Unit 3. Memory hierarchy

Computer Architecture

Area of Computer Architecture and Technology
Department of Computer Science and Engineering
University of Oviedo

Fall, 2015

Objectives

1.- Performance improvements

Organizational changes to improve performance

2.- Support for multitasking operating systems

Requirements to deal with this task

Table of contents

Ideal requirements of the memory system

1. Extremely large capacity
   - several programs, many instructions, much data
2. Any word immediately available
   - several memory accesses (at least one per instruction)
   - sometimes the CPU waits for the memory system

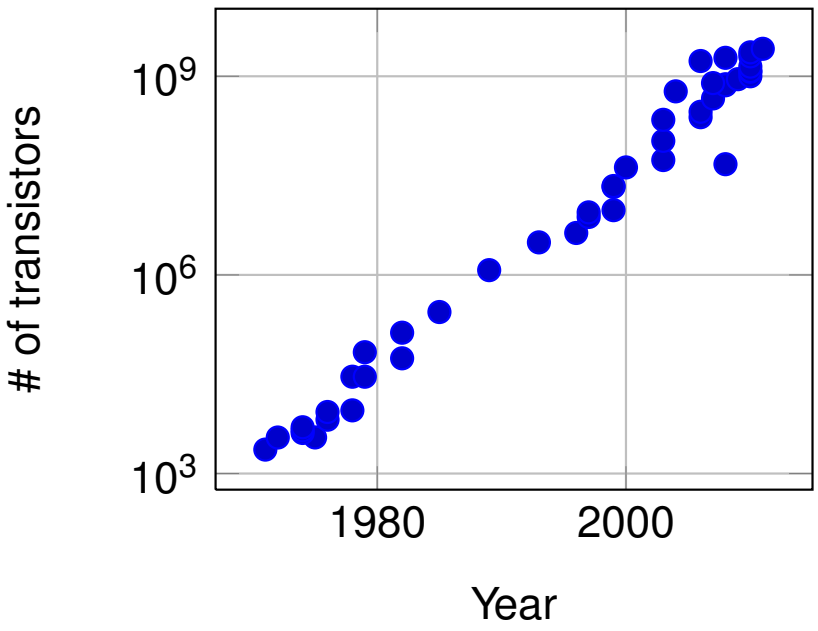
```
MOV R5, [R1]
```

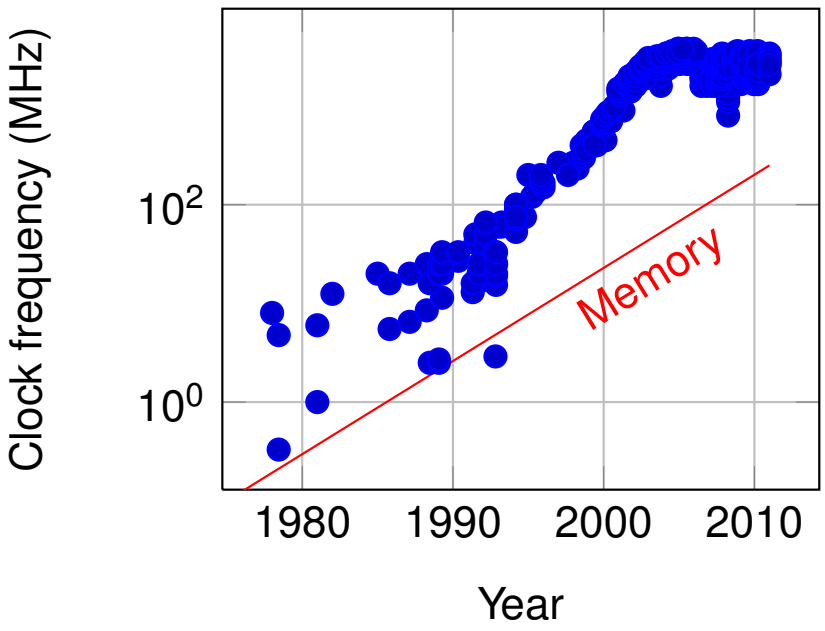
	Step	Control signals
Fetch	1	PC-IB, IB-MAR, TMPE_CLR, CARRY_IN, ADD, ALU-TMPS, READ
	2	TMPS-IB, IB-PC
	3	MDR-IB, IB-IR
Exec.	4	R1-IB, IB-MAR, READ
	5	Wait cycle
	6	MDR-IB, IB-R5, FIN

Performance differences

Moore's law

The number of transistors in an IC is doubled approximately every two years





Performance gap between memory and CPU is growing → memory wall

Memory-wall effect

Example: quantifying the effect

$$\text{Effective frequency} = \frac{1}{\text{Period} + (\text{T. mem. access} \times \text{Mem. access ratio})}$$

- CPU: 1 GHz
- Memory latency: 10 ns
- 1 out of 4 instructions have operands in memory

$$\text{Memory access ratio} = 1 + 0.25$$

$$\frac{1}{1\,\text{ns} + (10\,\text{ns} \times 1.25)} = 0.08\,\text{GHz}$$

This CPU behaves as a CPU with a clock frequency of 80 MHz (12.5 times slower)

Memory technologies

	Static RAM (SRAM)	Dynamic RAM (DRAM)	Hard disk
Basic cell	flip-flop	capacitor	magnetic material
Persistence	power needed	refresh needed	permanent
Latency	≈ 0.5 ns (CPU freq.)	≈ 10 ns	$\approx 10^7$ ns
Cost	high (6 transistors)	medium (1 transistor)	low

Cost

- basic cell material
- power to store a bit
- cost/bit

Memory technologies

	Static RAM (SRAM)	Dynamic RAM (DRAM)	Hard disk
Basic cell	flip-flop	capacitor	magnetic material
Persistence	power needed	refresh needed	permanent
Latency	≈ 0.5 ns (CPU freq.)	≈ 10 ns	$\approx 10^7$ ns
Cost	high (6 transistors)	medium (1 transistor)	low

Cost

- basic cell material
- power to store a bit
- cost/bit

Memory technologies

	Static RAM (SRAM)	Dynamic RAM (DRAM)	Hard disk
Basic cell	flip-flop	capacitor	magnetic material
Persistence	power needed	refresh needed	permanent
Latency	≈ 0.5 ns (CPU freq.)	≈ 10 ns	$\approx 10^7$ ns
Cost	high (6 transistors)	medium (1 transistor)	low

Cost

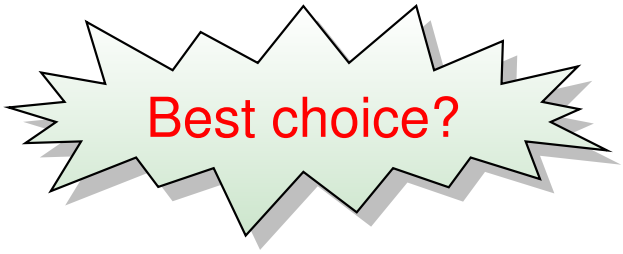
- basic cell material
- power to store a bit
- cost/bit

Memory technologies

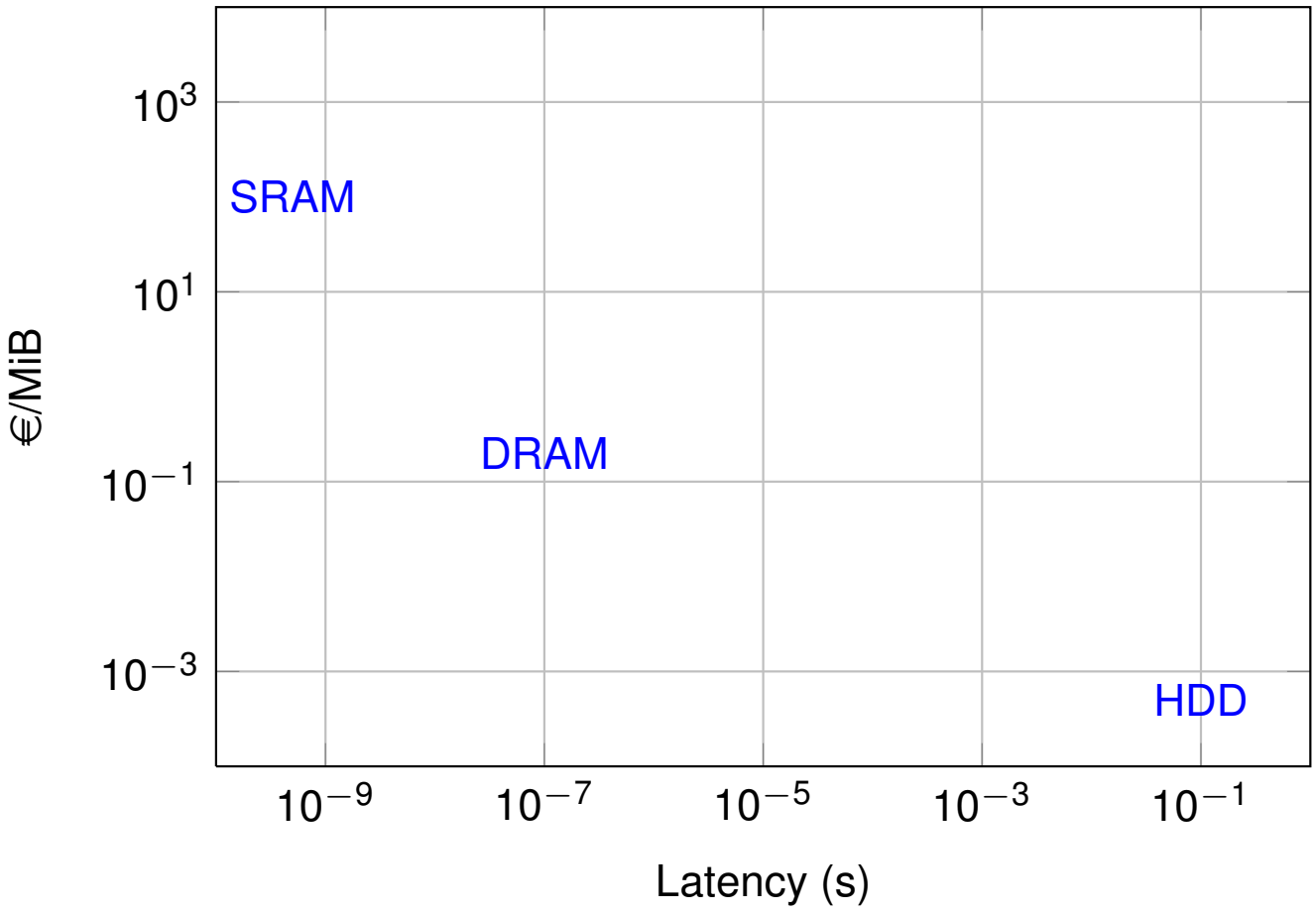
	Static RAM (SRAM)	Dynamic RAM (DRAM)	Hard disk
Basic cell	flip-flop	capacitor	magnetic material
Persistence	power needed	refresh needed	permanent
Latency	≈ 0.5 ns (CPU freq.)	≈ 10 ns	$\approx 10^7$ ns
Cost	high (6 transistors)	medium (1 transistor)	low

Cost

- basic cell material
- power to store a bit
- cost/bit



Memory technologies



No technology meets the objectives by its own

Memory hierarchy

Requirements

- High speed (latency and bandwidth)
- High capacity
- Low cost

Solution: memory hierarchy

Combine memory technologies in several levels

This solution will be successful if:

- Fast and small memories contain data with higher access probability
- Slow and big memories contain data with lower access probability

Memory hierarchy

Requirements

- High speed (latency and bandwidth)
- High capacity
- Low cost

Solution: memory hierarchy

Combine memory technologies in several levels

This solution will be successful if:

- Fast and small memories contain data with higher access probability
- Slow and big memories contain data with lower access probability

Principle of locality

Principle of locality

A correlation exists among memory accesses

Locality types
- Spatial: access to close items
- Temporal: access to recently accessed items

Presence
- Code
- Data

Principle of locality

A correlation exists among memory accesses

Locality types
- Spatial: access to close items
- Temporal: access to recently accessed items

Presence
- Code
  - spatial: sequential execution
  - temporal: execution of loops
- Data

Principle of locality

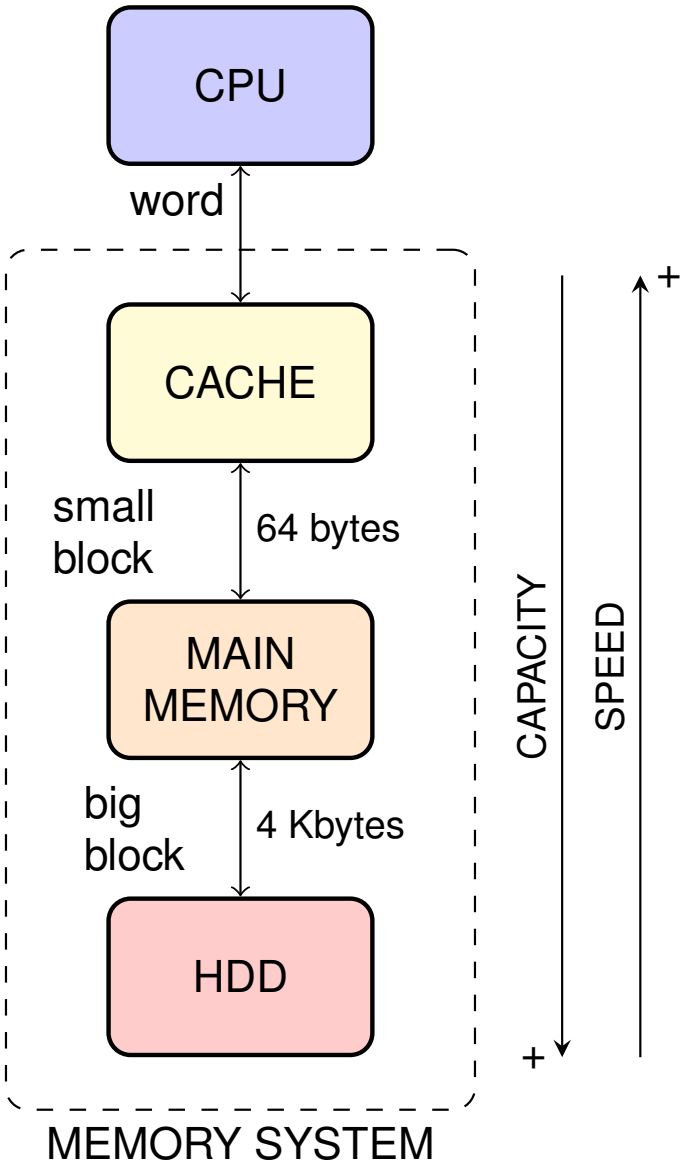
A correlation exists among memory accesses

Locality types
- Spatial: access to close items
- Temporal: access to recently accessed items

Presence
- Code
  - spatial: sequential execution
  - temporal: execution of loops
- Data
  - spatial: access to arrays
  - temporal: loop counters

Scheme

- Several levels in the hierarchy
- Different memory technologies are used
- The closer to the CPU, the faster
- The further from the CPU, the bigger



Operation

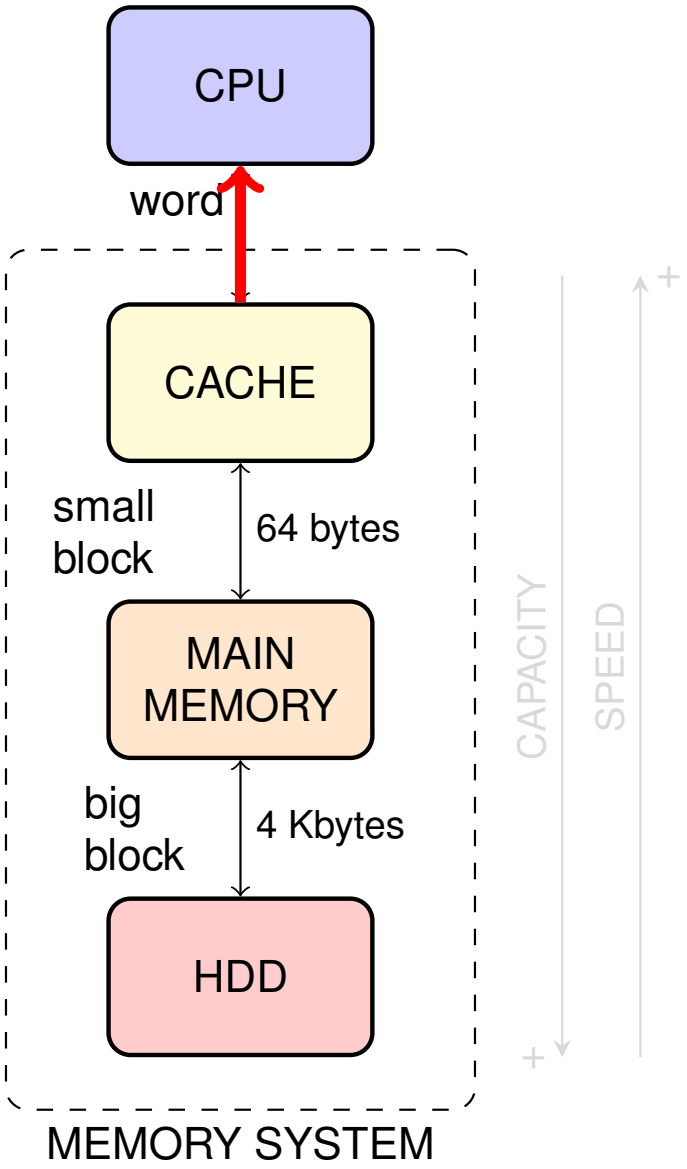
1. CPU wants to access a word
   - cache hit
   - cache miss
2. the word is fetched from a lower level in memory
3. misses can be chained

Scheme

- Several levels in the hierarchy
- Different memory technologies are used
- The closer to the CPU, the faster
- The further from the CPU, the bigger

Operation

1. CPU wants to access a word
   - cache hit
   - cache miss
2. the word is fetched from a lower level in memory
3. misses can be chained

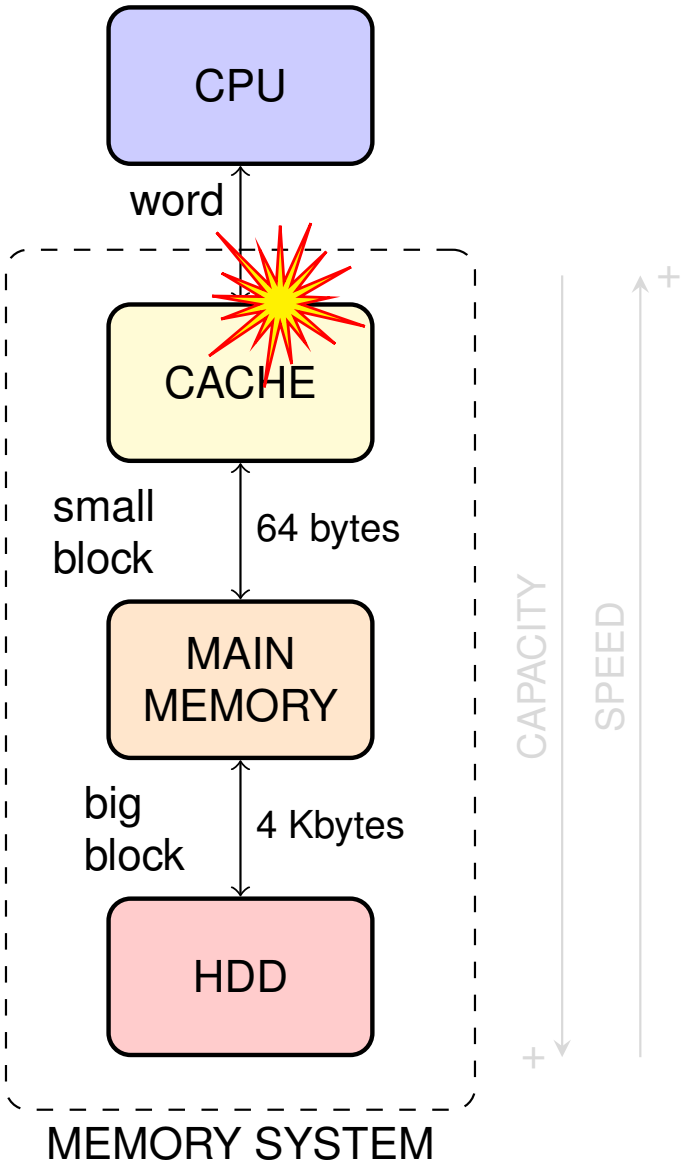


Scheme

- Several levels in the hierarchy
- Different memory technologies are used
- The closer to the CPU, the faster
- The further from the CPU, the bigger

Operation

1. CPU wants to access a word
   - cache hit
   - cache miss
2. the word is fetched from a lower level in memory
3. misses can be chained

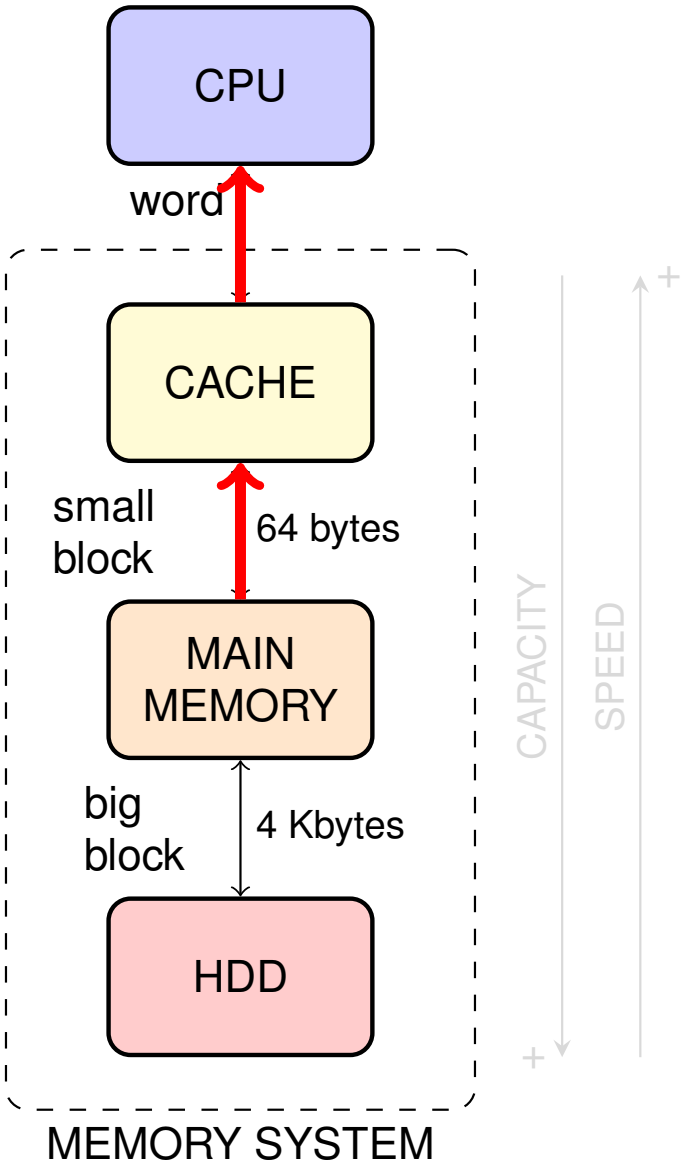


Scheme

- Several levels in the hierarchy
- Different memory technologies are used
- The closer to the CPU, the faster
- The further from the CPU, the bigger

Operation

1. CPU wants to access a word
   - cache hit
   - cache miss
2. the word is fetched from a lower level in memory
3. misses can be chained

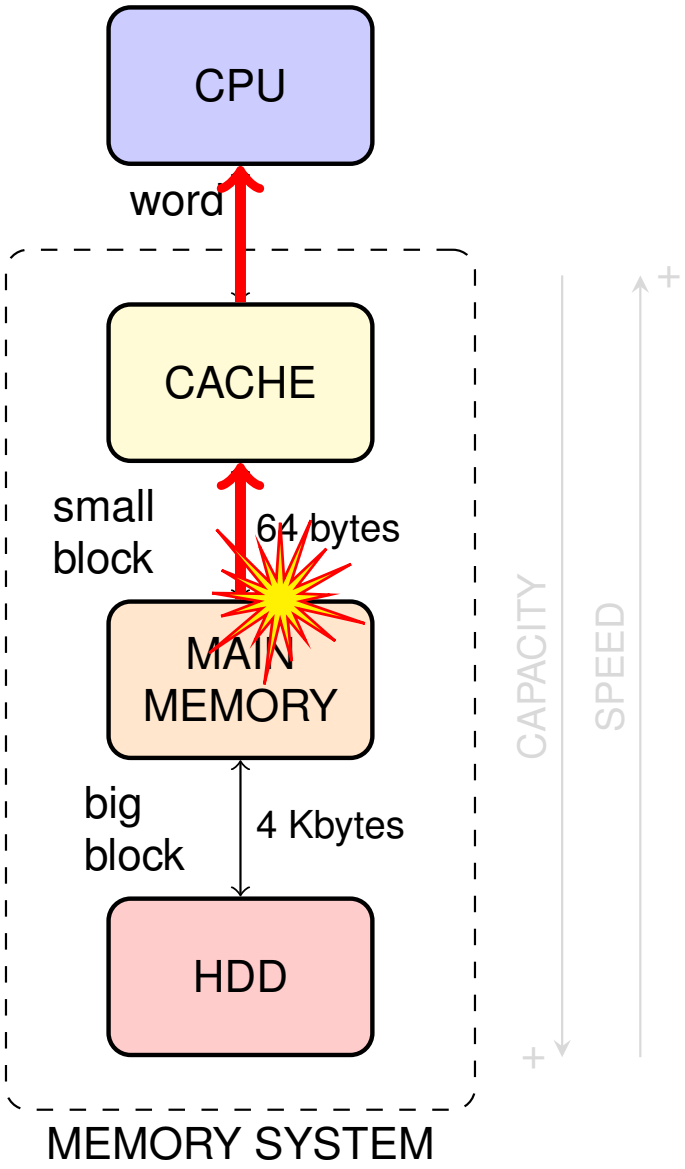


Scheme

- Several levels in the hierarchy
- Different memory technologies are used
- The closer to the CPU, the faster
- The further from the CPU, the bigger

Operation

1. CPU wants to access a word
   - cache hit
   - cache miss
2. the word is fetched from a lower level in memory
3. misses can be chained



Average memory access time

Example

A computer has the following memory hierarchy: cache, main memory and hard disk drive. The access time in each level is:

- $t_c = 1\,\text{ns}$
- $t_p = 10\,\text{ns}$ each byte
- $t_d = 10\,\text{ms}$ (for any block size)

The average hit rate in each level is:

- $A_c = 0.99 \Rightarrow 99\%$
- $A_p = 0.9999 \Rightarrow 99.99\%$

Cache block size (B_c) is 64 bytes

Which is the average access time to the memory system (t_{cpd})?

$$t_{cpd} = \underbrace{A_c \cdot t_c}_{\text{cache hit}}$$

Average memory access time

Example

A computer has the following memory hierarchy: cache, main memory and hard disk drive. The access time in each level is:

- $t_c = 1$ ns
- $t_p = 10$ ns each byte
- $t_d = 10$ ms (for any block size)

The average hit rate in each level is:

- $A_c = 0.99 \Rightarrow 99\%$
- $A_p = 0.9999 \Rightarrow 99.99\%$

Cache block size (B_c) is 64 bytes

Which is the average access time to the memory system (t_{cpd})?

$$t_{cpd} = \underbrace{A_c \cdot t_c}_{\text{cache hit}} + \underbrace{(1 - A_c) \times [t_{pd}]}_{\text{cache miss}}$$

Average memory access time

Example

A computer has the following memory hierarchy: cache, main memory and hard disk drive. The access time in each level is:

- $t_c = 1\,\text{ns}$
- $t_p = 10\,\text{ns}$ each byte
- $t_d = 10\,\text{ms}$ (for any block size)

The average hit rate in each level is:

- $A_c = 0.99 \Rightarrow 99\%$
- $A_p = 0.9999 \Rightarrow 99.99\%$

Cache block size (B_c) is 64 bytes

Which is the average access time to the memory system (t_{cpd})?

$$t_{cpd} = \underbrace{A_c \cdot t_c}_{\text{cache hit}} + \underbrace{(1 - A_c) \times \left[\underbrace{A_p \cdot t_p \cdot B_c}_{\text{main memory hit}} + \underbrace{(1 - A_p) \cdot t_d}_{\text{main memory miss}} \right]}_{\text{cache miss}}$$