Hardware & the Memory Hierarchy

Parke Godfrey

Slides: Thanks to

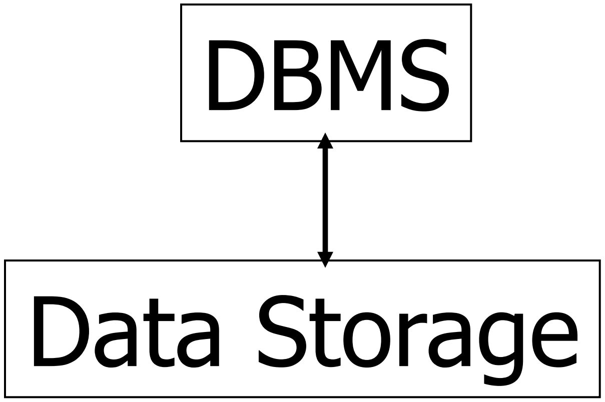
- Hector Garcia-Molina
- Jeffery Ullman

Aligned with the textbook, “*Database Systems: The Complete Book*”.

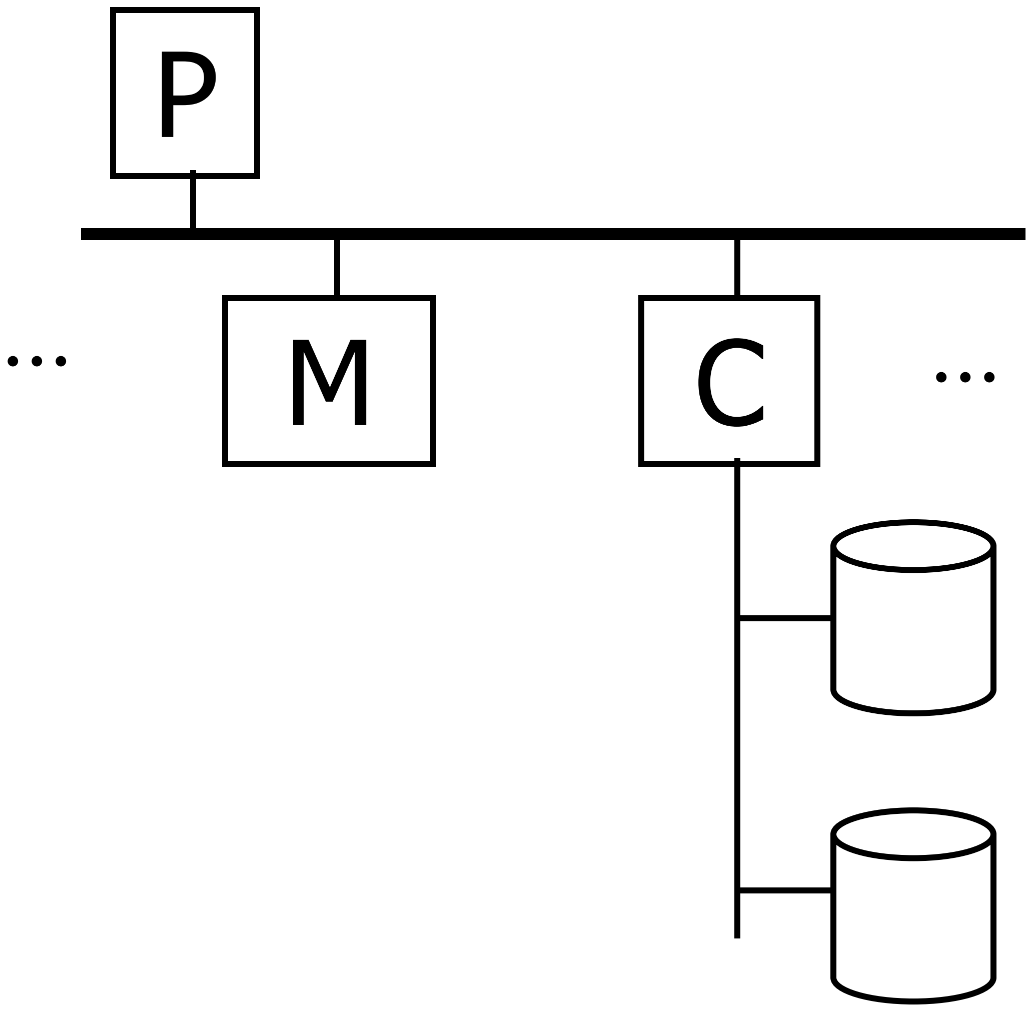
Outline

- Hardware: Disks
- Access Times
- Solid State Drives (SSDs)
- Optimizations
- Other Topics:
  - Storage costs
  - Using secondary storage
  - Disk failures

Hardware dictates our design choices



- Data lives in secondary storage.
  - non-volatile
  - cheaper per byte
  - for us, random access (per *block*)



Typical Computer

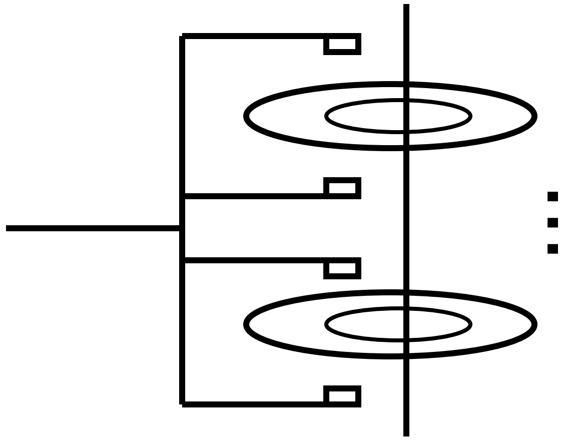
Secondary Storage

Secondary storage

Many flavors:

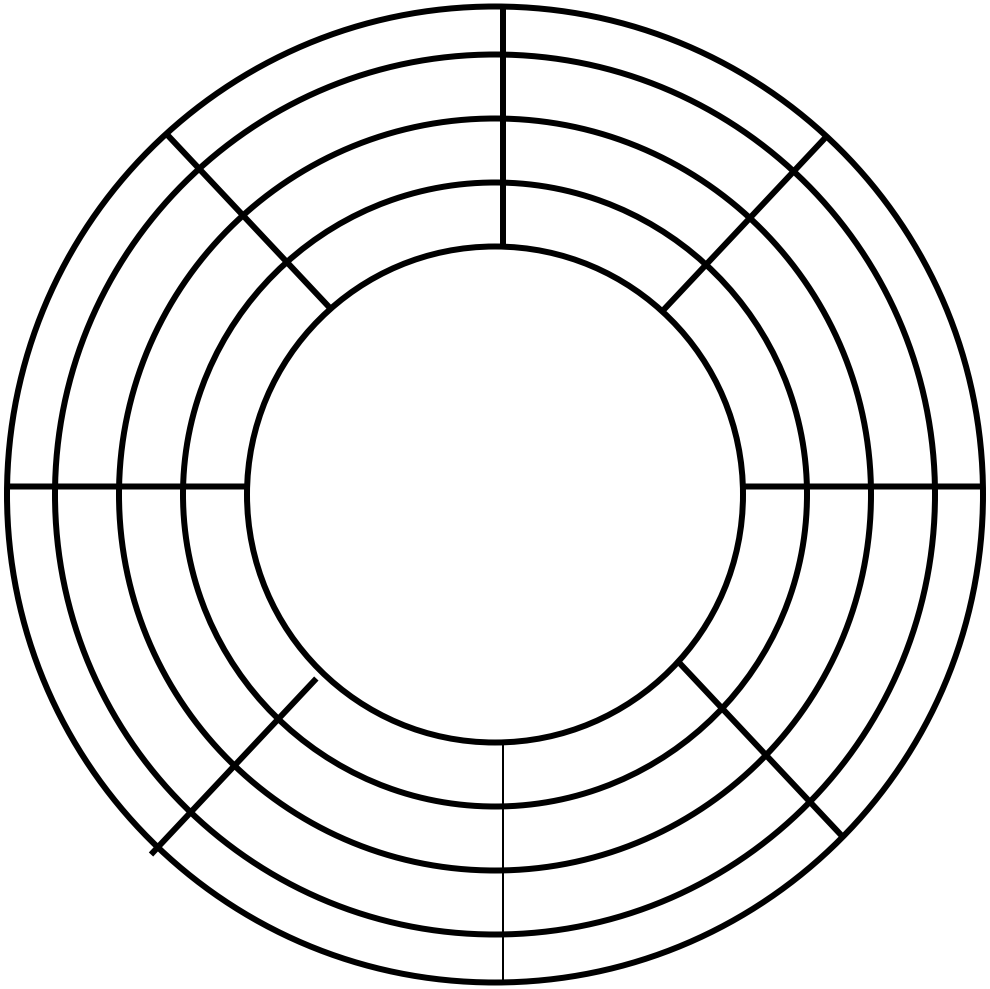
- Disk:	Floppy (hard, soft)
	Removable Packs
	Winchester
	SSD disks
	Optical, CD-ROM…
	Arrays
- Tape	Reel, cartridge
	Robots

Focus on: "Typical Disk"



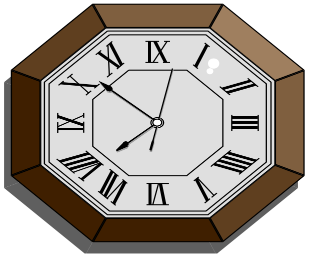
Terms: Platter, Head, Actuator
Cylinder, Track
Sector (physical),
Block (logical), Gap

Top View



Disk Access Time

I want block X →

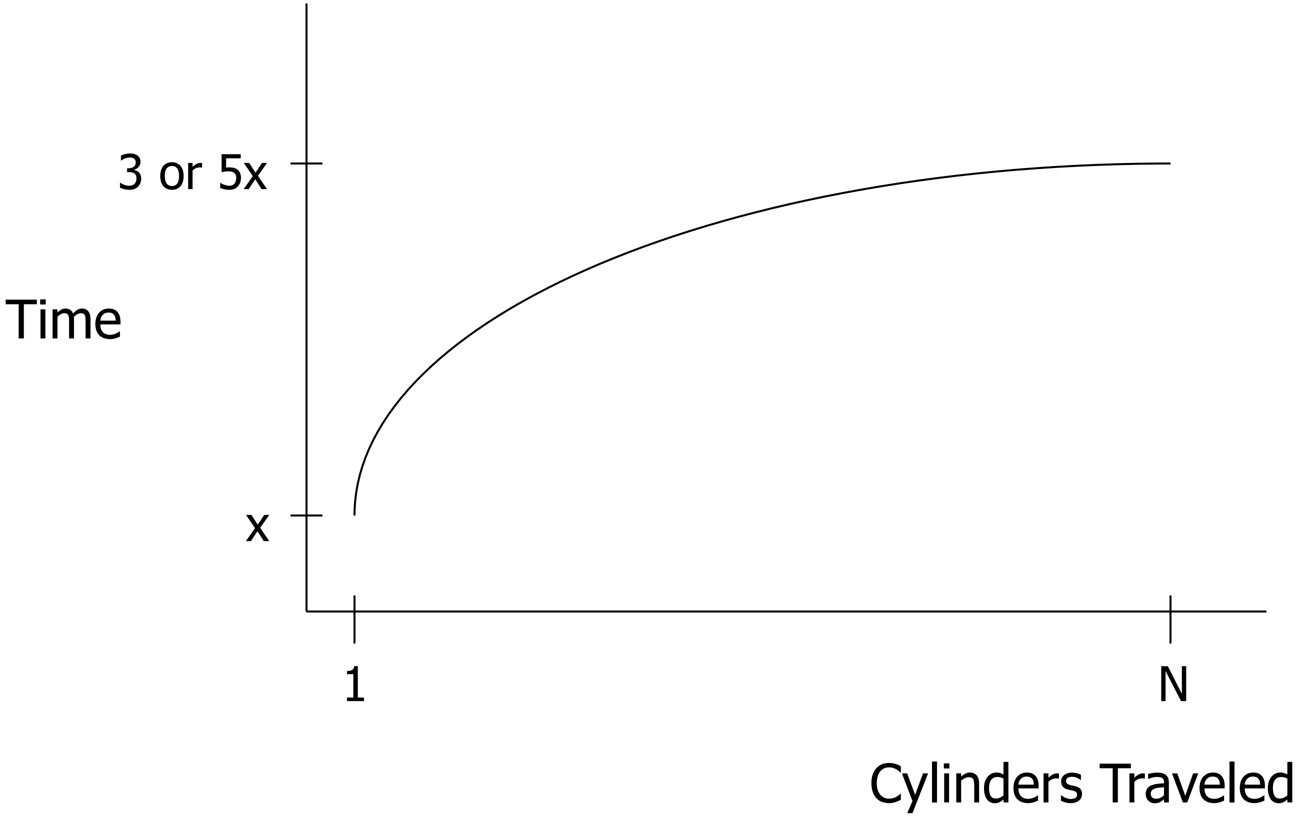


→ block x in memory

?

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

Seek Time



Average Random Seek Time

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

Average Random Seek Time

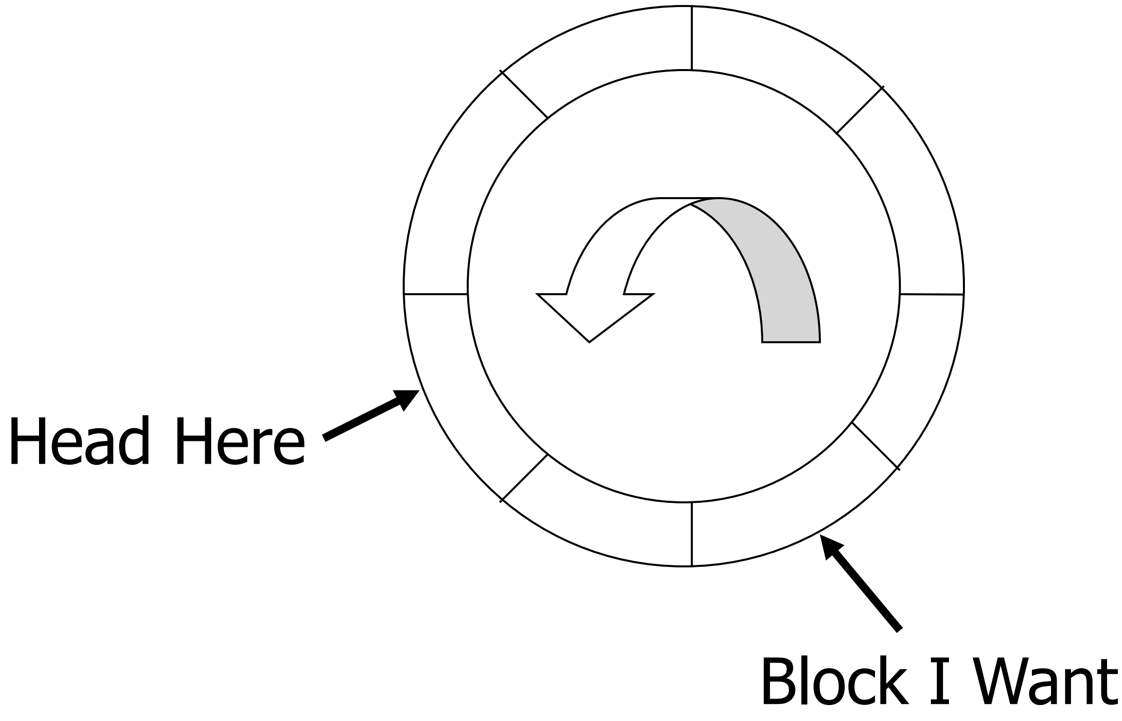
$$S = \frac{\displaystyle\sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \text{SEEKTIME } (i \rightarrow j)}{N(N-1)}$$

Typical Seek Time

- Ranges from
  - 4ms for high end drives
  - 15ms for mobile devices
- Typical SSD: ranges from
  - 0.08ms
  - 0.16ms

- Source: Wikipedia, "Hard disk drive performance characteristics"

Rotational Delay



Average Rotational Delay

R = 1/2 revolution

R=0 for SSDs

Typical HDD figures

HDD Spindle [rpm]	Average rotational latency [ms]
4,200	7.14
5,400	5.56
7,200	4.17
10,000	3.00
15,000	2.00

Source: Wikipedia, "Hard disk drive performance characteristics"

Transfer Rate: t

- value of t ranges from
  - up to 1000 Mbit/sec
  - 432 Mbit/sec 12x Blu-Ray disk
  - 1.23 Mbits/sec 1x CD
  - for SSDs, limited by interface e.g., SATA 3000 Mbit/s
- transfer time: $\frac{\text{block size}}{t}$

Other Delays

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

Other Delays

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

“Typical” Value: 0

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

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

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

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

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

Cost for Writing similar to Reading

…. unless we want to verify!

need to add (full) rotation + $\frac{\text{Block size}}{t}$

- To Modify a Block?

- To Modify a Block?

To Modify Block:

(a) Read Block

(b) Modify in Memory

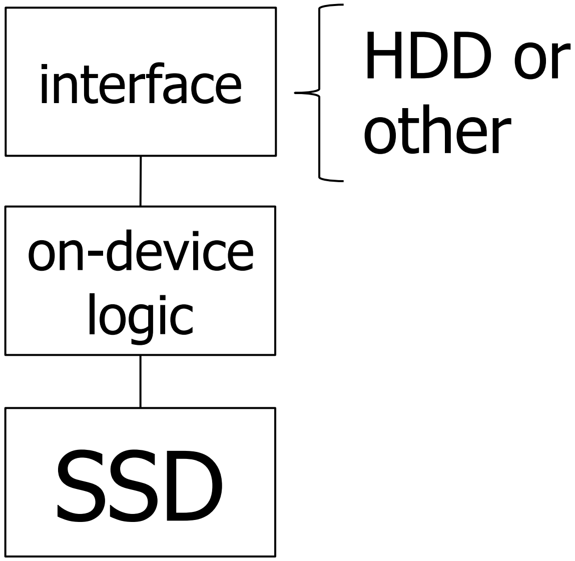
(c) Write Block

[(d) Verify?]

SSDs

Source: Reza Sadri, STEC ("the SSD Company")

- storage is block oriented (not random access)
- lots of errors
  - e.g., write of one block may cause an error of nearby block
  - e.g., a block can only be written a limited number of times
- logic masks most issues
  - e.g., using log structure
- sequential writes improve throughput (less bookkeeping)
  - latency for seq. writes = random writes
  - performance seq. reads = random reads



SSD vs Hard Disk Comparison (from Wikipedia)

- **Factors:** start up time, random access time, read latency time, data transfer rate, read performance, fragmentation, noise, temperature control, environmental factors, installation and mounting, magnetic fields, weight and size, reliability, secure writing, cost, capacity, R/W symmetry, power consumption.

Random Access Time

- **SSD:** Typically under 0.1 ms. As data can be retrieved directly from various locations of the flash memory, access time is usually not a big performance bottleneck.
- **Hand Drive:** Ranges from 2.9 (high end server drive) to 12 ms (laptop HDD) due to the need to move the heads and wait for the data to rotate under the read/write head

Data Transfer Rate

- **SSD:** In consumer products the maximum transfer rate typically ranges from about 100 MB/s to 600 MB/s, depending on the disk. Enterprise market offers devices with multi-gigabyte per second throughput.
- **Hard Disk:** Once the head is positioned, an enterprise HDD can transfer data at about 140 MB/s. In practice transfer speeds are lower due to seeking. Data transfer rate depends also upon rotational speed, which can range from 4,200 to 15,000 rpm and also upon the track (reading from the outer tracks is faster due higher).

Reliability

- **SSD:** Reliability varies across manufacturers and models with return rates reaching 40% for specific drives. As of 2011 leading SSDs have lower return rates than mechanical drives. Many SSDs critically fail on power outages; a December 2013 survey found that only some of them are able to survive multiple power outages.
- **Hard Disk:** According to a study performed by CMU for both consumer and enterprise-grade HDDs, their average failure rate is 6 years, and life expectancy is 9–11 years. Leading SSDs have overtaken hard disks for reliability, however the risk of a sudden, catastrophic data loss can be lower for mechanical disks.

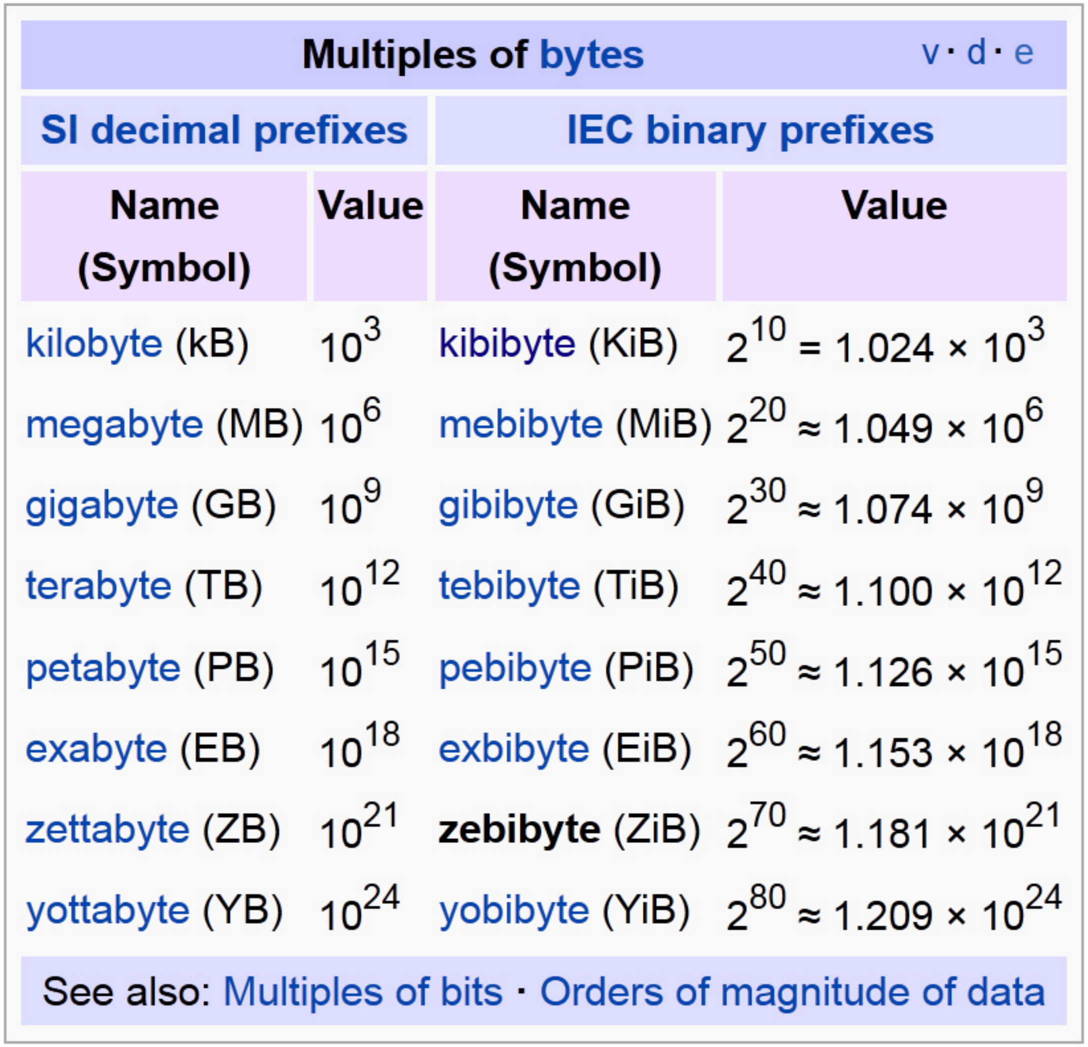
Cost and Capacity

- **SSD:** NAND flash SSDs have reached US$0.59 per GB. In 2013, SSDs were available in sizes up to 2 TB, but less costly 128 to 512 GB drives were more common.

- **Hard Drive:** HDDs cost about US$0.05 per GB for 3.5-inch and $0.10 per GB for 2.5-inch drives. In 2013, HDDs of up to 6 TB were available.

Kibibytes

- 1 kibibyte = 2^{10} bytes = 1024 bytes.



Multiples of bytes v · d · e

SI decimal prefixes		IEC binary prefixes	
Name (Symbol)	Value	Name (Symbol)	Value
kilobyte (kB)	10^3	kibibyte (KiB)	$2^{10} = 1.024 \times 10^3$
megabyte (MB)	10^6	mebibyte (MiB)	$2^{20} \approx 1.049 \times 10^6$
gigabyte (GB)	10^9	gibibyte (GiB)	$2^{30} \approx 1.074 \times 10^9$
terabyte (TB)	10^{12}	tebibyte (TiB)	$2^{40} \approx 1.100 \times 10^{12}$
petabyte (PB)	10^{15}	pebibyte (PiB)	$2^{50} \approx 1.126 \times 10^{15}$
exabyte (EB)	10^{18}	exbibyte (EiB)	$2^{60} \approx 1.153 \times 10^{18}$
zettabyte (ZB)	10^{21}	**zebibyte** (ZiB)	$2^{70} \approx 1.181 \times 10^{21}$
yottabyte (YB)	10^{24}	yobibyte (YiB)	$2^{80} \approx 1.209 \times 10^{24}$

See also: Multiples of bits · Orders of magnitude of data

from Wikipedia

Outline

- Hardware: Disks
- Access Times
- Solid State Drives
- Optimizations here
- Other Topics
  - Storage Costs
  - Using Secondary Storage
  - Disk Failures

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

Double Buffering

Problem: Have a File

» Sequence of Blocks B1, B2

Have a Program

» Process B1

» Process B2

» Process B3

⋮

Single Buffer Solution

(1) Read B1 → Buffer

(2) Process Data in Buffer

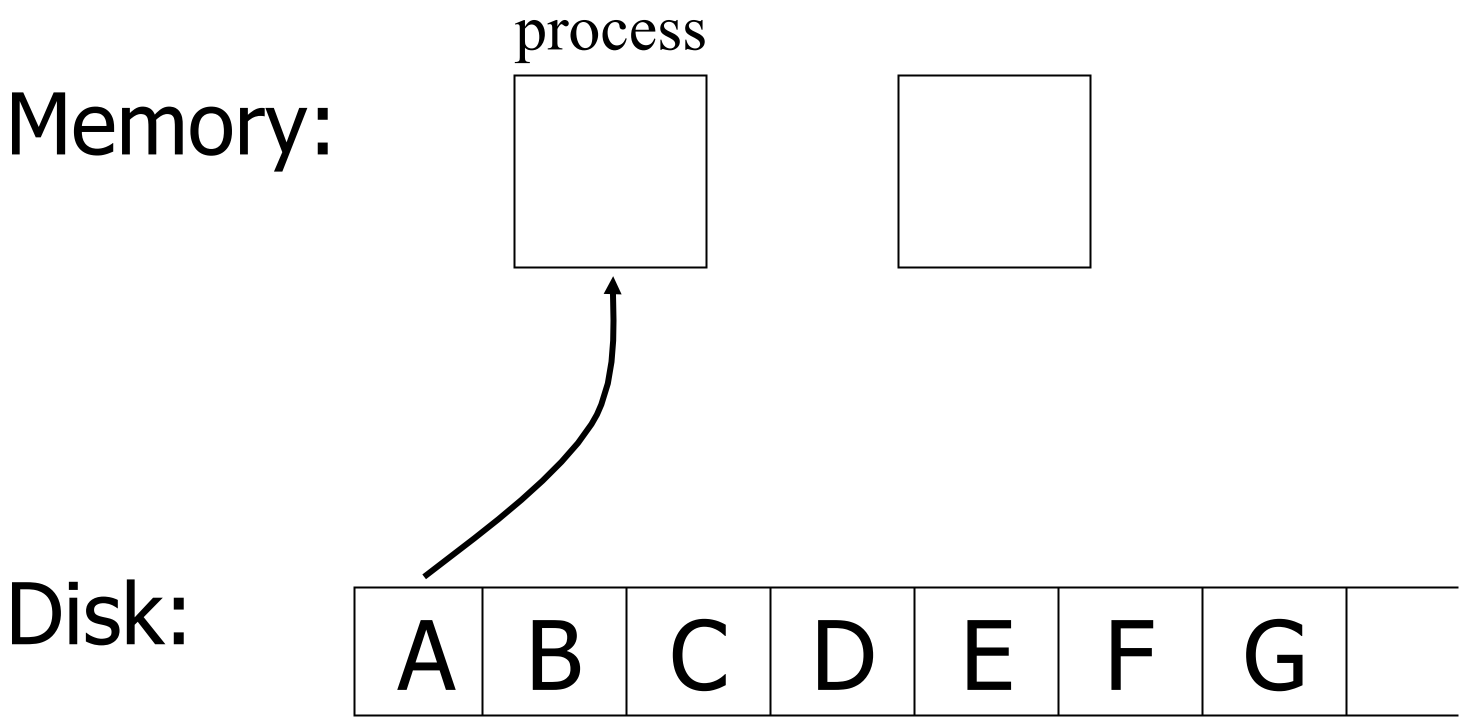
(3) Read B2 → Buffer

(4) Process Data in Buffer ...

Say P = time to process/block
R = time to read in 1 block
n = # blocks

Single buffer time = n(P+R)

Double Buffering



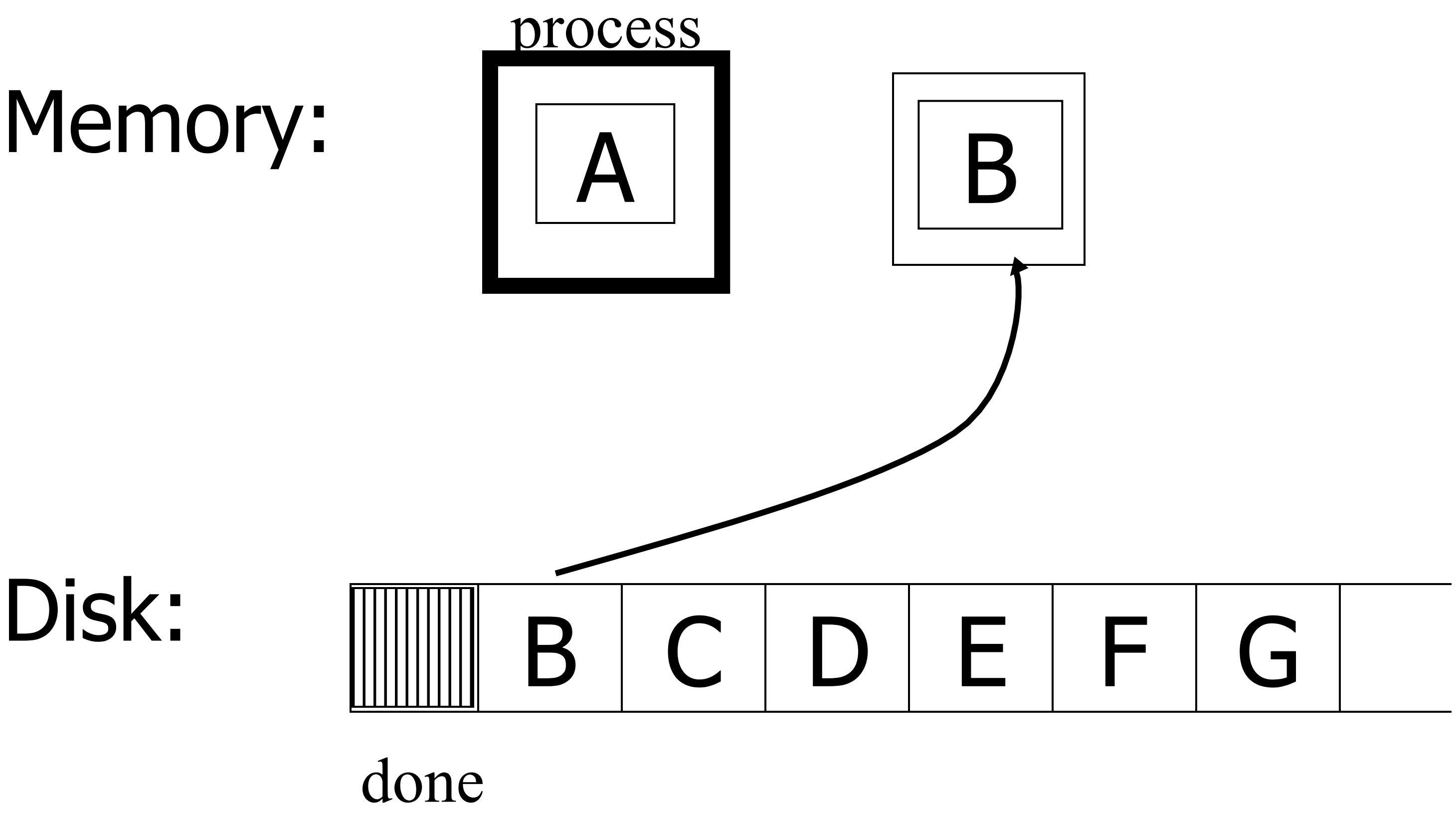
Memory:

process

Disk:

A B C D E F G

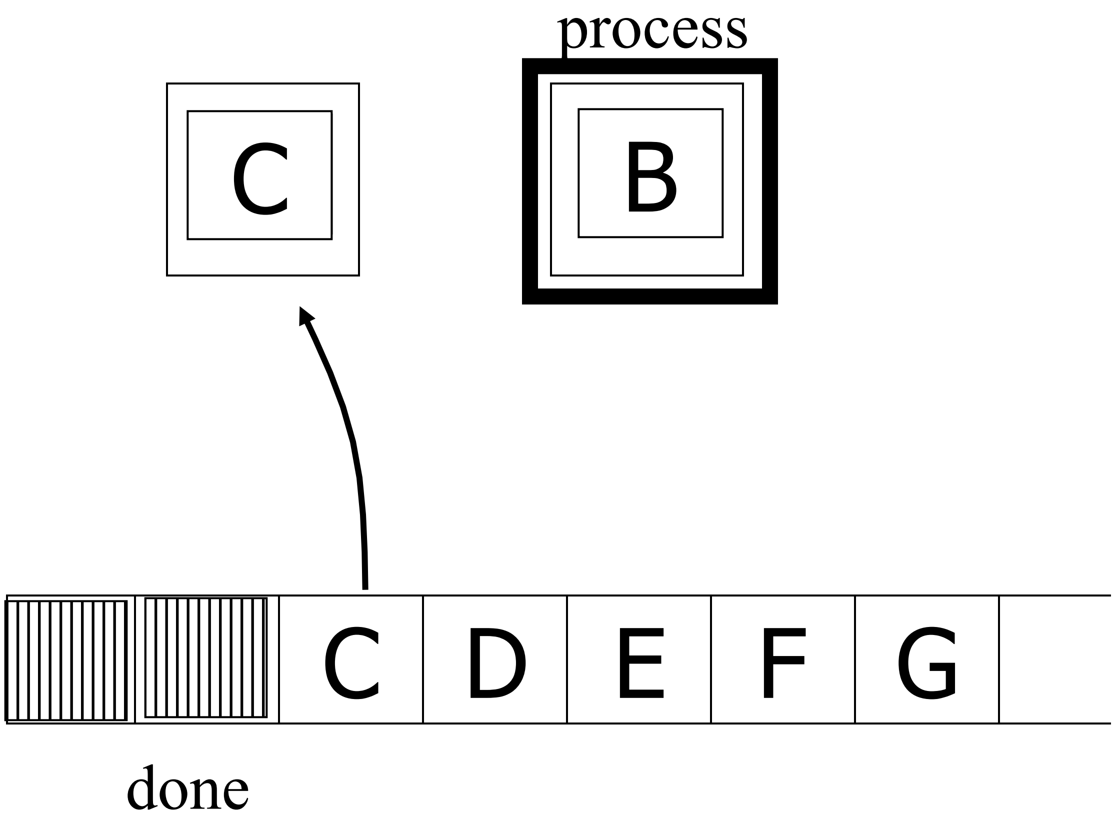
Double Buffering



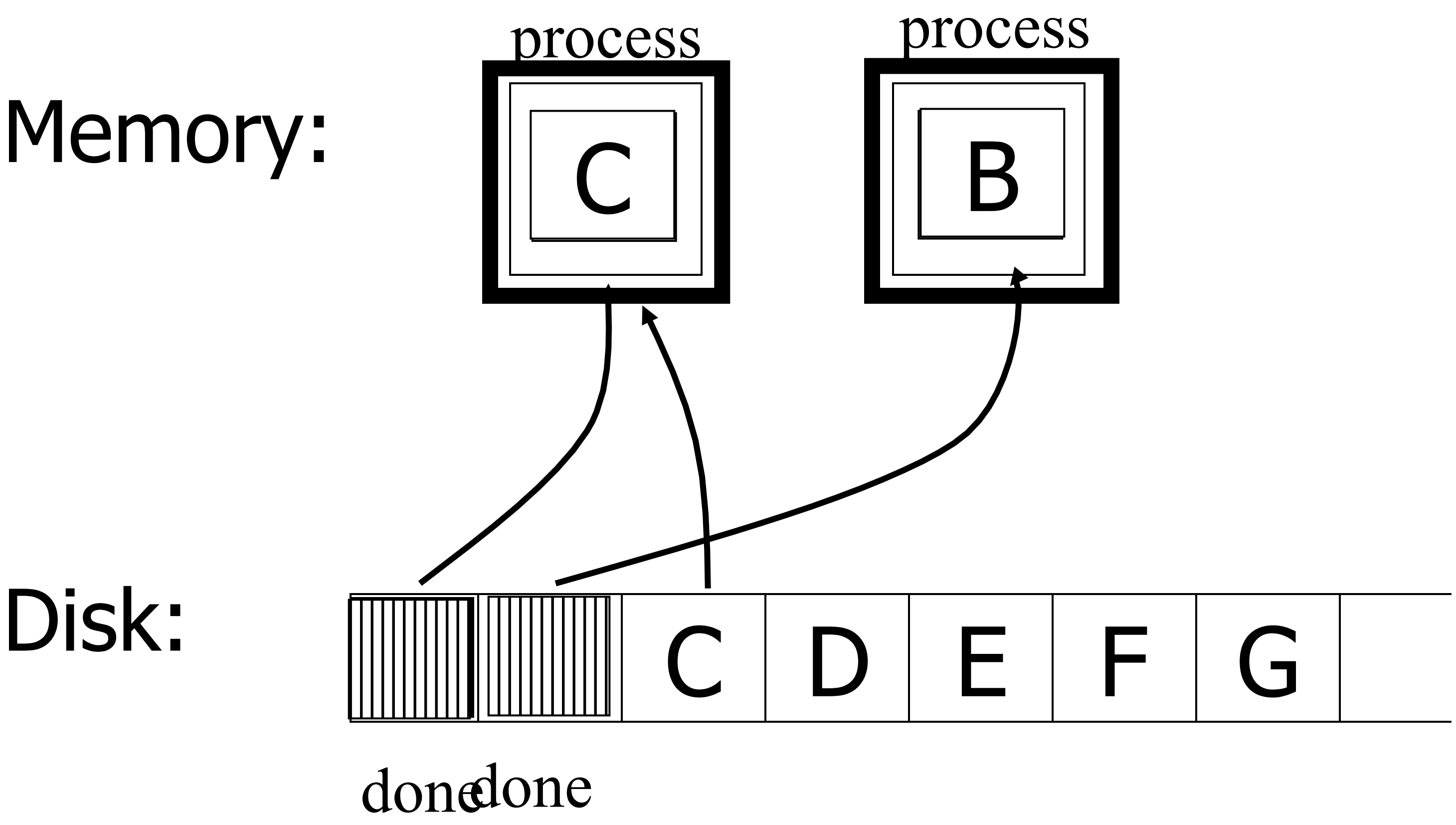
Double Buffering

Memory:

Disk:



Double Buffering



Say P ≥ R

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

What is processing time?

Say $P \geq R$

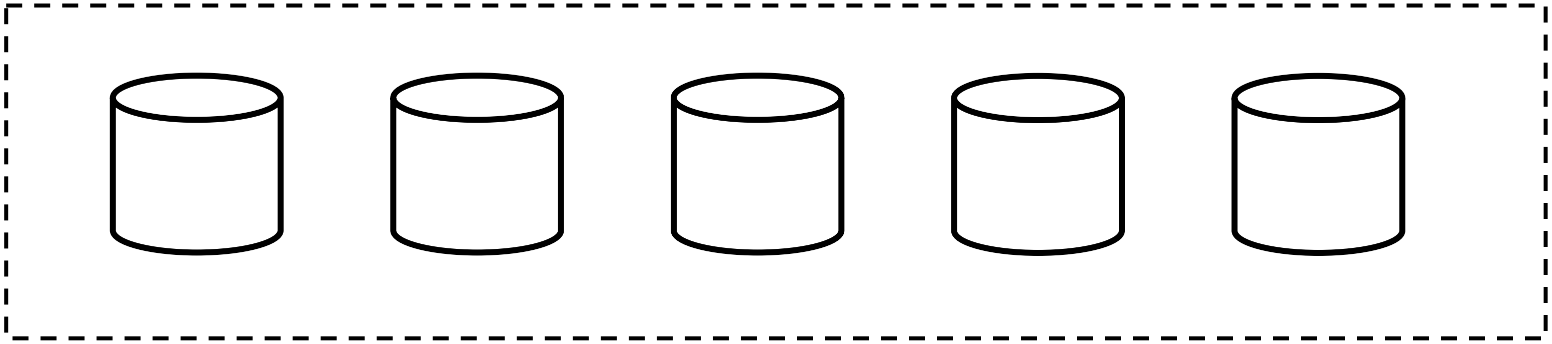
P = Processing time/block
R = IO time/block
n = # blocks

What is processing time?

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

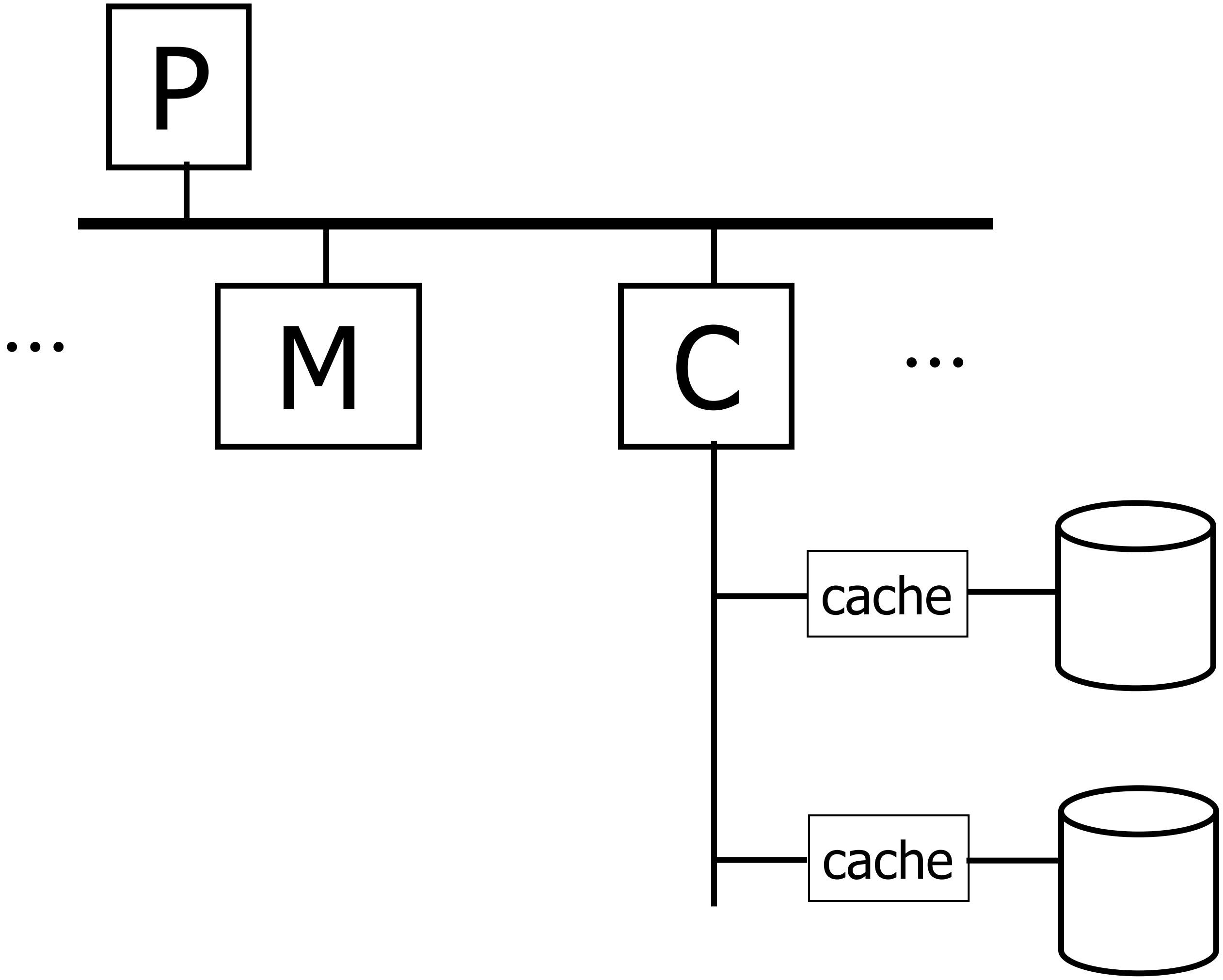
Disk Arrays

- RAIDs (various flavors)
- Block Striping
- Mirrored



logically one disk

On Disk Cache



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

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
  - In X seconds, unit can do XI IOs
  - So   CD = $D / XI

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

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

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

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)

Disk Failures (Sec 2.5)

- Partial → Total
- Intermittent → Permanent

Coping with Disk Failures

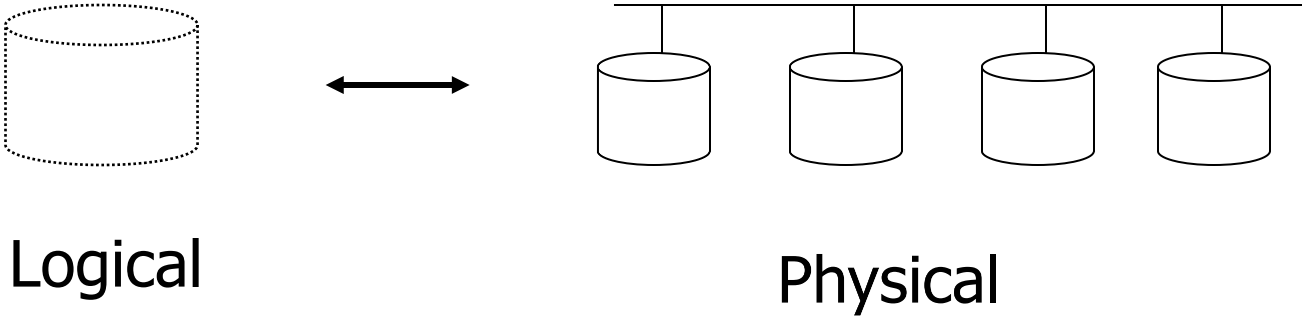
- Detection
  - e.g. Checksum

- Correction

  ⇒ Redundancy

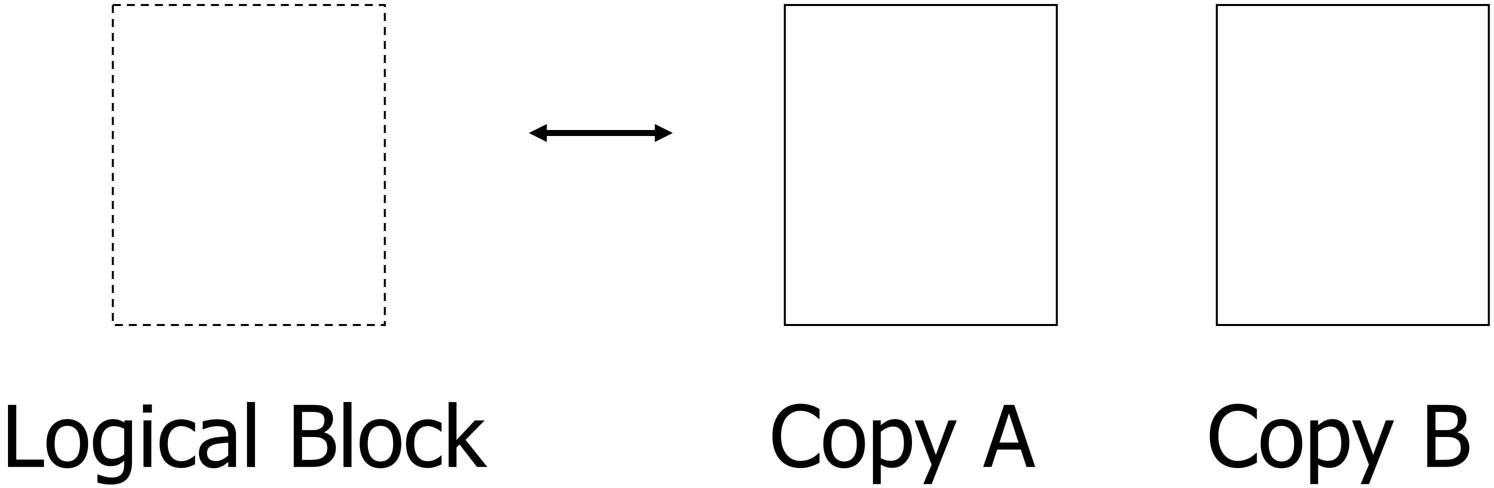
At what level do we cope?

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



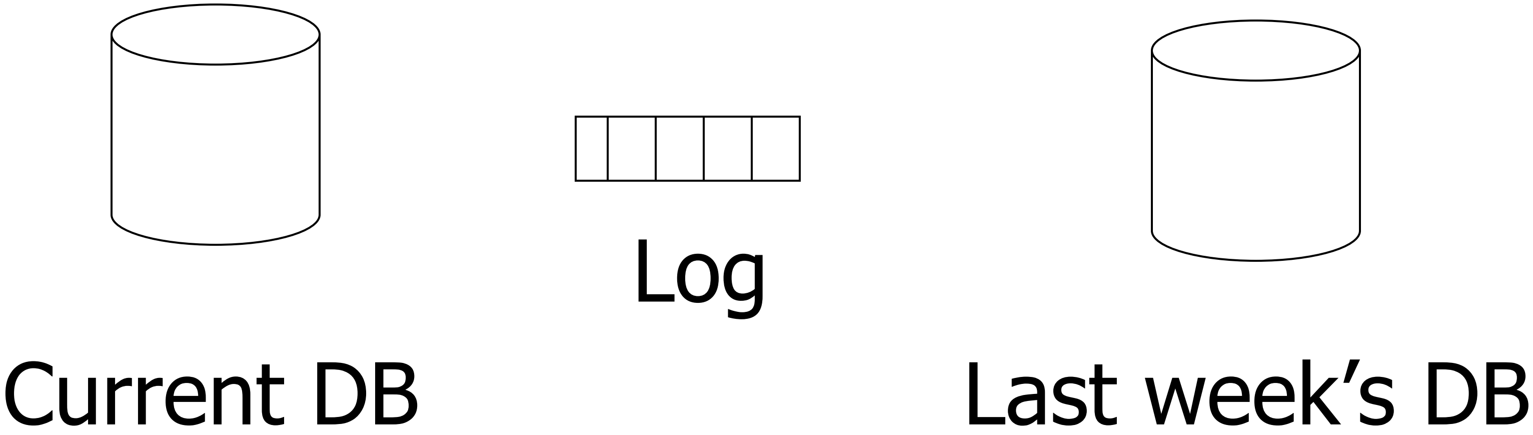
→ Operating System

e.g., Stable Storage



→ Database System

- e.g.,



Summary

- Secondary storage, mainly disks
- I/O times
- I/Os should be avoided,

  especially random ones…..

Outline

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

here