

Hardware

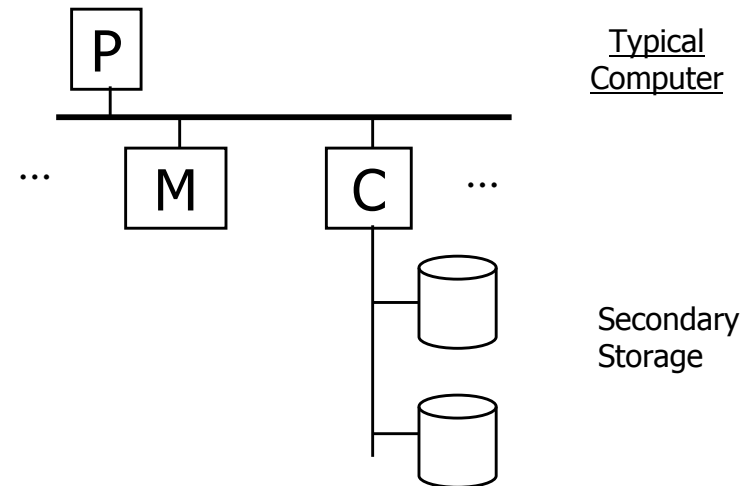
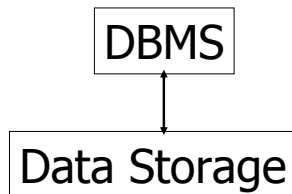
Read Sections 11.2, 11.3, 11.7
of Garcia-Molina et al.

Slides derived from those by Hector Garcia-Molina
Some images by Wikipedia

Outline

- Hardware: Disks
- Access Times
- Example - Megatron 747
- Reliability
- RAID

Hardware



Processor

Fast, slow, reduced instruction set,
with cache, pipelined...

Speed: 1000 → 10000 MIPS

Memory

Fast, slow, non-volatile, read-only,...

Access time: 10^{-6} → 10^{-9} sec.

1 μ s → 1 ns

5

Secondary storage

Hard Disks

Tertiary storage

Optical disks:

- CD-ROM

- DVD-ROM...

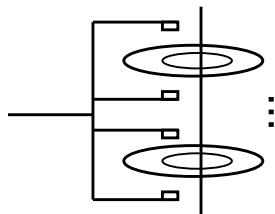
Tape

- Cartridges

Robots

6

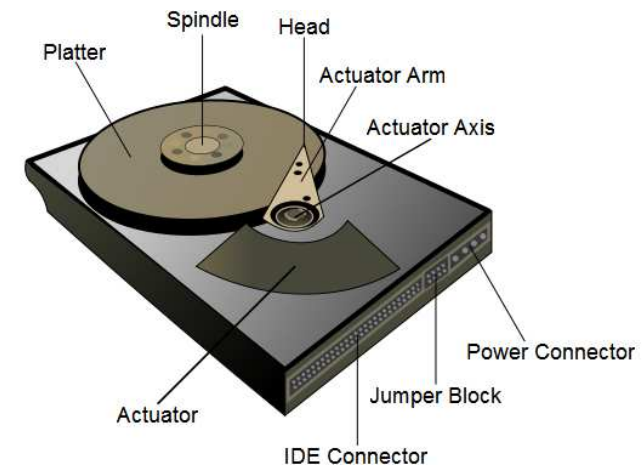
Focus on: "Typical Disk"



Terms: Platter, Surface, Head, Actuator
Cylinder, Track
Sector (physical),
Block (logical), Gap

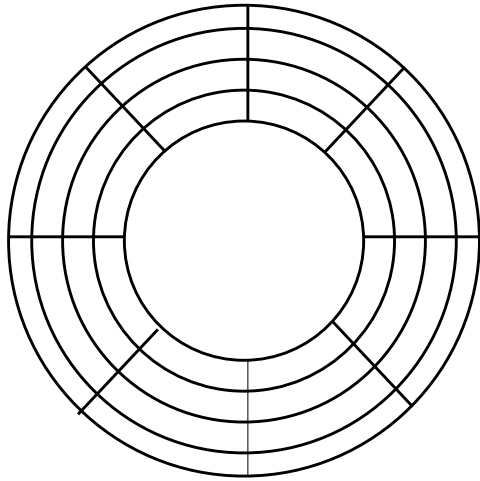
7

Disk Architecture



8

Top View



9

"Typical" Numbers

Diameter: 1 inch → 15 inches
(1 inch=2.54 cm:

2.5 cm → 38.1 cm)

Cylinders: 10000 → 50000

Surfaces: 2 → 30

(Tracks/cyl)

Sector Size: 512B → 50KB

Capacity: 72 GB → 2TB

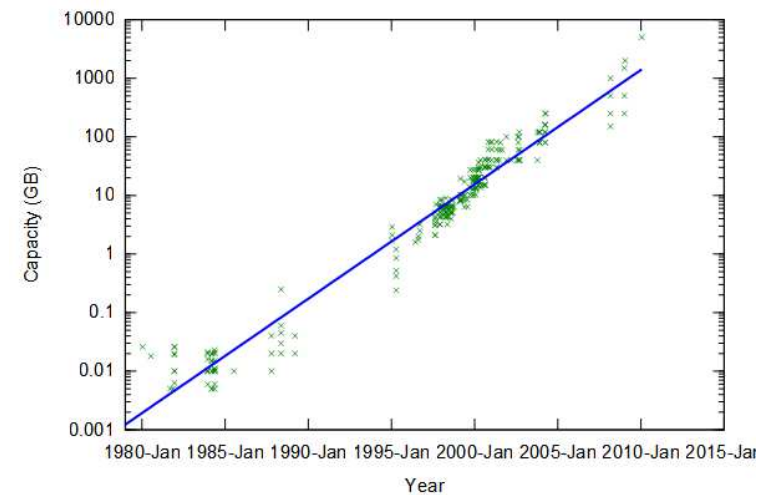
10

Diameter

- Form factors:
 - 8 inches
 - 5.25 inches
 - 3,5 inches
 - 2,5 inches
 - 1,8 inches
 - 1 inch

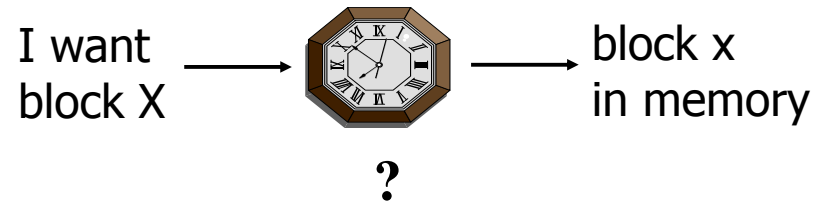
11

Capacity



12

Disk Access Time

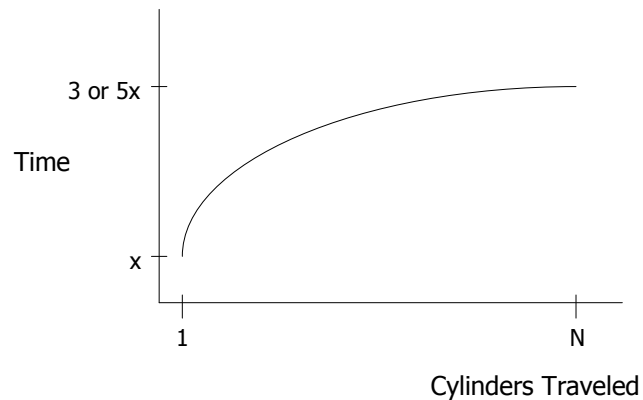


13

$$\text{Time} = \text{Seek Time} + \text{Rotational Delay} + \text{Transfer Time} + \text{Other}$$

14

Seek Time



15

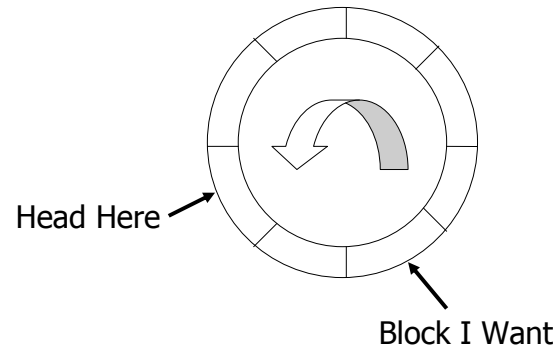
Average Random Seek Time

$$S = \frac{\sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N \text{SEEKTIME}(i \rightarrow j)}{N(N-1)}$$

“Typical” S: 3 ms → 10 ms

16

Rotational Delay



17

Average Rotational Delay

$R = 1/2$ revolution

"typical" $R = 4.17$ ms (7200 RPM)

$R = 3$ ms (10000 RPM)

$R = 2$ ms (15000 RPM)

18

Transfer Rate: t

- "typical" t : 60 MB/second
- transfer time: $\frac{\text{block size}}{t}$

19

Other Delays

- CPU time to issue I/O
- Contention for controller
- Contention for bus, memory


"Typical" Value: 0

20

- So far: Random Block Access
- What about: Reading "Next" block?

21

$$\text{Time to get block} = \frac{\text{Block Size}}{t} + \text{Negligible}$$



- skip gap

22

Cost for Writing similar to Reading

.... unless we want to verify!
 need to add (full) rotation + $\frac{\text{Block size}}{t}$

23

- To Modify a Block?

To Modify Block:

- (a) Read Block
- (b) Modify in Memory
- (c) Write Block
- [(d) Verify?]

24

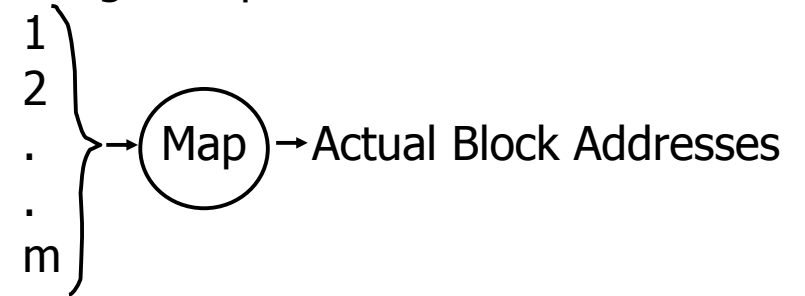
Block Address:

- Physical Device
- Cylinder (Track) #
- Surface #
- Sector

25

Complication: Bad Blocks

- Messy to handle
- May map via software to integer sequence



26

An Example

Megatron 747 Disk

- 3.5 in diameter
- 8 platters, 16 surfaces
- $2^{14}=16,384$ tracks per surface (16,384 cylinders)
- $2^7=128$ sectors per track
- $2^{12}=4096$ bytes per sector
- Capacity
 - Disk= $2^4*2^{14}*2^7*2^{12}=2^{37}=128\text{GB}$
 - Single track= $2^7*2^{12}=512\text{KB}$

27

Megatron 747 Disk

- Rotation speed: 7200 RPM
- Average seek time: 8.5 ms

28

Layout

- Radius: 1.75 inches
- The tracks occupy the outer inch
- The inner 0.75 inch is unoccupied
- Track density in the radial direction: 16,384 tracks per inch
- 10% overhead between blocks

29

Density of bits

- Outermost track
 - Length= $3.5\pi \approx 11$ inches
 - One track = 512KB = 4Mbits
 - 90% of 11 inches holds 4Mbits
 - Density=420,000 bits per inch
- Innermost track
 - 90% of 4.71 inches holds 4Mbits
 - Density \approx 1Mbit per inch

30

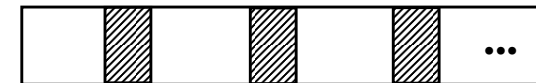
Density of bits

- To avoid such a high difference of density, the disk stores more sectors on the outer track than on the inner tracks
 - 96 sectors per track in the inner third
 - 128 in the middle third
 - 160 in the outer third
- The density varies from 742,000 bits per inch to 530,000 bits per inch

31

7200 RPM \rightarrow 120 revolutions / sec
 \rightarrow 1 rev. = 8.33 msec.

One track:



Time over useful data: $(8.33)(0.9) = 7.5$ ms.
Time over gaps: $(8.33)(0.1) = 0.833$ ms.
Transfer time 1 sector = $7.5/128 = 0.059$ ms.
Trans. time 1 sector+gap = $8.33/128 = 0.065$ ms.

32

Burst Bandwidth

4 KB in 0.059 ms.

$$BB = 4/0.059 = 68 \text{ KB/ms.}$$

or

$$\begin{aligned} BB &= 68 \text{ KB/ms} \times 1000 \text{ ms/1sec} \\ &\quad \times 1\text{MB}/1024\text{KB} \\ &= 68,000/1024 = 66.4 \text{ MB/sec} \end{aligned}$$

33

Sustained bandwidth (over track)

512 KB in 8.33 ms.

$$SB = 512/8.33 = 61.5 \text{ KB/ms}$$

or

$$SB = 61.5 \times 1000/1024 = 60 \text{ MB/sec.}$$

34

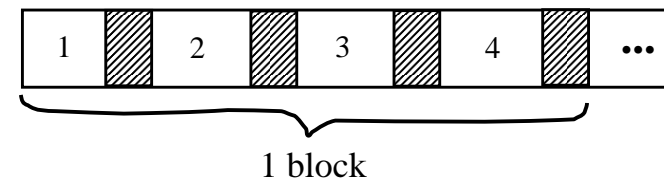
T_1 = Time to read one random block

$$T_1 = \text{seek} + \text{rotational delay} + TT$$

$$= 8.5 + (8.33/2) + 0.059 = 12.72 \text{ ms.}$$

35

Suppose OS deals with 16 KB blocks



$$\begin{aligned} T_4 &= 8.5 + (8.33/2) + 0.059 \times 1 + (0.065) \\ &\quad \times 3 = 12.92 \text{ ms} \end{aligned}$$

[Compare to $T_1 = 12.72 \text{ ms}$]

36

T_T = Time to read a full track
(start at any block)

$$T_T = 8.5 + (0.065/2) + 8.33^* = 16.86 \text{ ms}$$

↑
to get to first block

* Actually, a bit less; do not have to read last gap.

37

Block Size Selection?

- Big Block → Amortize I/O Cost



- Big Block ⇒ Read in more useless stuff!
and takes longer to read

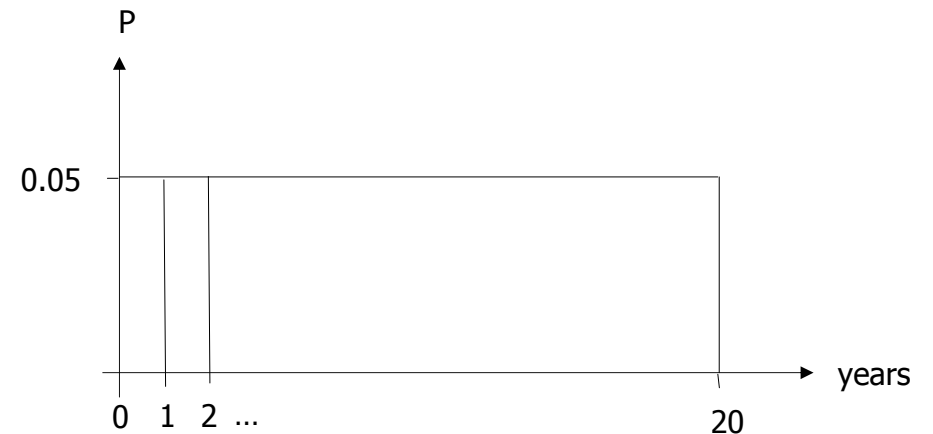
38

Reliability

- Measured by the Mean Time to Failure (MTTF):
 - Length of time by which 50% of a population of disks will have failed catastrophically (head crash, no longer readable)
 - For modern disks, the MTTF is 10 years
 - This means that, on average, after 10 years it will crash
 - We can assume that every year 5% of the disks fail (uniform distribution assumption)
 - Probability that a disk fails in one year $P_F=5\%=1/20$

39

Probability of Failure



40

MTTF

- Expected value of the failure year:
- $MTTF = E(\text{year}) =$
 $= 0.05 * 1 + \dots + 0.05 * 20 =$
 $= 0.05 * 20 * (20 + 1) / 2 = 21 / 2 \approx 10$

41

Disk Arrays

- Redundant Arrays of Inexpensive Disks (RAID)
- Two aims: increase speed and reliability

42

RAID 0

- Uses "block level striping"
 - Blocks that are consecutive for the OS are distributed evenly across different disks

RAID 0

A1 A2 consecutive blocks: A1-A8
A3 A4
A5 A6
A7 A8

43

RAID 0

- Improves reading and writing speed
 - With two disks, two blocks can be read at the same time
 - A request for block "A1" would be serviced by disk 1. A simultaneous request for block A3 would have to wait, but a request for A2 could be serviced concurrently
- Reduces reliability: if one disk fails, the data is lost.

44

RAID 0

- $P(\text{data loss}) = P(\text{disk1 fails or disk2 fails}) =$
 $= P(\text{disk1 fails}) + P(\text{disk2 fails}) - P(\text{disk1 fails and disk2 fails}) =$
 $= P_F + P_F - P_F * P_F = 2P_F - P_F^2 =$
 $= 2 * 0.05 - 0.0025 = 0.0975$

45

RAID 0

- Number of years = $1 / 0.0975 \approx 10$
- $MTTF = E(\text{year}) =$
 $\approx 0.0975 * 10 * (10 + 1) / 2 \approx 11 / 2 \approx 5.5$

46

RAID 1

- Creates an exact copy (or **mirror**) of a set of data on two or more disks.
- Typically, a RAID 1 array contains two disks
- Improved
 - Reading speed: two blocks can be read at the same time
 - Reliability: if one disk crashes, we can use the other
- Writing speed remains the same

47

RAID 1

RAID 1

A1	A1
A2	A2
A3	A3
A4	A4

48

RAID 1 Reliability

- Two disks with MTTF of 10 years
- What is the MTTF resulting in data loss?
- Data loss happens when one disk fails and the other fails as well while we are replacing the first.
- Supposing it takes 3 hours to replace the first disk. This is $1/2920$ of a year
- $P(\text{fails rep}) = 1/2920 = 3.42E-04$

49

RAID 1 Reliability

- The probability that the second disk fails while replacing the first is
 $P(\text{fails1 and fails2 rep}) =$
 $= 5E-2 * 5E-2 * 3.42E-04 = 8.55E-07$
- $P(\text{data loss}) = P(\text{fails1 and fails2 rep or fails2 and fails1 rep}) =$

50

RAID 1 Reliability

$$\begin{aligned} &= P(\text{fails1 and fails2 rep}) + P(\text{fails2 and fails1 rep}) - P(\text{fails2 and fails1 rep and fails1 and fails2 rep}) = \\ &\approx 2 * 8.55E-07 \\ &= 1.71E-06 \end{aligned}$$

51

RAID 1 Reliability

- Number of years $= 1/1.71E-06 \approx 584795$
- $MTTF = E(\text{years}) =$
 $= 1.71E-06 * 584795 * 584796 / 2 =$
 $= 584796 / 2 = 292398$

52

RAID 4

- Uses block-level striping with a dedicated parity disk.

RAID 4

A1	A2	A3	Ap	Consecutive blocks
B1	B2	B3	Bp	A1-A3, B1-B3,
C1	C2	C3	Cp	C1-C3, D1-D3
D1	D2	D3	Dp	

53

Parity block

- Bit i of the block in position j on the parity disk is the parity bit of the bits in position i in the blocks in position j in the other disks
- Eg., blocks of one byte, blocks A1-A3
Disk1 11110000
Disk2 10101010
Disk3 00111000
Disk4 01100010 (parity disk)

54

RAID 4

- Improves reading time: multiple blocks can be read at the same time
- Improves reliability: if one disk fails, we can reconstruct its content (assuming the others are correct)

55

RAID 4

- Problem:
 - When writing a block, we need to read and write the parity disk's block
 - This creates a bottleneck

56

RAID 5

- Uses block-level striping with parity data distributed across all member disks.

RAID 5

A1 A2 A3 Ap

B1 B2 Bp B3

C1 Cp C2 C3

Dp D1 D2 D3

57

RAID 5

- Reading and reliability as RAID 4
- Writing improved because the parity blocks are not all on one disk

58

RAID 6

- Uses block-level striping with dual parity data distributed across all member disks.

RAID 6

A1 A2 A3 Ap Aq

B1 B2 Bp Bq B3

C1 Cp Cq C2 C3

Dp Dq D1 D2 D3

59

RAID 6

- p and q blocks are computed with two different algorithms, e.g.
 - parity and Reed-Solomon
 - orthogonal dual parity
 - diagonal parity

60

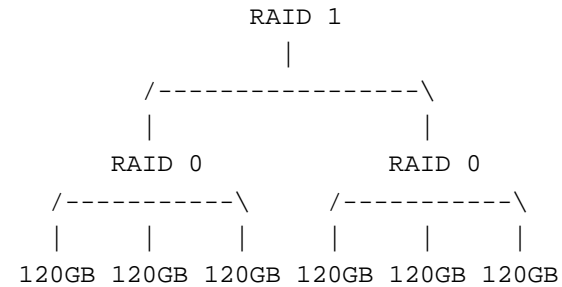
RAID 6

- It is able to recover from the loss of two disks
- Writing improved because the parity blocks are not all on one disk

61

Nested RAID Levels

- RAID 0+1:



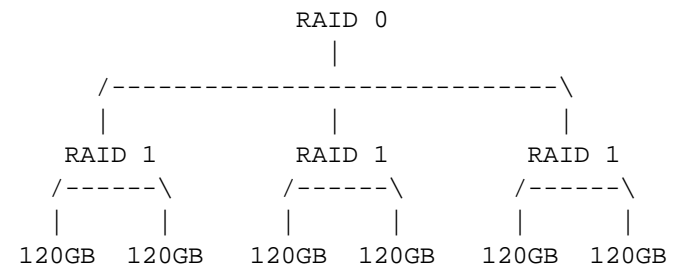
62

RAID 0+1

- If a disk fails, it can be rebuilt from the corresponding disk in the other RAID 0 batch
- If two disk fails from the same stripe, no recovery

63

RAID 1+0 o RAID 10



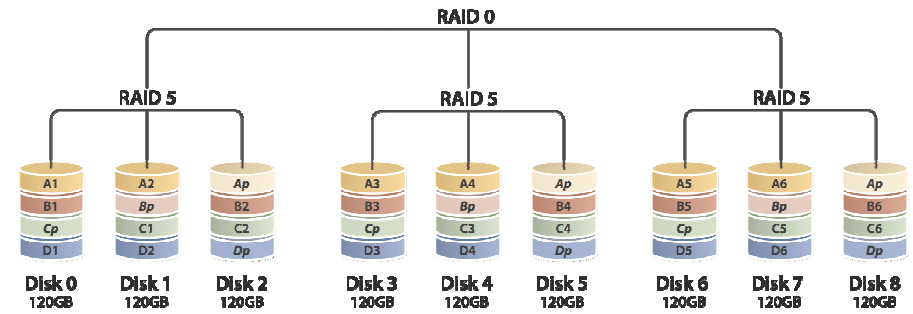
64

RAID 1+0 o RAID 10

- If a disk fails, it can be rebuilt from the corresponding disk in the other RAID 1 batches
- If two disk fails from the same RAID 1 batch, no recovery

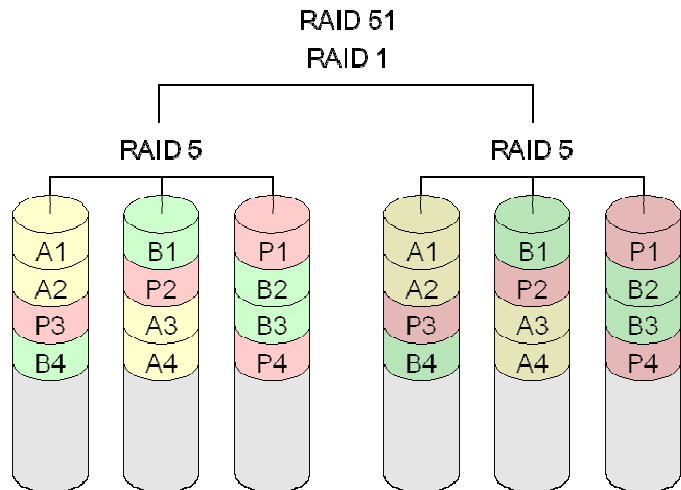
65

RAID 5+0



66

RAID 5+1



67