Operating System

BITS Pilani
K K Birla Goa Campus

Dr. Lucy J. Gudino
Dept. of CS and IS

External Memory

- Magnetic Disk
- RAID

Magnetic Disk

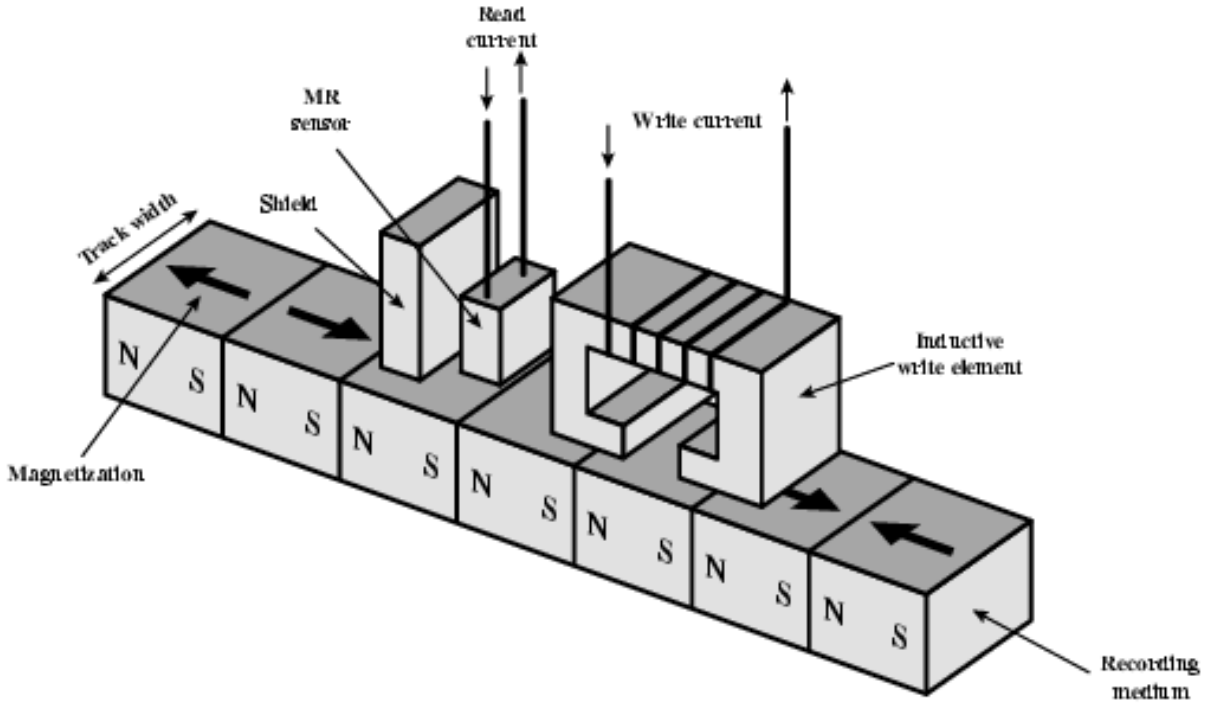
- Off line storage, non volatile
- Provide bulk storage
- is a circular platter constructed of non-magnetic material ➔ substrate
  - Substrate is coated with magnetizable material (Iron oxide)
  - Platter diameter range from 1.8 to 5.25inches
  - Disks rotate at 5400RPM to 15,000 RPM
  - Read/Write head “flies” just above the platter
  - Traditional system used aluminium for substrate
  - Now glass as substrate

Contd

- Glass as substrate (Advantages)
  - Improved surface uniformity
    - Increases reliability
  - Reduction in surface defects
    - Reduced read/write errors
  - Ability to support lower fly heights (2 to 3 nm)
  - Better stiffness to reduce disk dynamics
  - Greater ability to withstand shock and damage

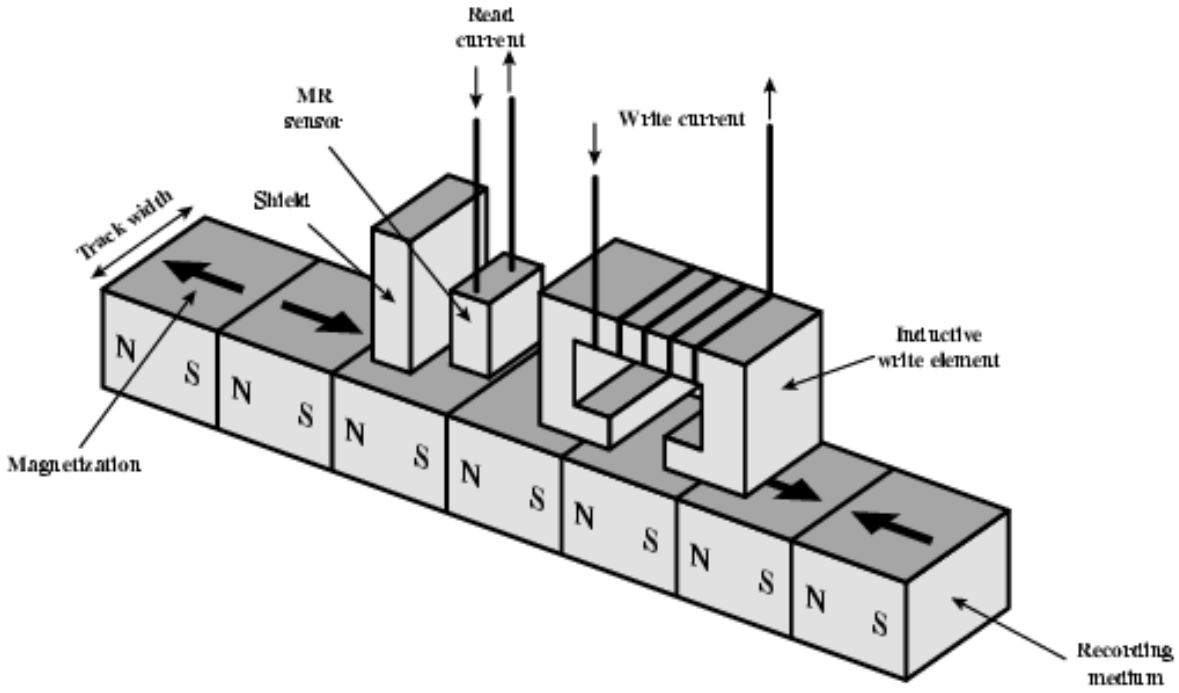
Read and Write Mechanisms

- Recording & retrieval via conductive coil called a head
- May be single read/write head or separate ones
- During read/write, head is stationary, platter rotates



Read and Write Mechanisms ...

- Write Operation
  - Basic principle : Current through coil produces magnetic field
  - Electric pulses sent to head
  - Magnetic pattern recorded on surface below

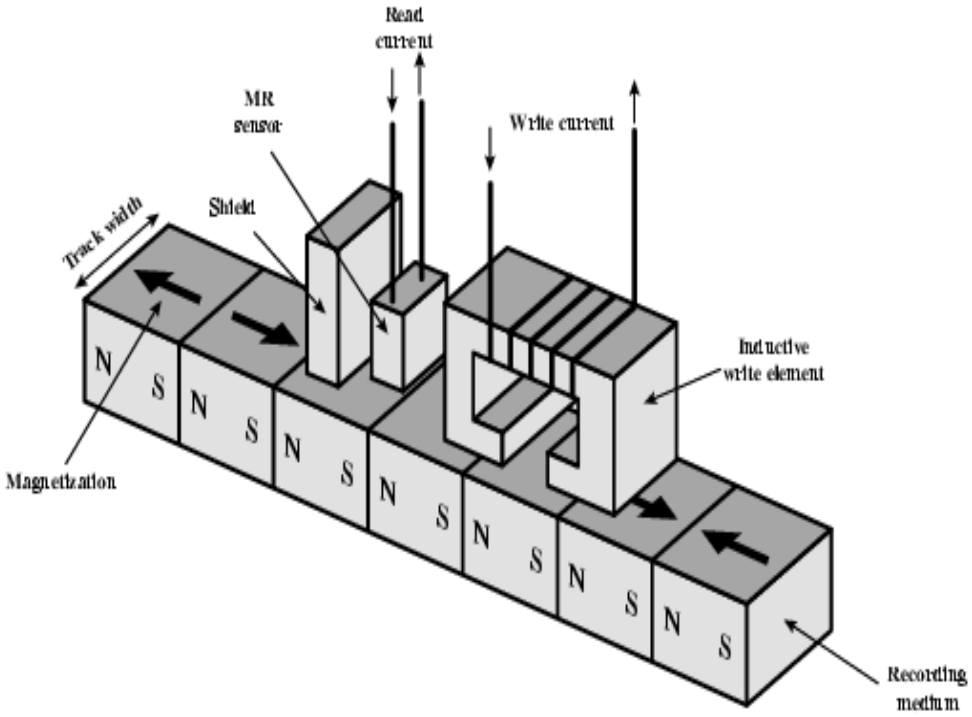


Contd…

- Read (traditional)
  - Magnetic field moving relative to coil produces current in the coil
  - Coil is the same for read and write
  - used in floppy disk and older rigid disk systems

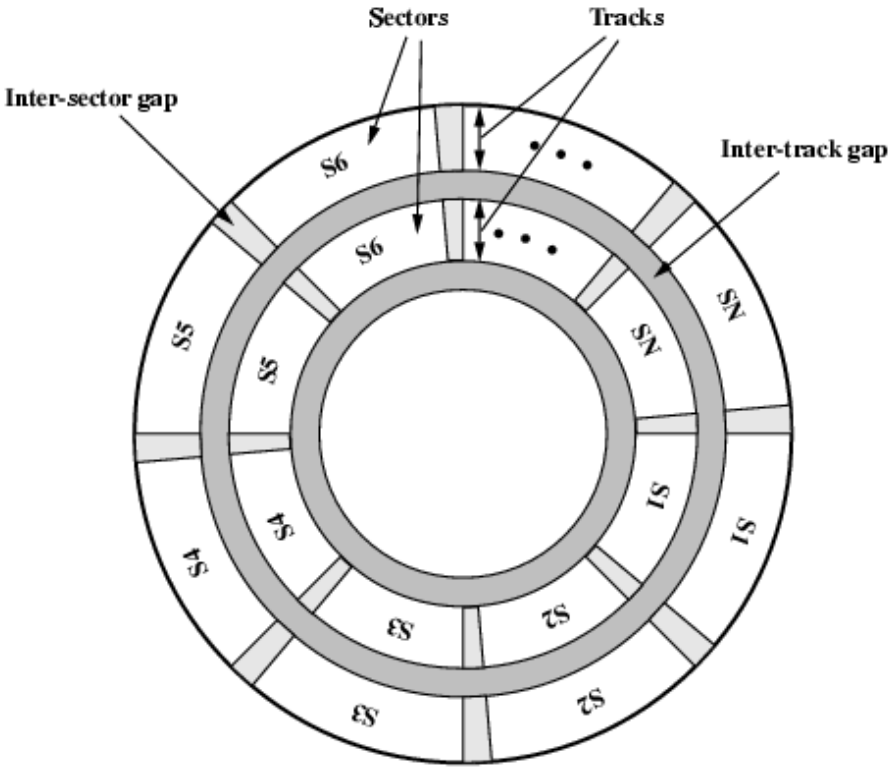
Contd…

- Read (contemporary)
  - Separate read head, close to write head
  - Partially shielded magneto resistive (MR) sensor
  - Electrical resistance depends on direction of magnetic field



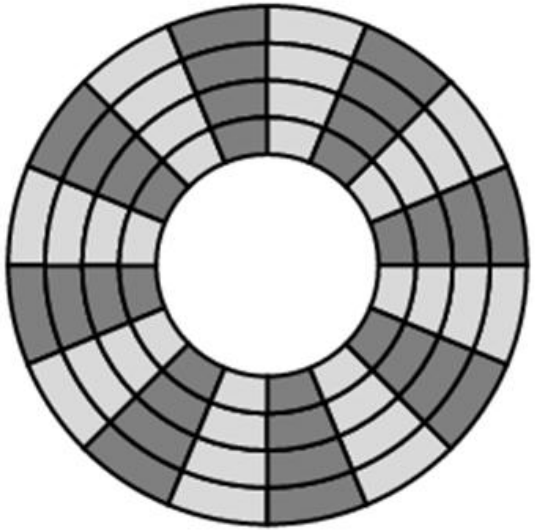
Data Organization and Formatting

- Concentric rings or tracks
  - Gaps between tracks : prevents/minimizes errors due to misalignment of the head or simply interference of magnetic field
  - Reduce gap to increase capacity
  - Same number of bits per track
  - Constant angular velocity
  - 10,000 to 50,000 tracks per surface
- Tracks divided into sectors
  - fixed or variable length
  - contemporary systems use fixed length system with 512 byte sector size
  - 100 to 500 sector



contd...

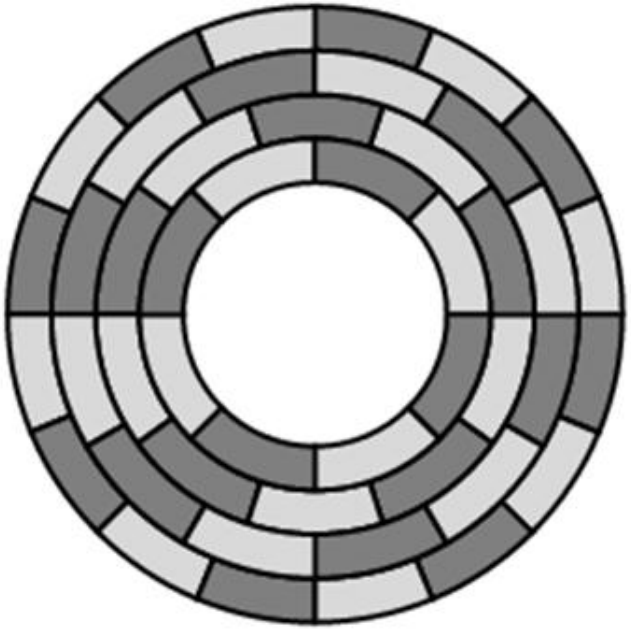
- Rotate disk at constant angular velocity (CAV)
  - Gives pie shaped sectors and concentric tracks
  - Individual tracks and sectors addressable
  - Move head to given track and wait for given sector
  - Waste of space on outer tracks
    - Lower data density
- Can use zones to increase capacity
  - Each zone has fixed bits per track
  - More complex circuitry

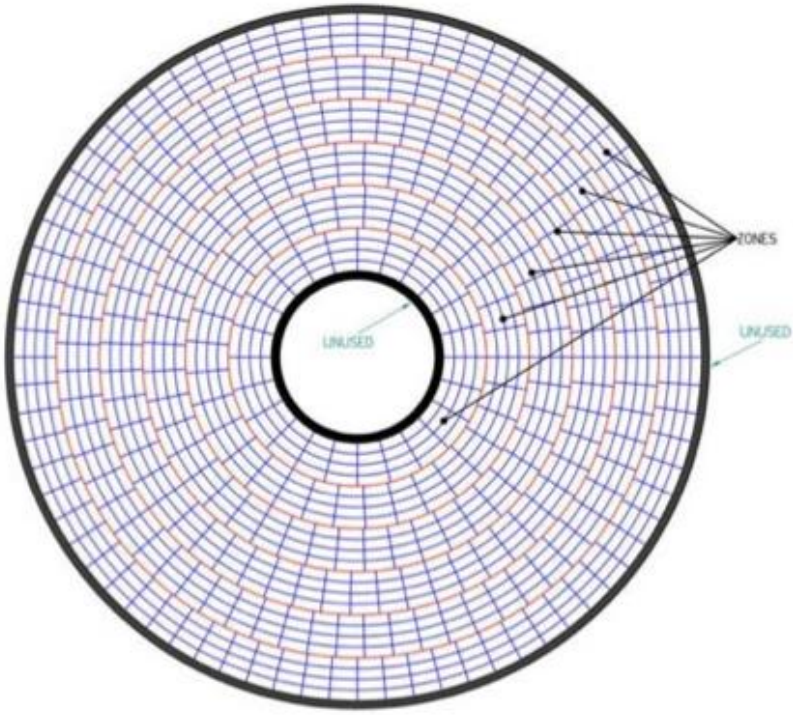


(a) Constant angular velocity

Multiple Zone Recording

- Also known as zoned bit recording
- Grouping the tracks in to sets called zones (16 is typical)
- Within a zone, the number of bits per track is constant.
- Tracks in the inner zones contain the fewest sectors, and tracks in the outer zones contain the most sectors.

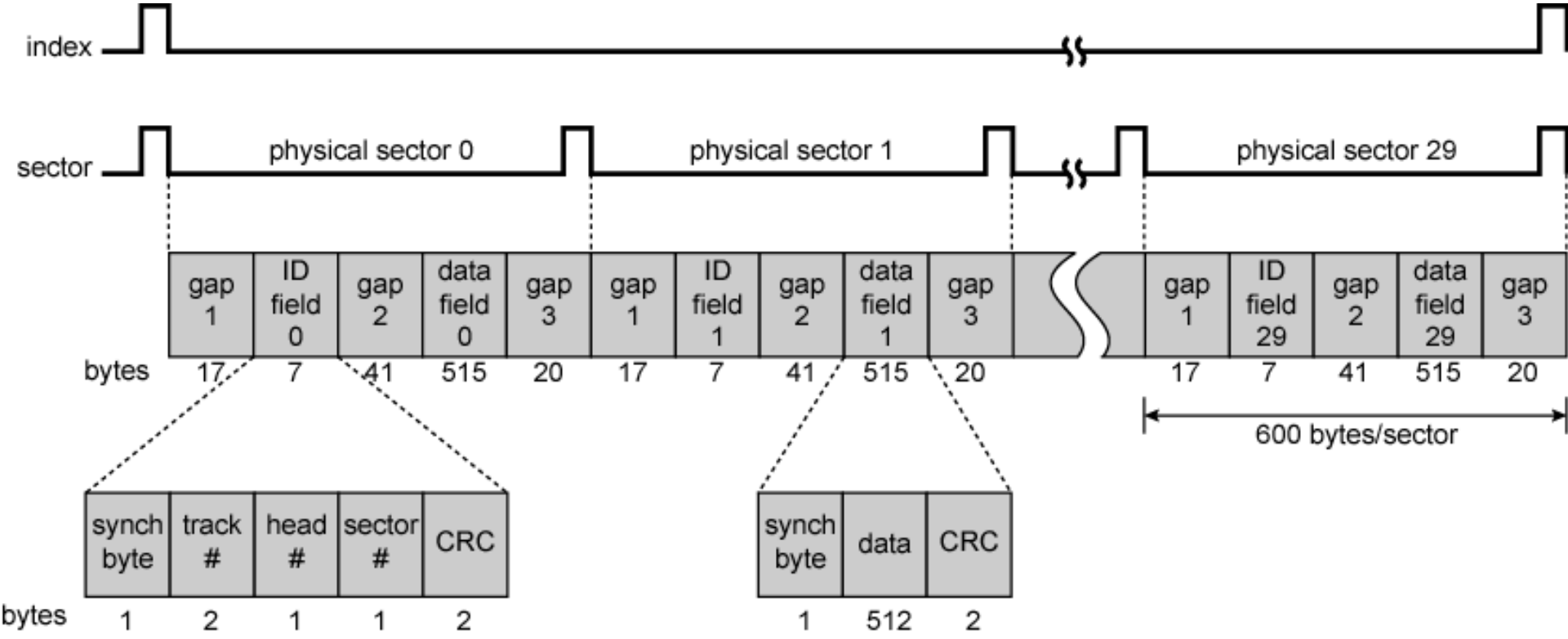




Finding Sectors

- Must be able to identify start of track and sector
- Format disk
  - Additional information not available to user
  - Marks tracks and sectors

Winchester Disk Format
Seagate ST506



Physical Characteristics of Disk Systems

- Head motion: Fixed (rare) or movable head
- Disk portability: Removable or fixed
- Sided: Single or double (usually) sided
- Platter: Single or multiple platter
- Head mechanism
  - Contact (Floppy)
  - Fixed gap
  - Flying (Winchester)

Fixed/Movable Head Disk

- Fixed head
  - One read write head per track
  - Heads mounted on fixed ridged arm
- Movable head
  - One read write head per side
  - Mounted on a movable arm

Removable or Not

Removable disk

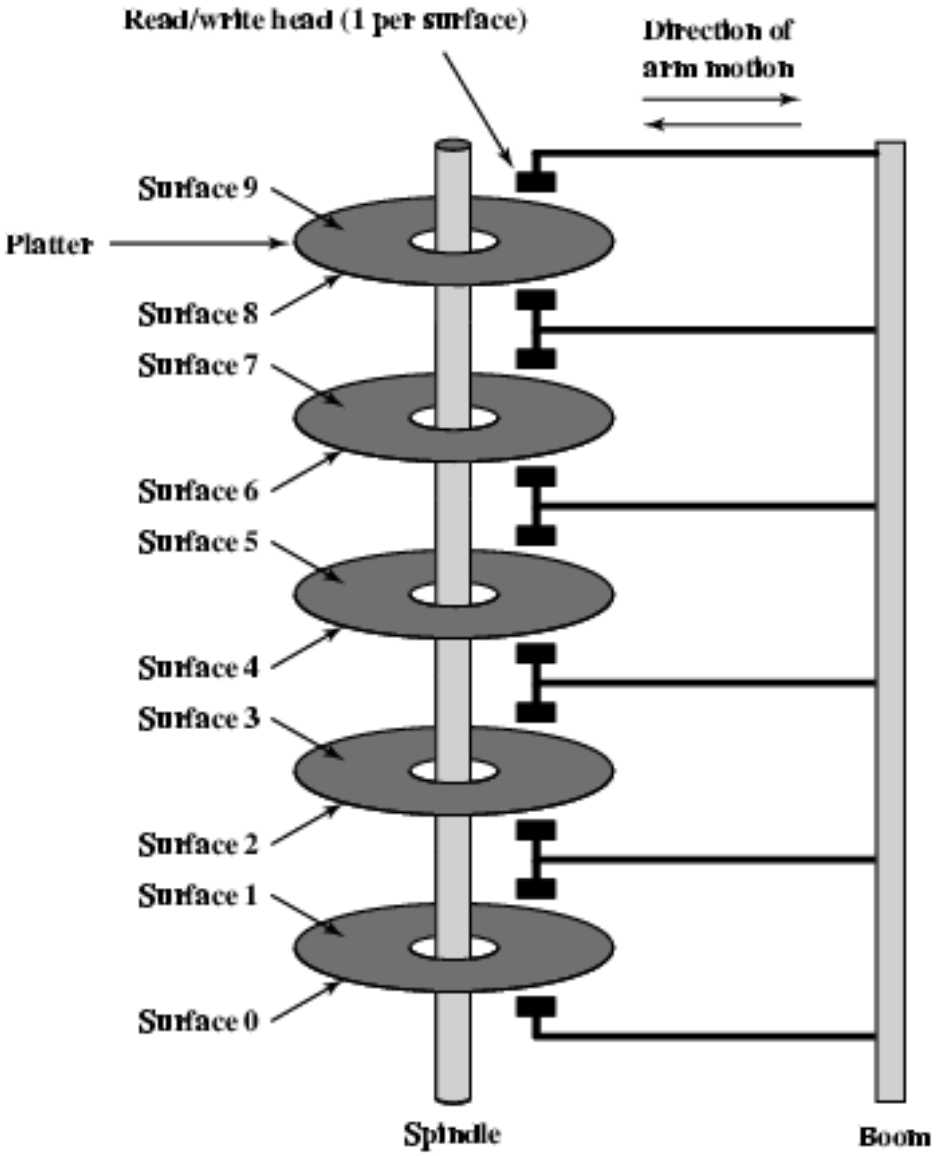
- Can be removed from drive and replaced with another disk
- Provides unlimited storage capacity
- Easy data transfer between systems

Nonremovable disk

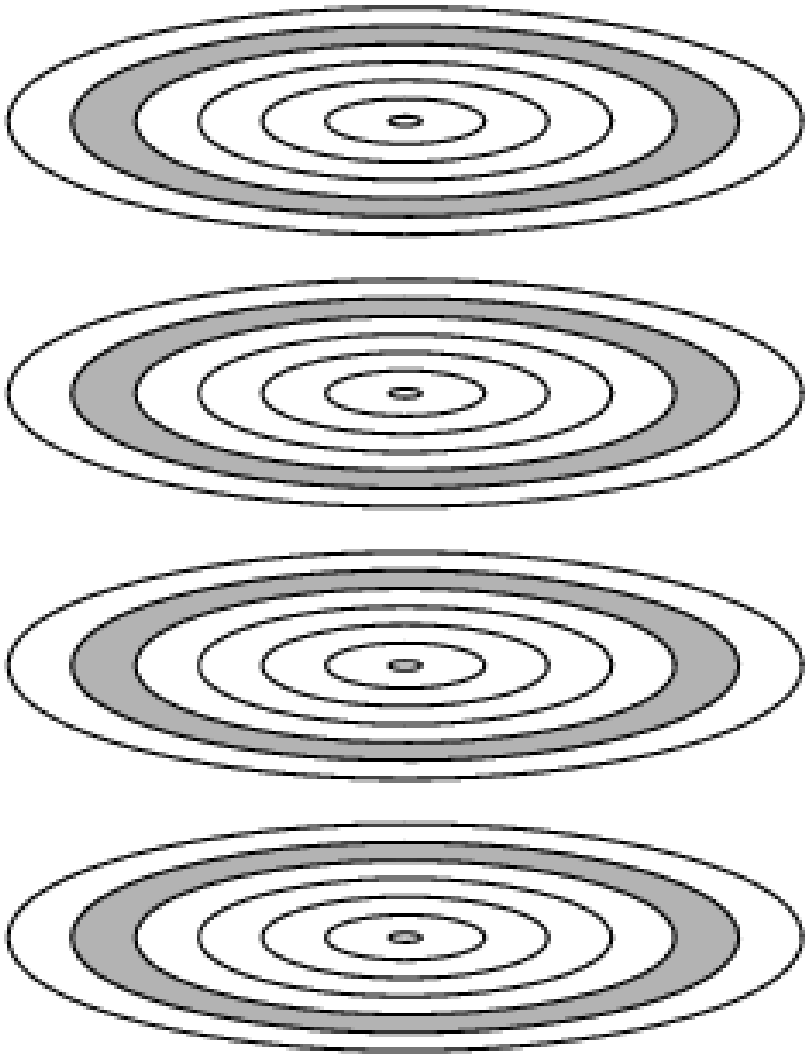
- Permanently mounted in the drive

Multiple Platter

- One head per side
- Heads are joined and aligne
- Aligned tracks on each platte form cylinders
- Cylinder : the set of all the tracks in the same relative position on the platter is referred to as a cylinder



Tracks and Cylinders



Floppy Disk

- 8”, 5.25”, 3.5”
- Small capacity
  - Up to 1.44Mbyte (2.88M never popular)
- Slow
- Cheap
- Obsolete?

Winchester Hard Disk (1)

- Relationship between data density, the size of the air gap and size of the head
  - Narrower head ➔ narrower track ➔ greater data density
  - can not keep head very close to platter
- Winchester Hard Disk:
  - Developed by IBM in Winchester (USA)
  - Sealed unit
  - One or more platters (disks)
  - head is an aerodynamic foil rests lightly on the platter
  - Heads fly on boundary layer of air as disk spins
  - Very small head to disk gap
  - Getting more robust

Capacity

- Capacity: maximum number of bits that can be stored.
- Vendors express capacity in units of gigabytes (GB), where 1 GB =10^9 Byte
- Capacity is determined by these technology factors:
  - Recording density (bits/in): number of bits that can be squeezed into a 1 inch segment of a track.
  - Track density (tracks/in): number of tracks that can be squeezed into a 1 inch radial segment.
  - Areal density (bits/in2): product of recording and track density

Contd…

- Modern disks partition tracks into disjoint subsets called recording zones
  - Each track in a zone has the same number of sectors, determined by the circumference of innermost track.
  - Each zone has a different number of sectors/track

Computing disk capacity

- Capacity =(# bytes/sector) x (Avg # sectors/track) x (# tracks/surface) x (# surfaces/platter) x (# platters/disk)
- Example:
  - 512 bytes/sector
  - 300 sectors/track (on average)
  - 20,000 tracks/surface
  - 2 surfaces/platter
  - 5 platters/disk
  - Capacity = 512 x 300 x 20000 x 2 x 5 = 30.72GB

Disk Performance Parameters

- Seek time
  - Moving head to correct track
- Rotational latency/ delay
  - the time necessary for the desired sector to rotate to the disk head
- Access time = Seek + Latency
- Transfer rate: is the rate at which data flow between the drive and the computer

Contd…

- Transfer time (T=b/rN)
- Total average access time

  $T_a = T_s + T_r + T_t$

  Here
  - T_s is Average seek time
  - T_r is the Average rotational time = 1/2r
  - T_t is the transfer time = b/rN
  - r is rotation speed in revolution per second
  - b number of bytes to be transferred
  - N number of bytes on a track

Example

- Average seek time=4ms
- Rotation speed= 15,000 rpm
- 512 bytes per sector
- No. of sectors per track=500
- Want to read a file consisting of 2500 sectors.
- Calculate the time to read the entire file

  A) File is stored sequentially.

  B) File is stored randomly

Problem

FORMULA

Solution:

r = 15000 /60
1/2r = 2ms

A) File is stored sequentially: on adjacent tracks

Total average access time = $T_a = T_s + 1/2r + b/rN$

Time required to read first track:

Average Seek time	= 4ms
Average rotational delay	= 2ms
Read 500 sectors	= 4ms
Total	10ms

500x512
---
250 x500x512

To read remaining tracks , seek time is very very small.

Average rotational delay	= 2ms
Read 500 sectors	= 4ms
Total	6ms

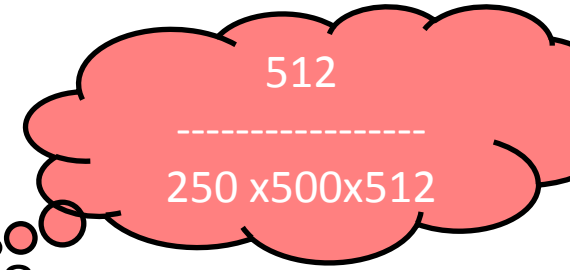
To read entire file : 10 x 1 + 6 x 4 = 34 ms

Solution…

b) File is stored Randomly: Sectors are randomly distributed over the disk

Time required to read each sector:

Average Seek time	= 4ms
Average rotational delay	= 2ms
Read 1 sector	= 0.008ms
Total	6.008ms



To read entire file : 2500x6.008= 15.02 seconds

RAID

- Main draw back of Magnetic disks :provide a single, serial bit stream during data transfer
- Main focus: decreasing access time, increasing bandwidth during data flow and reliability
  - achieved using multiple disks operating independently and parallel
- RAID: storage technology that combines multiple disk drive components into a logical unit.
  - Redundant Array of Independent Disks or Redundant Array of Inexpensive Disks
  - Set of physical disks viewed as single logical drive by O/S
- Data distributed across physical drives

Contd…

- Main advantage:
  - Performance improvement due to parallelism by multiple disks
  - fault tolerance due to replication of data
- 7 levels in common use: Zero through six
  - defines different disk organization techniques

Three main characteristics

1. RAID is a set of physical disk drives viewed by the operating system as a single logical drive.

2. Data are distributed across the physical drives of an array in a scheme known as striping

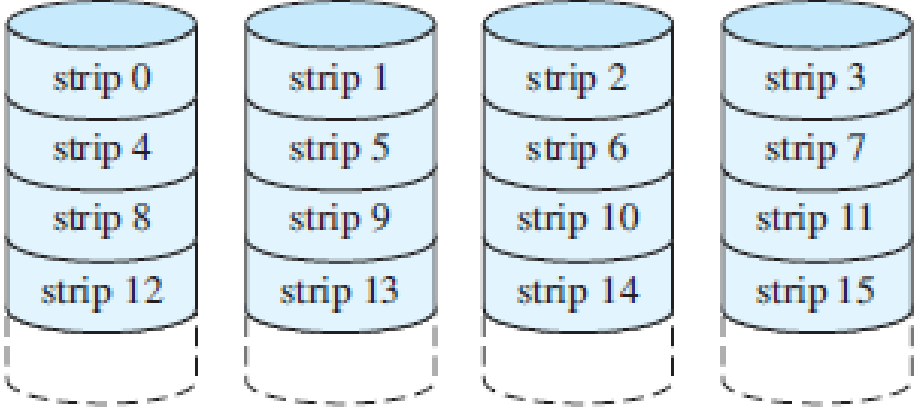
3. Redundant disk capacity is used to store parity information, which guarantees data recoverability in case of a disk failure.

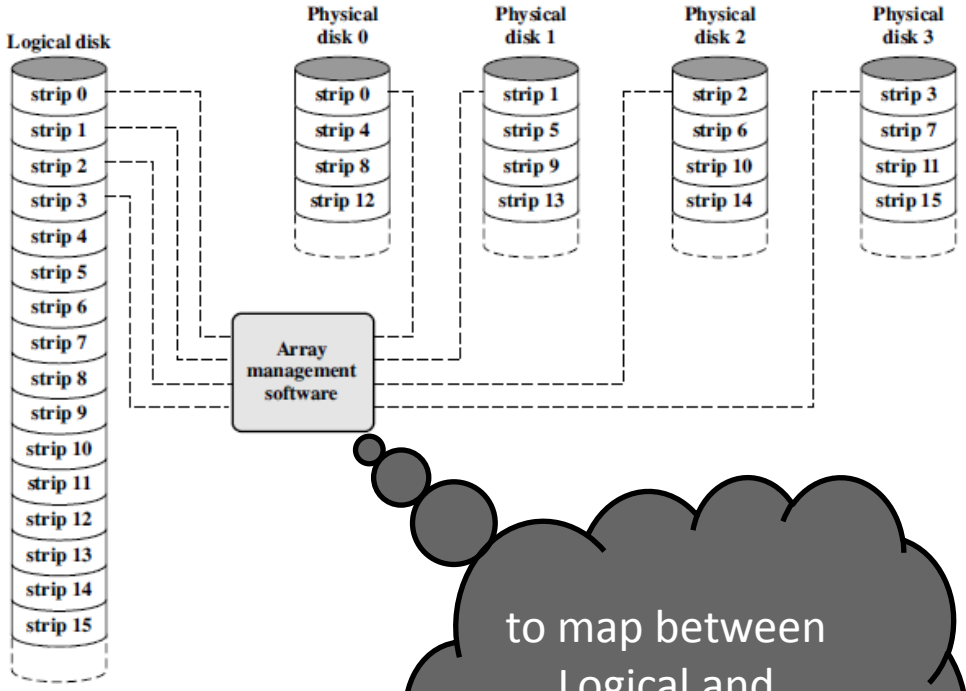
Striping

- Bit level striping: Splitting the bits of each byte across multiple disks
- Block level striping: Blocks of a file striped across multiple disks
- Advantage : Improvement in performance via parallelism

RAID 0

- Not a true member of RAID
  - No redundancy
- Data striped across all disks
- Round Robin striping
- Increase in speed
  - Multiple data requests probably not on same disk
  - Disks seek in parallel
  - Parallel access to N disks





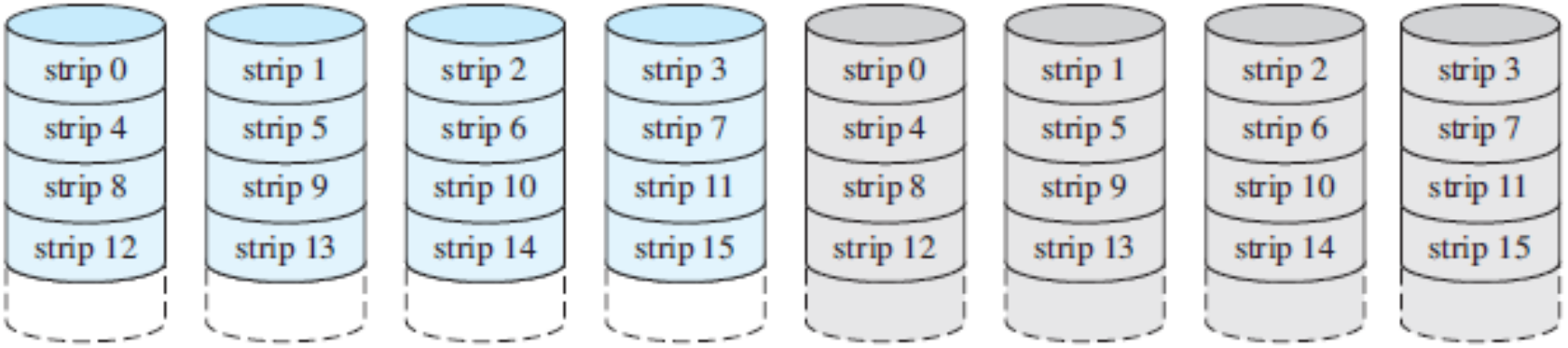
to map between Logical and physical disk space

RAID 0

- High data transfer capacity : 2 requirements
  - a high transfer capacity must exist along the entire path between host memory and the individual disk drives.
    - internal controller buses, host system I/O buses, I/O adapters, and host memory buses.
- the application must make I/O requests that drive the disk array efficiently
  - usually large contiguous data

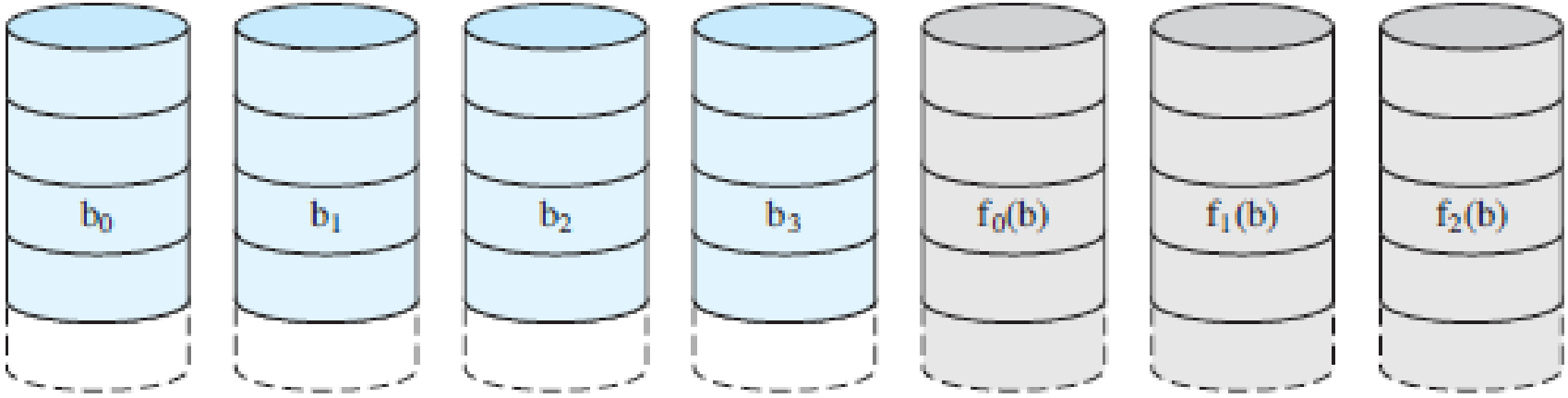
RAID 1

- Mirrored Disks
- Data is striped across disks
- 2 copies of each stripe on separate disks
- Read from either
- Write to both
- Recovery is simple
  - replace faulty disk & re-mirror
- Expensive
  - limited to drives that store system software other highly critical files



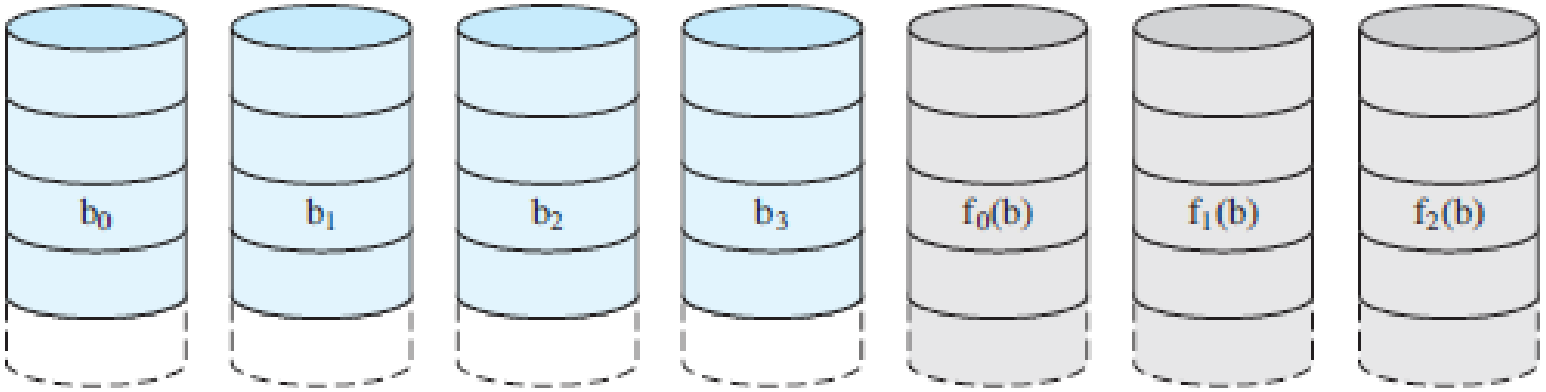
RAID 2

- Also known as memory style error correcting code organization (ECC)
- Provides parallel access
- Disks are synchronized
  - Spindles of individual disks are synchronized
- Very small stripes
  - Often single byte/word



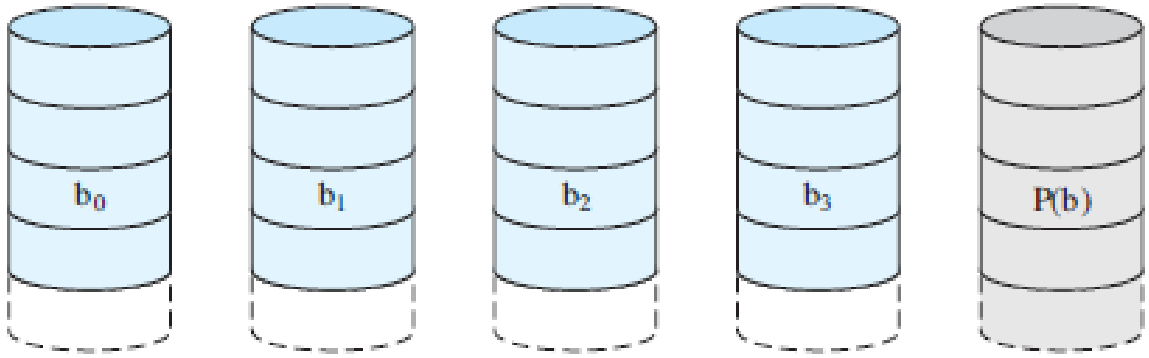
Contd…

- Error correction calculated across corresponding bits on each disk➔ Hamming code
  - correction → single bit, detect → double – bit errors
- Multiple parity disks store Hamming code error correction in corresponding positions
- Lots of redundancy
  - Expensive
  - Not used



RAID 3

- Similar to RAID 2
- Also known as bit interleaved parity organization
- Only one redundant disk, no matter how large the array
- Simple parity bit for each set of corresponding bits
- Data on failed drive can be reconstructed from surviving data and parity info
- Very high transfer rates as RAID 0
- Reduced storage overhead



RAID 3…

Consider 5 drives : X0 to X4

X0 – X3 ➔ data

X4 ➔ Parity

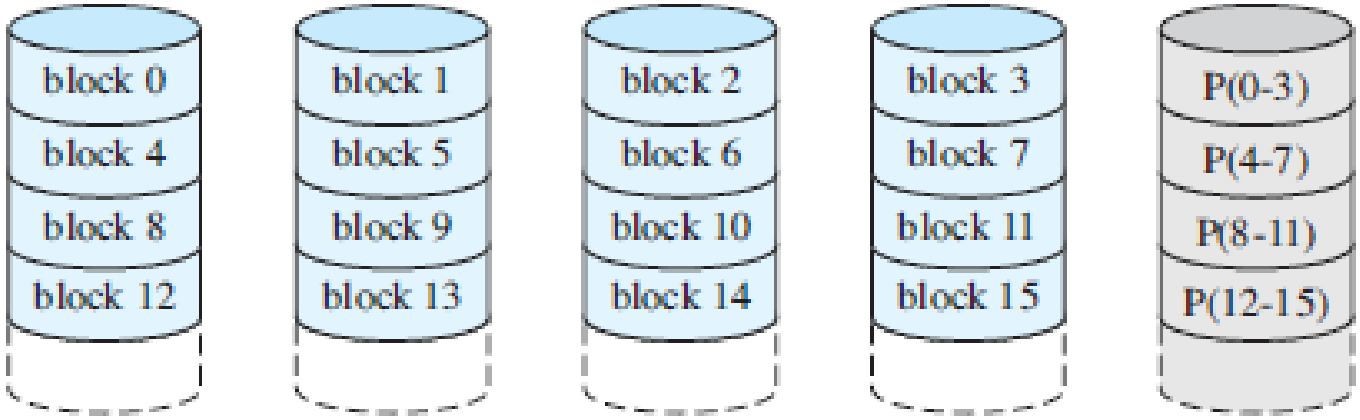
$$X4(i) = X3(i) \oplus X2(i) \oplus X1(i) \oplus X0(i)$$

If X1 fails

$$X1(i) = X3(i) \oplus X2(i) \oplus X0(i) \oplus X4(i)$$

RAID 4

- Also known as “Block interleaved parity organization”
- Each disk operates independently
- Good for application with high I/O request rate
- Large stripes
- Bit by bit parity calculated across stripes on each disk
- Parity stored on parity disk



RAID 4…

- Write penalty: update user data as well as corresponding parity bits

$$X4(i) = X3(i) \oplus X2(i) \oplus X1(i) \oplus X0(i)$$

- Update in X1

$$X4'(i) = X3(i) \oplus X2(i) \oplus X1'(i) \oplus X0(i)$$

$$X4'(i) = X3(i) \oplus X2(i) \oplus X1'(i) \oplus X0(i) \oplus X1(i) \oplus X1(i)$$

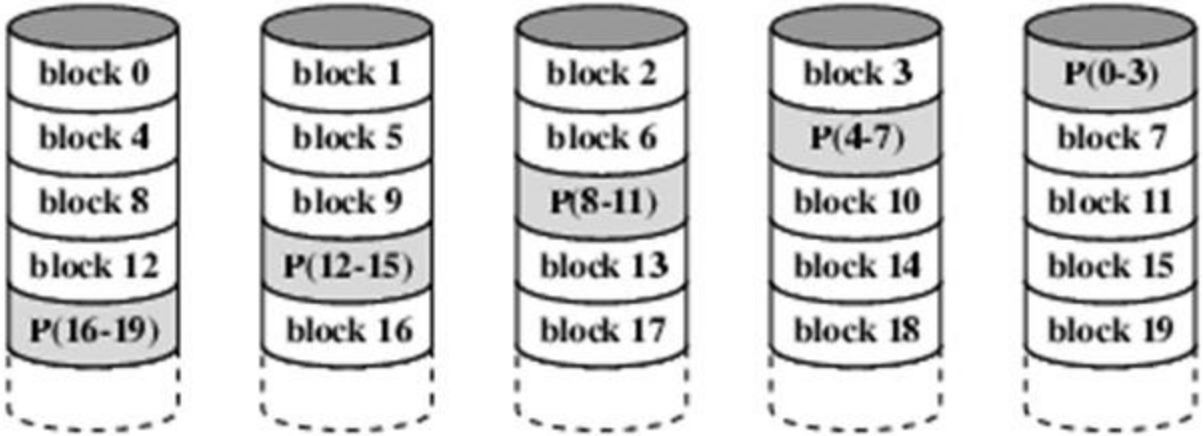
$$X4'(i) = X3(i) \oplus X2(i) \oplus X1(i) \oplus X0(i) \oplus X1'(i) \oplus X1(i)$$

$$X4'(i) = X4(i) \oplus X1'(i) \oplus X1(i)$$

- each strip write involves two reads and two writes.

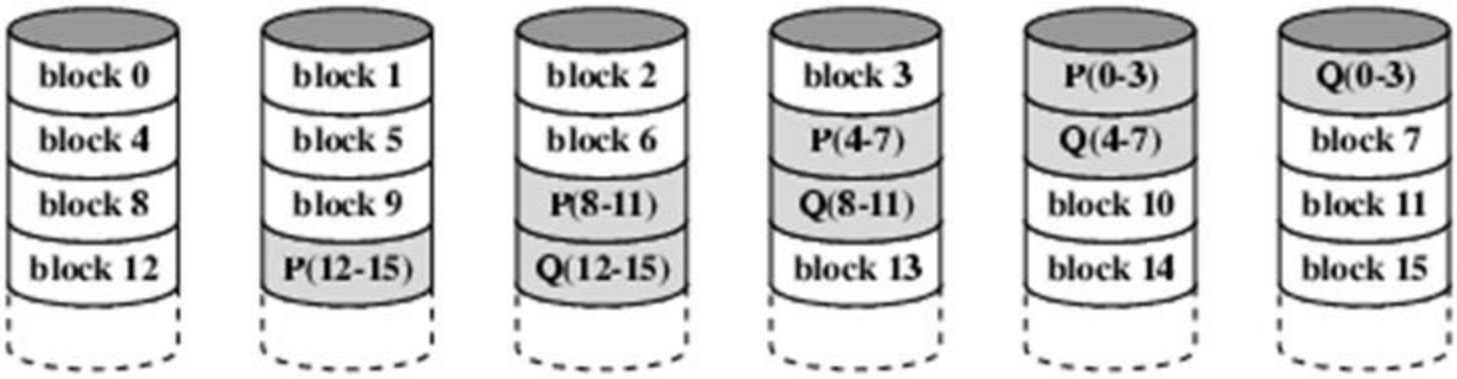
RAID 5

- Like RAID 4
- Parity striped across all disks
- Also known as “ Block-interleaved distributed parity”
- Round robin allocation for parity stripe
- Avoids RAID 4 bottleneck at parity disk
- Commonly used in network servers



RAID 6

- Also known as P+Q redundancy scheme
- Two parity calculations
  - provides extra redundant info to guard against multiple disk failures
  - Reed Solomon code instead of parity
- Stored in separate blocks on different disks
- User requirement of N disks needs N+2
- High data availability
  - Three disks need to fail for data loss
  - Significant write penalty



Category	Level	Description	Disks required	Data availability	Large I/O data transfer capacity	Small I/O request rate
Striping	0	Nonredundant	N	Lower than single disk	Very high	Very high for both read and write
Mirroring	1	Mirrored	$2N$	Higher than RAID 2, 3, 4, or 5; lower than RAID 6	Higher than single disk for read; similar to single disk for write	Up to twice that of a single disk for read; similar to single disk for write
Parallel access	2	Redundant via Hamming code	$N + m$	Much higher than single disk; comparable to RAID 3, 4, or 5	Highest of all listed alternatives	Approximately twice that of a single disk
	3	Bit-interleaved parity	$N + 1$	Much higher than single disk; comparable to RAID 2, 4, or 5	Highest of all listed alternatives	Approximately twice that of a single disk
Independent access	4	Block-interleaved parity	$N + 1$	Much higher than single disk; comparable to RAID 2, 3, or 5	Similar to RAID 0 for read; significantly lower than single disk for write	Similar to RAID 0 for read; significantly lower than single disk for write
	5	Block-interleaved distributed parity	$N + 1$	Much higher than single disk; comparable to RAID 2, 3, or 4	Similar to RAID 0 for read; lower than single disk for write	Similar to RAID 0 for read; generally lower than single disk for write
	6	Block-interleaved dual distributed parity	$N + 2$	Highest of all listed alternatives	Similar to RAID 0 for read; lower than RAID 5 for write	Similar to RAID 0 for read; significantly lower than RAID 5 for write