




CS 525: Advanced Database Organization

02: Hardware





Boris Glavic

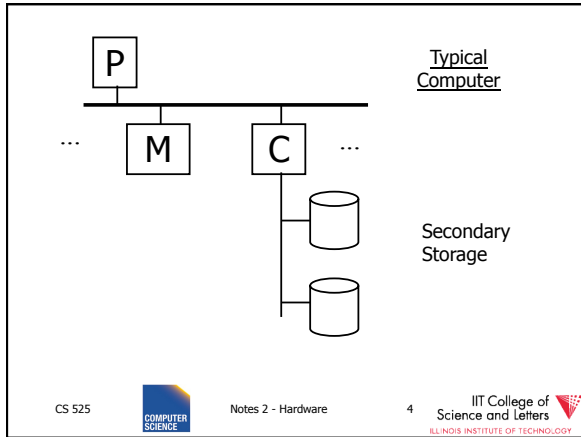
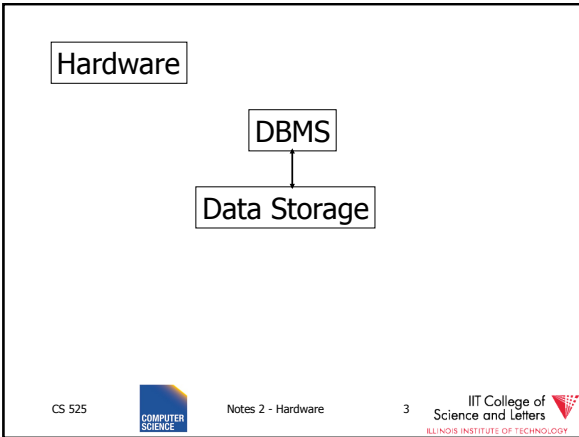
Slides: adapted from a [course](#) taught by [Hector Garcia-Molina](#), Stanford InfoLab

CS 525

Notes 2 - Hardware
1


Outline

- Hardware: Disks
- Access Times
- Example - Megatron 747
- Optimizations
- Other Topics:
 - Storage costs
 - Using secondary storage
 - Disk failures

CS 525

Notes 2 - Hardware
2




Processor

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



Speed: 100 → 500 → 1000 MIPS

Memory

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

Access time: 10^{-6} → 10^{-9} sec.



$1 \mu\text{s}$ → 1 ns

CS 525

Notes 2 - Hardware
5


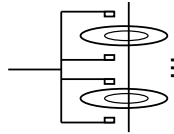
Secondary storage

Many flavors:

- Disk: Floppy (hard, soft)
Removable Packs
Winchester
Ram disks
Optical, CD-ROM...
Arrays
- Tape: Reel, cartridge
Robots

CS 525

Notes 2 - Hardware
6


Focus on: "Typical Disk"



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

CS 525

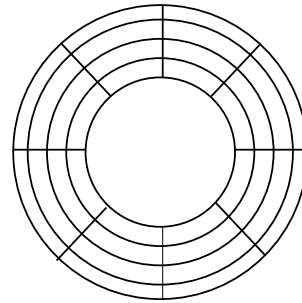


Notes 2 - Hardware

7

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Top View



CS 525



Notes 2 - Hardware

8

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

"Typical" Numbers

Diameter: 1 inch \rightarrow 15 inches
Cylinders: 100 \rightarrow 2000
Surfaces: 1 (CDs) \rightarrow
(Tracks/cyl) 2 (floppies) \rightarrow 30
Sector Size: 512B \rightarrow 50K
Capacity: 360 KB (old floppy)
 \rightarrow 1 TB (I use)

CS 525

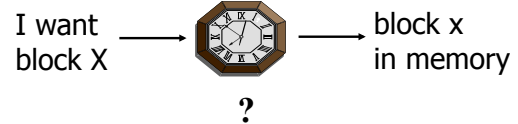


Notes 2 - Hardware

9

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Disk Access Time



CS 525



Notes 2 - Hardware

10

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Time = Seek Time +
Rotational Delay +
Transfer Time +
Other

CS 525

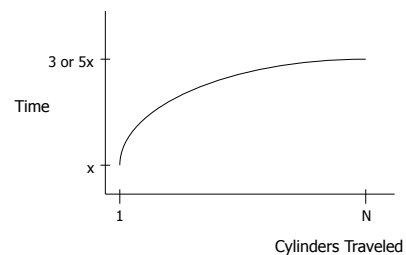


Notes 2 - Hardware

11

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Seek Time



CS 525



Notes 2 - Hardware

12

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Average Random Seek Time

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

CS 525



Notes 2 - Hardware

13

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

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: 10 ms → 40 ms

CS 525

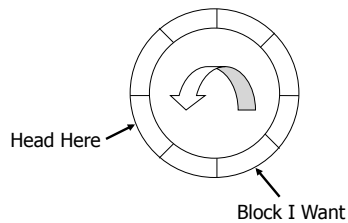


Notes 2 - Hardware

14

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Rotational Delay



CS 525



Notes 2 - Hardware

15

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Average Rotational Delay

$R = 1/2$ revolution

“typical” R = 8.33 ms (3600 RPM)

CS 525



Notes 2 - Hardware

16

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Transfer Rate: t

- “typical” t: 10’s → 100’s MB/second
- transfer time: $\frac{\text{block size}}{t}$

CS 525



Notes 2 - Hardware

17

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Other Delays

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

CS 525



Notes 2 - Hardware

18

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Other Delays

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

“Typical” Value: 0

CS 525



Notes 2 - Hardware

19

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Other Delays (now and near future)

- Increasing amount of parallelism
- Contention can become a problem
- -> need rethink approach to scale

CS 525



Notes 2 - Hardware

20

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

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

CS 525



Notes 2 - Hardware

21

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

If we do things right (e.g., Double Buffer,
Stagger
Blocks...)

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

- skip gap
- switch track
- once in a while,
next cylinder

CS 525



Notes 2 - Hardware

22

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Rule of Thumb Random I/O: Expensive
 Sequential I/O: Much less

- Ex: 1 KB Block
 - » Random I/O: ~ 20 ms.
 - » Sequential I/O: ~ 1 ms.

CS 525



Notes 2 - Hardware

23

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Cost for Writing similar to Reading

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

CS 525



Notes 2 - Hardware

24

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

- To Modify a Block?

CS 525



Notes 2 - Hardware

25

- To Modify a Block?

To Modify Block:

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

CS 525



Notes 2 - Hardware

26

Block Address:

- Physical Device
- Cylinder #
- Surface #
- Sector

CS 525

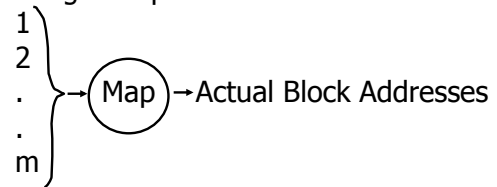


Notes 2 - Hardware

27

Complication: Bad Blocks

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



CS 525



Notes 2 - Hardware

28

An Example Megatron 747 Disk (old)

- 3.5 in diameter
- 3600 RPM
- 1 surface
- 16 MB usable capacity (16×2^{20})
- 128 cylinders
- seek time: average = 25 ms.
adjacent cyl = 5 ms.

CS 525



Notes 2 - Hardware

29

- 1 KB blocks = sectors
- 10% overhead between blocks
- capacity = 16 MB = $(2^{20})16 = 2^{24}$
- # cylinders = $128 = 2^7$
- bytes/cyl = $2^{24}/2^7 = 2^{17} = 128 \text{ KB}$
- blocks/cyl = $128 \text{ KB} / 1 \text{ KB} = 128$

CS 525



Notes 2 - Hardware

30

3600 RPM → 60 revolutions / sec
 → 1 rev. = 16.66 msec.

One track:



CS 525



Notes 2 - Hardware

31

IIT College of Science and Letters
 ILLINOIS INSTITUTE OF TECHNOLOGY

3600 RPM → 60 revolutions / sec
 → 1 rev. = 16.66 msec.

One track:



Time over useful data: $(16.66)(0.9) = 14.99$ ms.
 Time over gaps: $(16.66)(0.1) = 1.66$ ms.
 Transfer time 1 block = $14.99/128 = 0.117$ ms.
 Trans. time 1 block+gap = $16.66/128 = 0.13$ ms.

CS 525



Notes 2 - Hardware

32

IIT College of Science and Letters
 ILLINOIS INSTITUTE OF TECHNOLOGY

Burst Bandwidth

1 KB in 0.117 ms.

$$BB = 1/0.117 = 8.54 \text{ KB/ms.}$$

or

$$BB = 8.54 \text{ KB/ms} \times 1000 \text{ ms/1sec} \times 1 \text{ MB}/1024 \text{ KB} \\ = 8540/1024 = 8.33 \text{ MB/sec}$$

CS 525



Notes 2 - Hardware

33

IIT College of Science and Letters
 ILLINOIS INSTITUTE OF TECHNOLOGY

Sustained bandwidth (over track)

128 KB in 16.66 ms.

$$SB = 128/16.66 = 7.68 \text{ KB/ms}$$

or

$$SB = 7.68 \times 1000/1024 = 7.50 \text{ MB/sec.}$$

CS 525



Notes 2 - Hardware

34

IIT College of Science and Letters
 ILLINOIS INSTITUTE OF TECHNOLOGY

T_1 = Time to read one random block

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

$$= 25 + (16.66/2) + .117 = 33.45 \text{ ms.}$$

CS 525

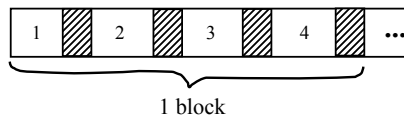


Notes 2 - Hardware

35

IIT College of Science and Letters
 ILLINOIS INSTITUTE OF TECHNOLOGY

Suppose OS deals with 4 KB blocks



$$T_4 = 25 + (16.66/2) + (.117) \times 1 \\ + (.130) \times 3 = 33.83 \text{ ms}$$

[Compare to $T_1 = 33.45$ ms]

CS 525



Notes 2 - Hardware

36

IIT College of Science and Letters
 ILLINOIS INSTITUTE OF TECHNOLOGY

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

$$T_T = 25 + (0.130/2) + 16.66^* = 41.73 \text{ ms}$$

↑
to get to first block

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

CS 525



Notes 2 - Hardware

37

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

The NEW Megatron 747

- 8 Surfaces, 3.5 Inch diameter
 - outer 1 inch used
- $2^{13} = 8192$ Tracks/surface
- 256 Sectors/track
- $2^9 = 512$ Bytes/sector

CS 525

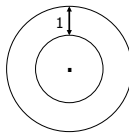


Notes 2 - Hardware

38

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

- 8 GB Disk
- If all tracks have 256 sectors
 - Outermost density: 100,000 bits/inch
 - Inner density: 250,000 bits/inch



CS 525



Notes 2 - Hardware

39

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

- Outer third of tracks: 320 sectors
- Middle third of tracks: 256
- Inner third of tracks: 192

- Density: 114,000 → 182,000 bits/inch

CS 525



Notes 2 - Hardware

40

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Timing for new Megatron 747 (Ex 2.3)

- Time to read 4096-byte block:
 - MIN: 0.5 ms
 - MAX: 33.5 ms
 - AVE: 14.8 ms

CS 525



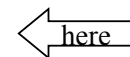
Notes 2 - Hardware

41

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Outline

- Hardware: Disks
- Access Times
- Example: Megatron 747
- Optimizations
- Other Topics
 - Storage Costs
 - Using Secondary Storage
 - Disk Failures



CS 525



Notes 2 - Hardware

42

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Optimizations (in controller or O.S.)

- Disk Scheduling Algorithms
 - e.g., elevator algorithm
- Track (or larger) Buffer
- Pre-fetch
- Arrays
- Mirrored Disks
- On Disk Cache

CS 525



Notes 2 - Hardware

43

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Double Buffering

Problem: Have a File

» Sequence of Blocks B1, B2

Have a Program

» Process B1

» Process B2

» Process B3

⋮

CS 525



Notes 2 - Hardware

44

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Single Buffer Solution

- (1) Read B1 → Buffer
- (2) Process Data in Buffer
- (3) Read B2 → Buffer
- (4) Process Data in Buffer ...

CS 525



Notes 2 - Hardware

45

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Say P = time to process/block

R = time to read in 1 block

n = # blocks

Single buffer time = $n(P+R)$

CS 525

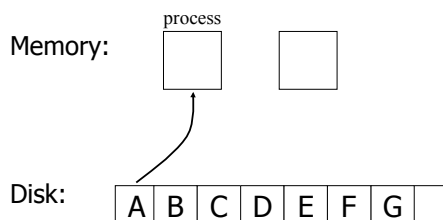


Notes 2 - Hardware

46

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Double Buffering



CS 525

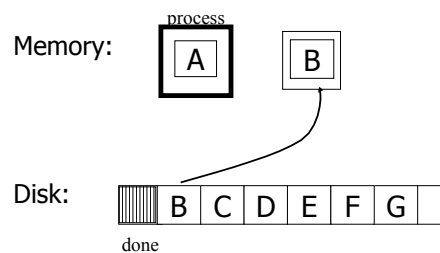


Notes 2 - Hardware

47

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Double Buffering



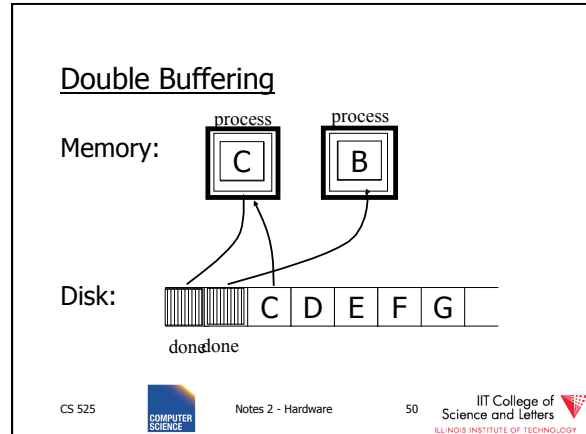
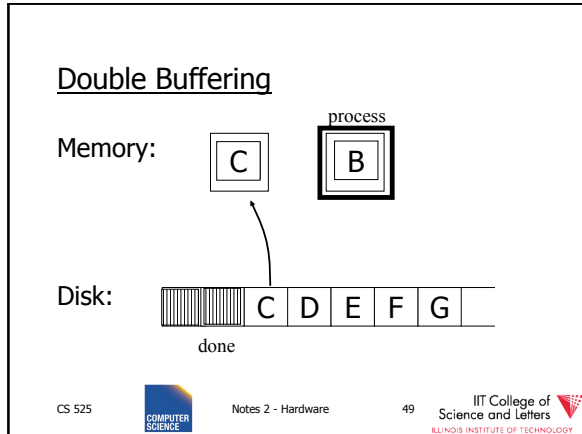
CS 525



Notes 2 - Hardware

48


IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY



Say $P \geq R$

P = Processing time/block
R = IO time/block
n = # blocks

What is processing time?


CS 525  Notes 2 - Hardware 51 IIT College of Science and Letters ILLINOIS INSTITUTE OF TECHNOLOGY

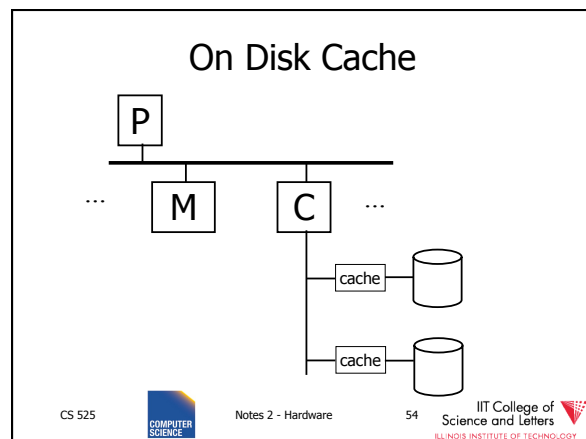
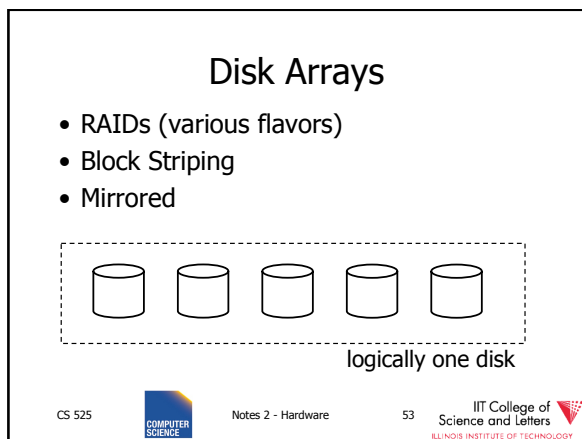
Say $P \geq R$

P = Processing time/block
R = IO time/block
n = # blocks

What is processing time?

- Double buffering time = $R + nP$
- Single buffering time = $n(R+P)$

CS 525  Notes 2 - Hardware 52 IIT College of Science and Letters ILLINOIS INSTITUTE OF TECHNOLOGY



Block Size Selection?

- Big Block → Amortize I/O Cost, Less Management Overhead

Unfortunately...

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

CS 525



Notes 2 - Hardware

55

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Trend

- As memory prices drop, blocks get bigger ...

CS 525

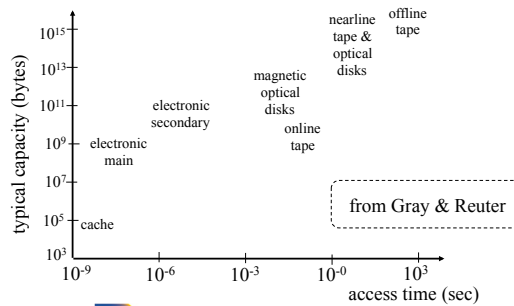


Notes 2 - Hardware

56

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Storage Cost



CS 525

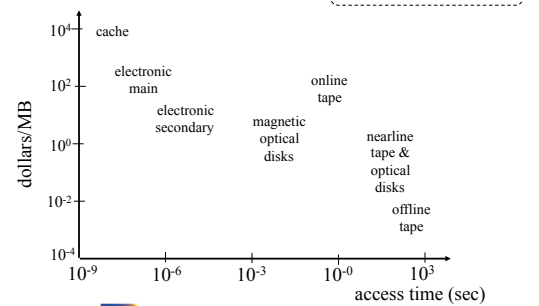


Notes 2 - Hardware

57

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Storage Cost



CS 525



Notes 2 - Hardware

58

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Using secondary storage effectively

- Example: Sorting data on disk
- Conclusion:
 - I/O costs dominate
 - Design algorithms to reduce I/O
- Also: How big should blocks be?

CS 525



Notes 2 - Hardware

59

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Five Minute Rule

- THE 5 MINUTE RULE FOR TRADING MEMORY FOR DISC ACCESSES
Jim Gray & Franco Putzolu
May 1985
- The Five Minute Rule, Ten Years Later
Goetz Graefe & Jim Gray
December 1997

CS 525



Notes 2 - Hardware

60

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Five Minute Rule

- Say a page is accessed every X seconds
- CD = cost if we keep that page on disk
 - \$D = cost of disk unit
 - I = numbers IOs that unit can perform per second
 - In X seconds, unit can do XI IOs
 - So CD = \$D / XI

CS 525



Notes 2 - Hardware

61

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Five Minute Rule

- Say a page is accessed every X seconds
- CM = cost if we keep that page on RAM
 - \$M = cost of 1 MB of RAM
 - P = numbers of pages in 1 MB RAM
 - So CM = \$M / P

CS 525



Notes 2 - Hardware

62

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Five Minute Rule

- Say a page is accessed every X seconds
- If CD is smaller than CM,
 - keep page on disk
 - else keep in memory
- Break even point when CD = CM, or

$$X = \frac{\$D P}{I \$M}$$

CS 525



Notes 2 - Hardware

63

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Using '97 Numbers

- P = 128 pages/MB (8KB pages)
- I = 64 accesses/sec/disk
- \$D = 2000 dollars/disk (9GB + controller)
- \$M = 15 dollars/MB of DRAM
- X = 266 seconds (about 5 minutes)
(did not change much from 85 to 97)

CS 525



Notes 2 - Hardware

64

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Disk Failures

- Partial → Total
- Intermittent → Permanent

CS 525



Notes 2 - Hardware

65

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Coping with Disk Failures

- Detection
 - e.g. Checksum
- Correction
 - ⇒ Redundancy

CS 525



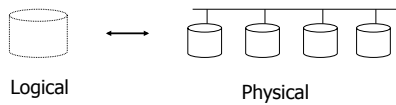
Notes 2 - Hardware

66

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

At what level do we cope?

- Single Disk
 - e.g., Error Correcting Codes
- Disk Array



CS 525

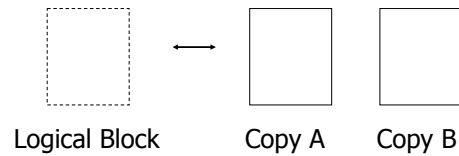


Notes 2 - Hardware

67

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

→ Operating System e.g., Stable Storage



CS 525



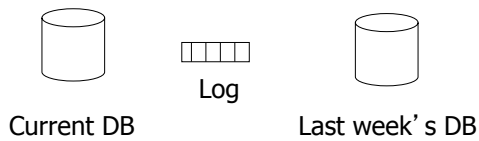
Notes 2 - Hardware

68

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

→ Database System

- e.g.,



CS 525



Notes 2 - Hardware

69

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Summary

- Secondary storage, mainly disks
- I/O times + formulas
 - Sequential vs. random
- I/Os should be avoided,
especially random ones.....
- OS optimizations
- Disk errors

CS 525



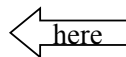
Notes 2 - Hardware

70

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Outline

- Hardware: Disks
- Access Times
- Example: Megatron 747
- Optimizations
- Other Topics
 - Storage Costs
 - Using Secondary Storage
 - Disk Failures



CS 525



Notes 2 - Hardware

71

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Outlook - Hardware

- Disk Access is the main limiting factor
- However, to implement fast DBMS
 - need to understand other parts of the hardware
 - Memory hierarchy
 - CPU architecture: pipelining, vector instructions, OOE, ...
 - SSD storage
 - need to understand how OS manages hardware
 - File access, VM, Buffering, ...

CS 525

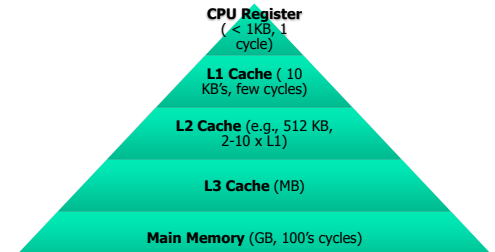


Notes 2 - Hardware

72

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Memory Hierarchy



CS 525



Notes 2 - Hardware

73

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Memory Hierarchy

- **Compare:** Disk vs. Main Memory
- Reduce accesses to main memory
- Cache conscious algorithms

CS 525



Notes 2 - Hardware

74

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

Increasing Amount of Parallelism

- Contention on, e.g., Memory
- NUMA
- Algorithmic Challenges
 - How to parallelize algorithms?
 - Sometime: Completely different approach required
 - -> Rewrite large parts of DBMS

CS 525



Notes 2 - Hardware

75

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY

New Trend: Software/Hardware Co-design

- Actually, revived trend: database machines (80's)
- New goals: power consumption
- Design specific hardware and write special software for it
- E.g., Oracle Exadata, Oracle Labs

CS 525



Notes 2 - Hardware

76

IIT College of
Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY