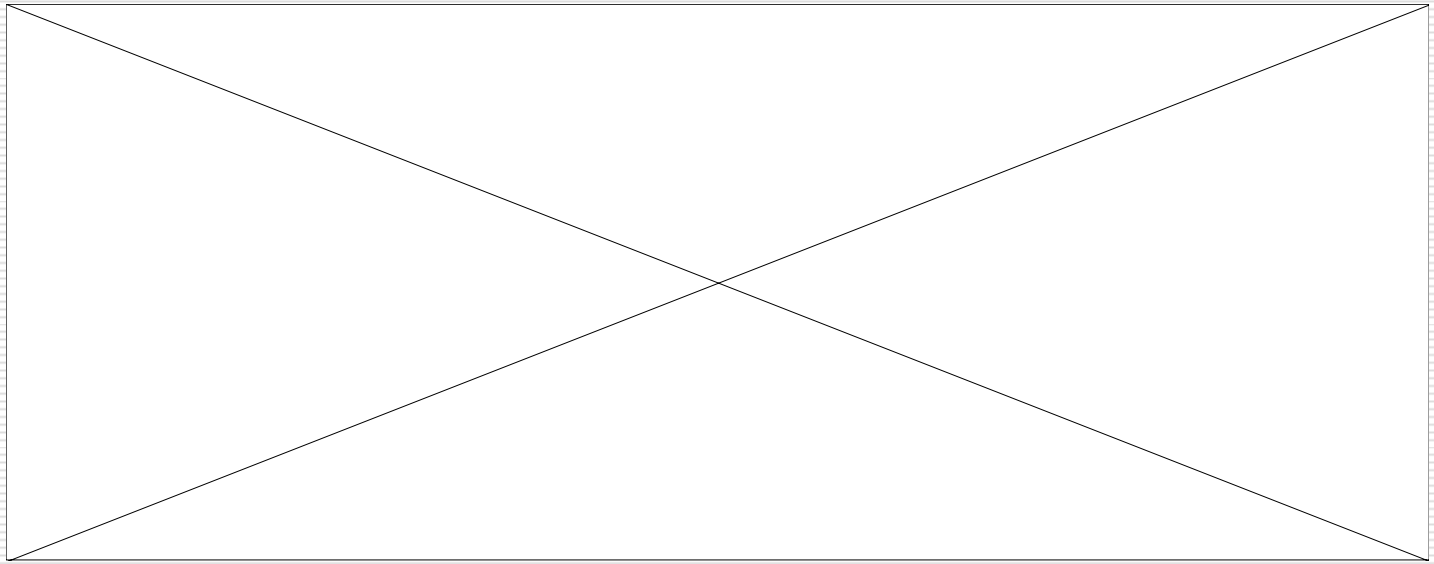
RAID Level 5

- 块交织分布式奇偶结构
- 奇偶块存放在不同的磁盘上

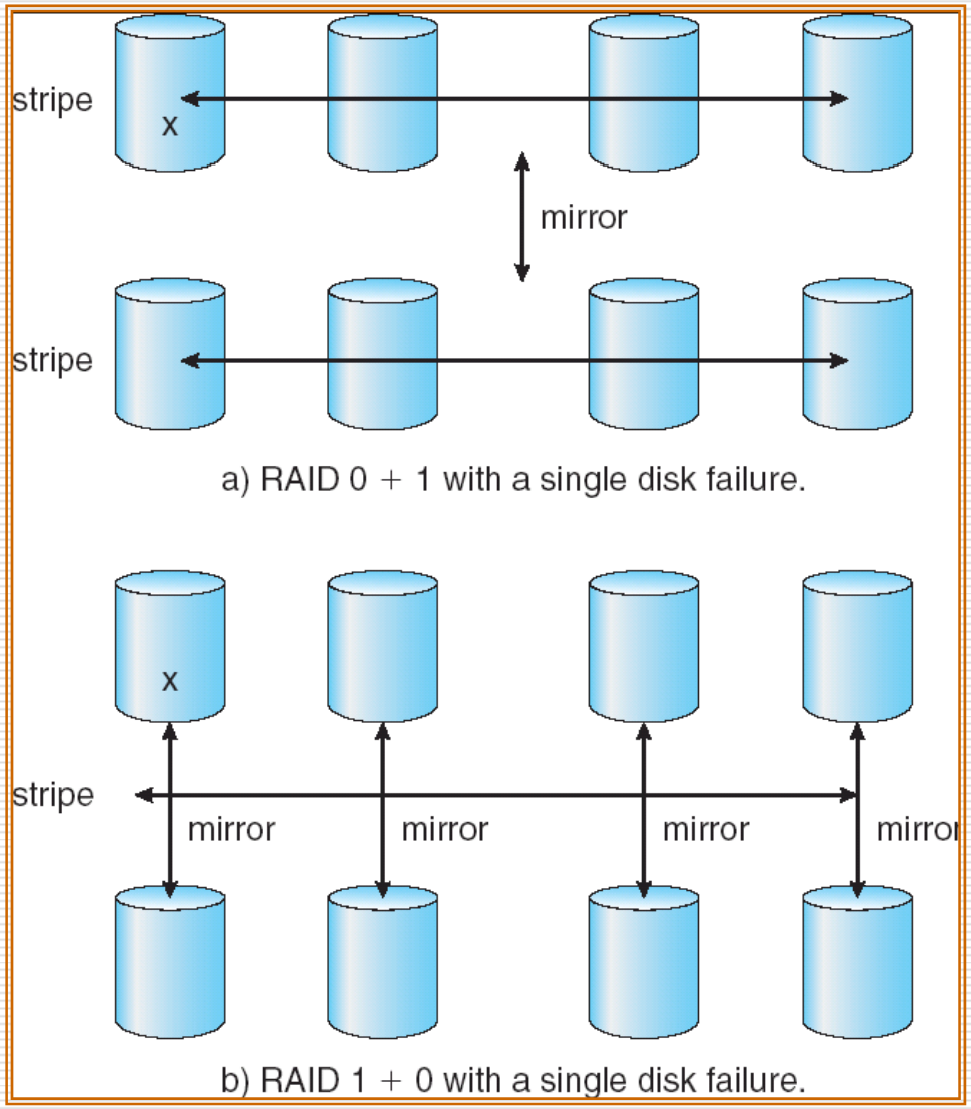


RAID Level 6

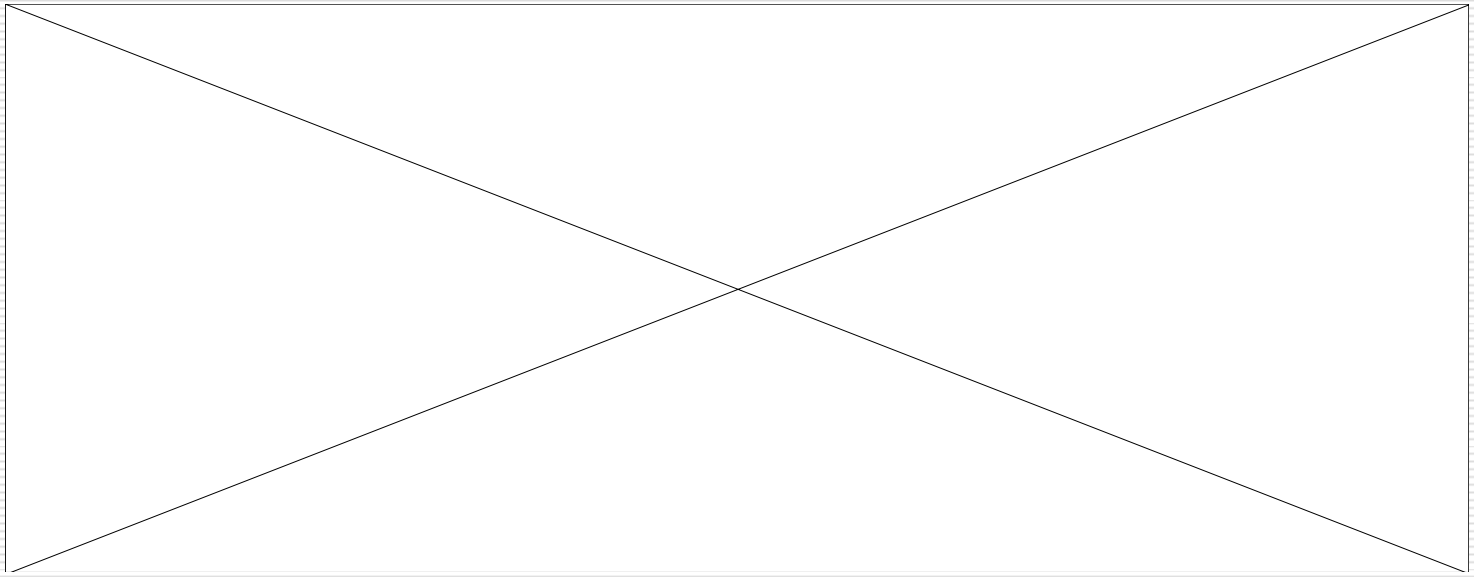
- P+Q冗余方案，可以解决多个磁盘出错
- 没有使用奇偶校验，而是使用了差错纠正码
  - e.g. reed-selomon code



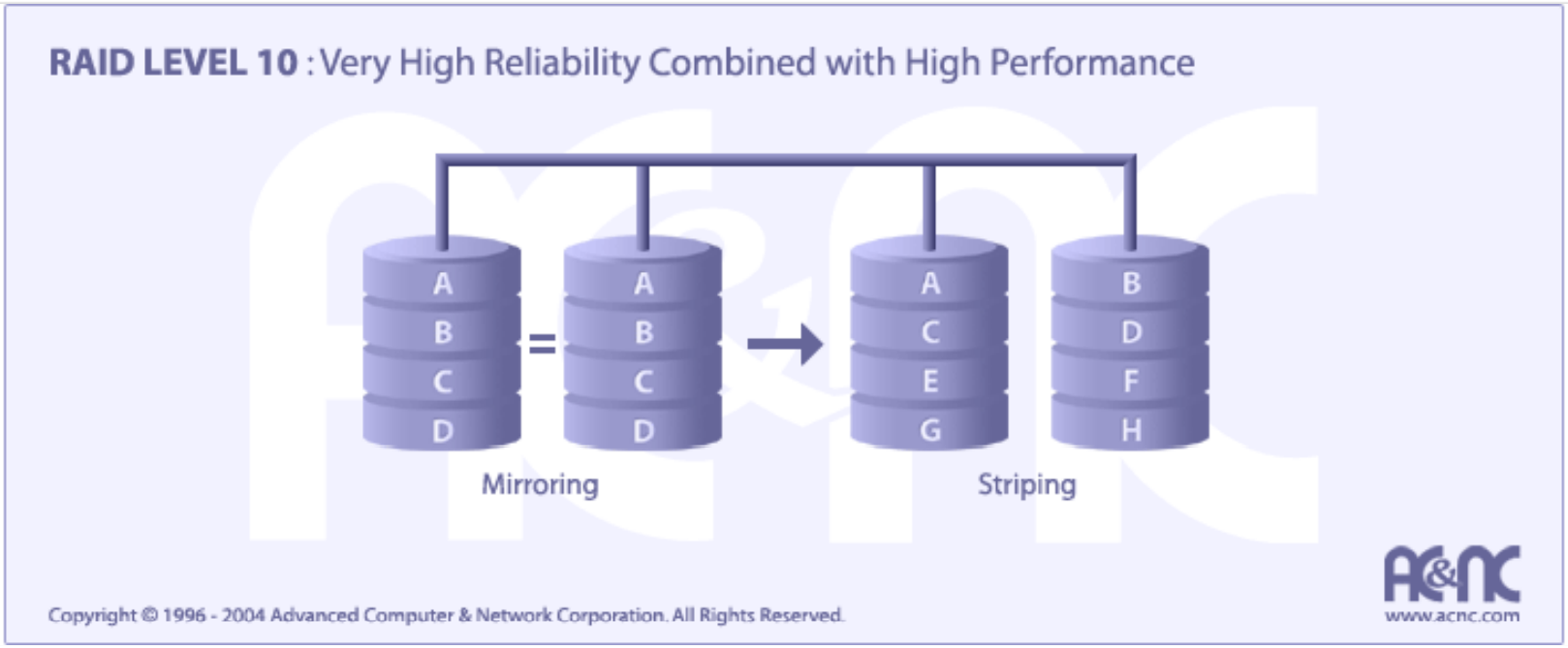
RAID (0 + 1) and (1 + 0)



RAID 0+1



RAID 1 + 0


RAID LEVEL 10 : Very High Reliability Combined with High Performance
A
B
C
D
=
A
B
C
D
A
C
E
G
B
D
F
H
Mirroring
Striping
Copyright © 1996 - 2004 Advanced Computer & Network Corporation. All Rights Reserved.
AC&NC
www.acnc.com

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.

Operation example

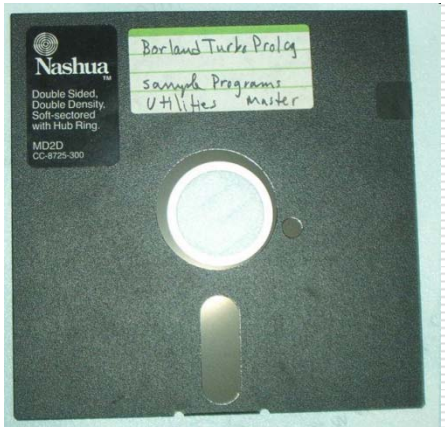
- An output operation can be executed as follows:
  - Write it onto the first physical block
  - Write it onto the second physical block
  - Declare the operation successfully

12.9 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.

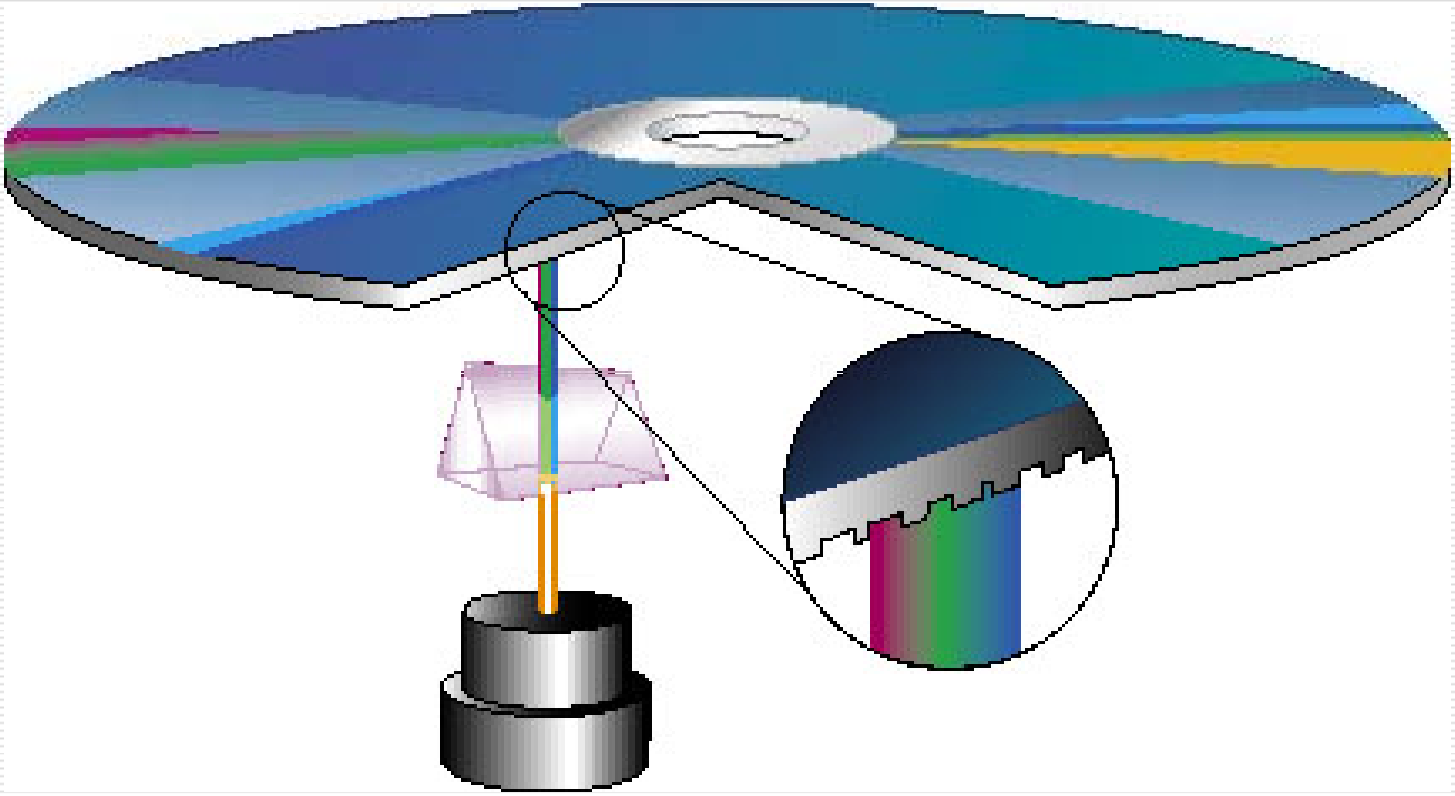


Removable Disks (Cont.)

- 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.

Removable Disks (Cont.)

- 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.

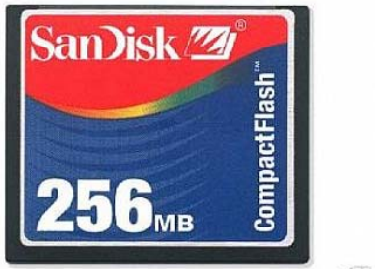
- **CD-ROM** (Compact Disc)
  - Optical drives that *read* CD-ROMs
- **CD-R recordable**
  - *WORM* media (write-once, read many)
- **CD-RW rewritable**
  - Can read CD-ROMs and write, erase and
  - rewrite data onto CD-R & CD-RW disks
- **DVD** (Digital Versatile Disks)
  - Store & distribute all kinds of data
  - Hold between 3.8 and 17 gigabytes of information
- **DVD-ROM** drives
  - DVD/CD-RW-Can play DVD movies, read DVD data disks, read standard CD-ROMs, and play audio CDs
  - Because they're read-only, they can't record data, music, or movies
- **DVD-RAM drives**
  - Can read, erase, and write data (but not DVD video) on multi-gigabyte DVD-R (some times CD-R or CD-RW) media
- DVD-RW
- DVD+RW
- DVD+MRW

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.

Solid-State Storage Devices

- Flash memory is an erasable memory chip
  - Sizes range from 16 MB to many GBs
  - Compact alternative to disk storage
  - Contains no moving parts
  - Designed for specific applications such as storing pictures in digital cameras
  - Likely to replace disk and tape storage



Operating System Issues

- 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.

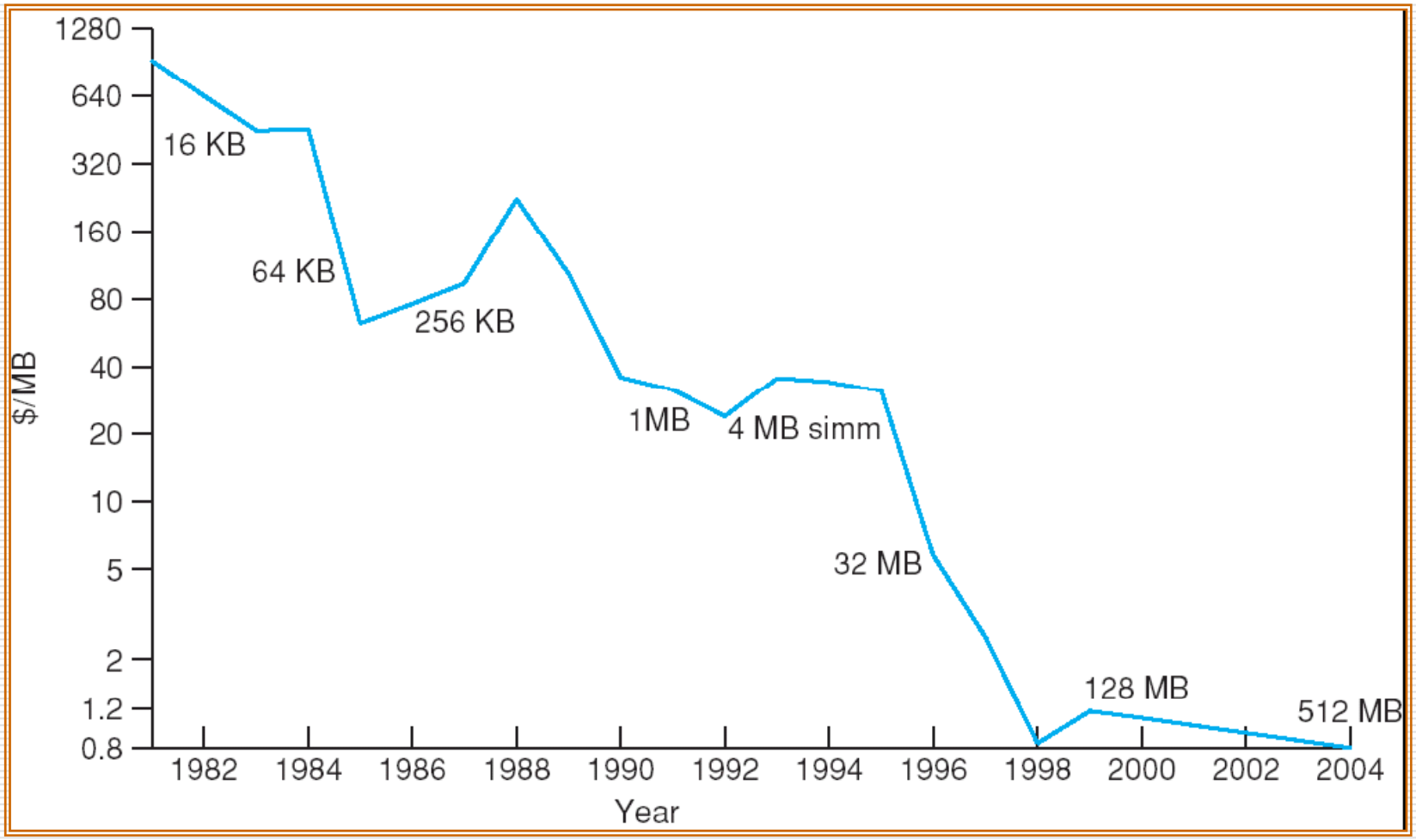
Reliability

- 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.

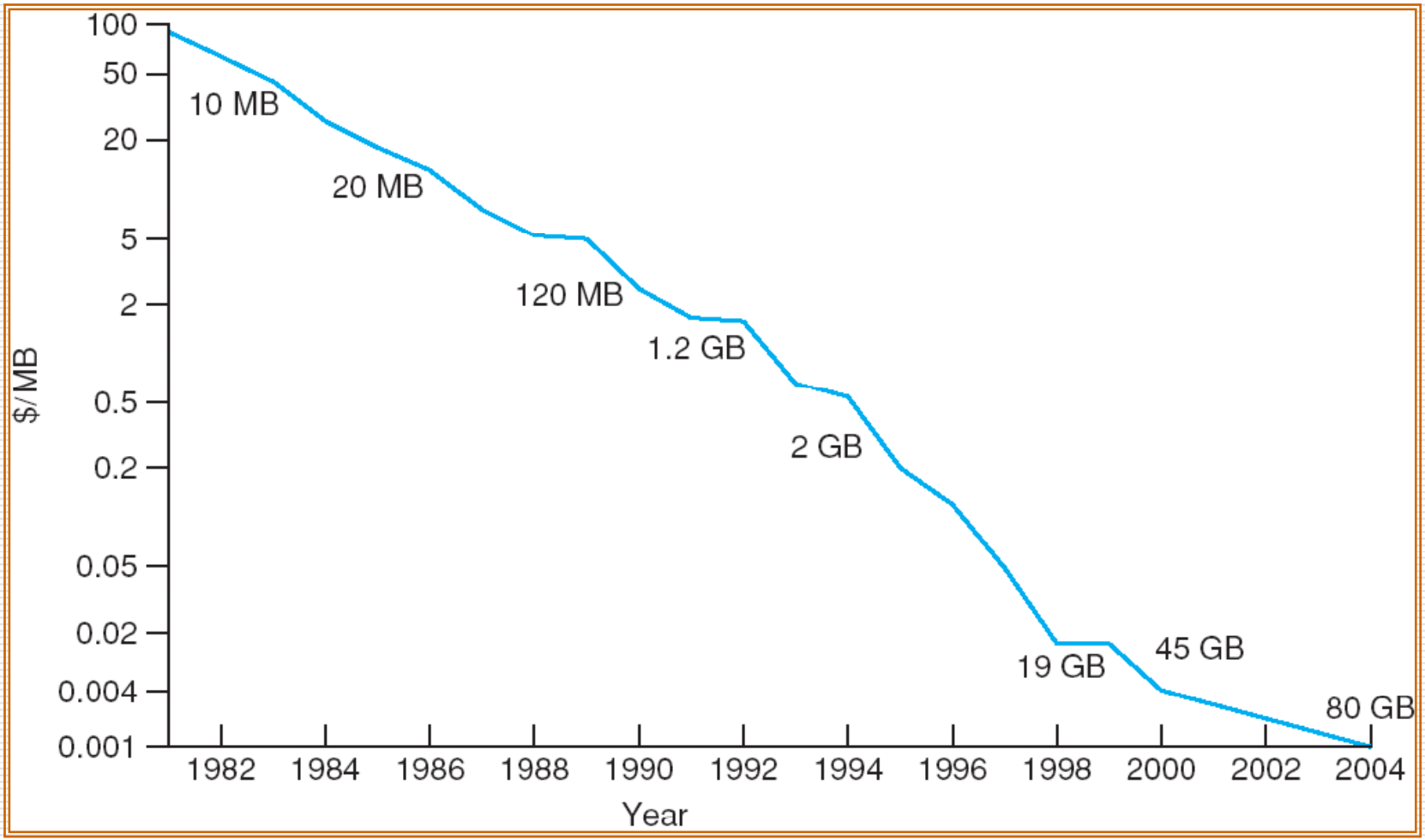
Cost

- 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.

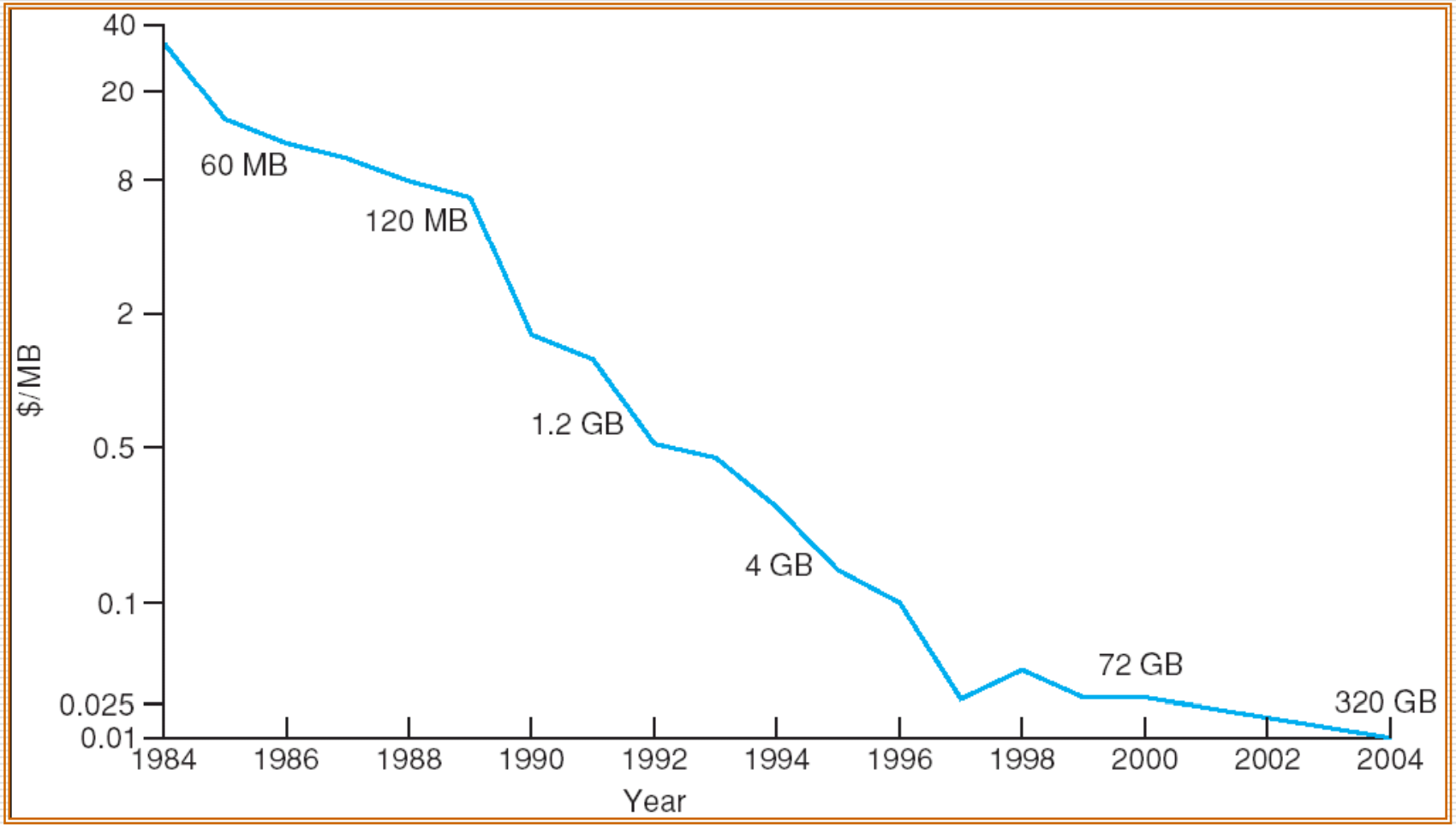
Price per Megabyte of DRAM, From 1981 to 2004



Price of Magnetic Hard Disk, From 1981 to 2004



Price of a Tape Drive, From 1984-2000



Assignment

- 2, 4, 10

End of Chapter 12

Any Question?