CSC 212

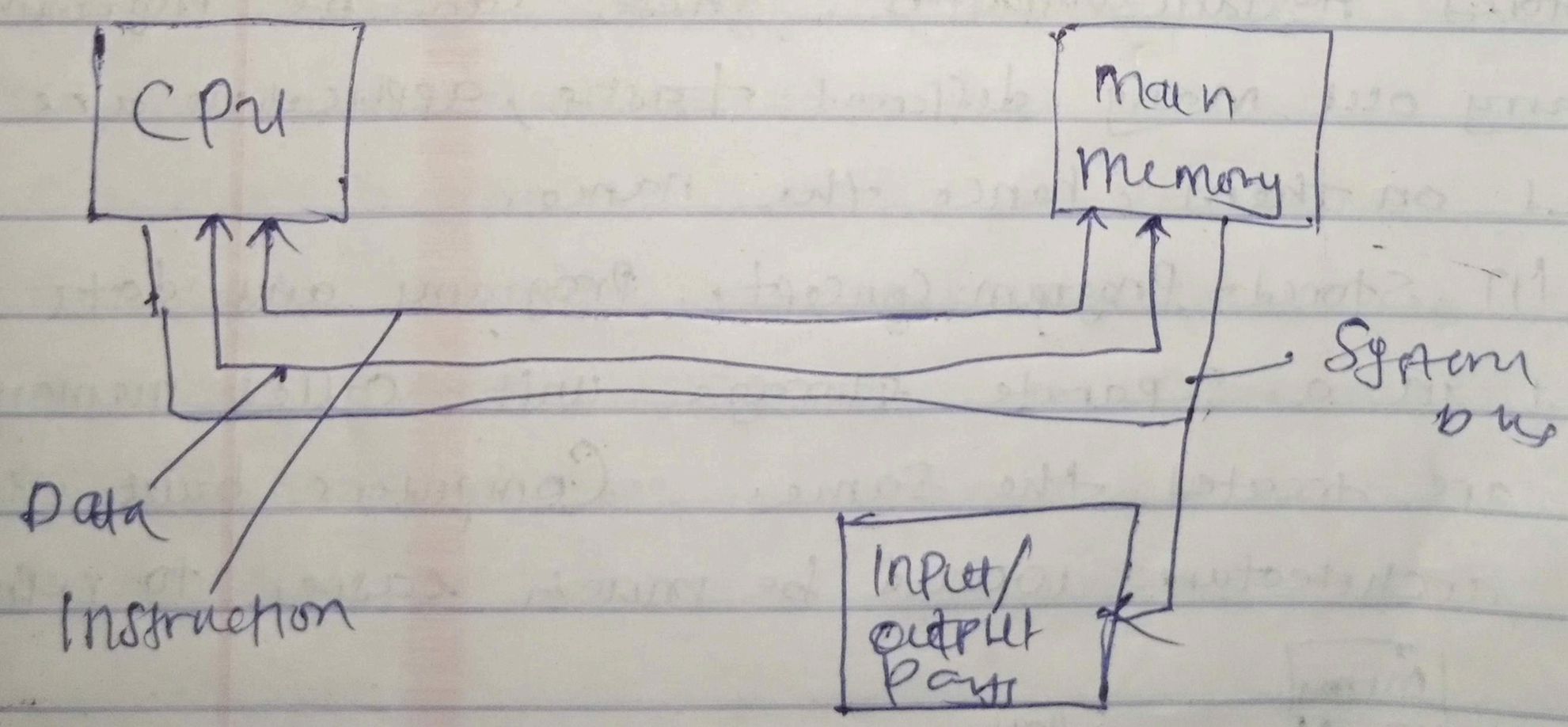
2 units

COMPUTER ARCHITECTURE

Course outline

* Level of machine design; gates, registers and processor levels.
* Number systems and computer Arithmetic, CPU design, instruction sets, von neumann architecture, multiplication and division algorithms and implementation, floating point processors. Parallelism, multiprocessor etc
Control unit design, hardwired & microprogrammed control.
* Memory Design, hierarchical memory design, Cache, associative and interleaved memory.

An Overview of the Computer System



Interfaces Between CPU & Main Memory

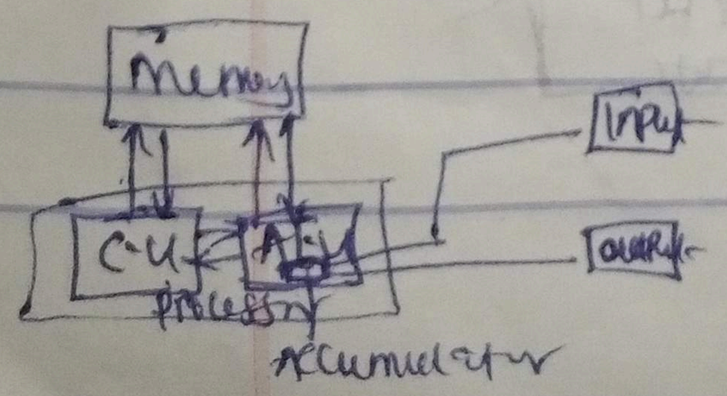
* Computer Architecture refers to the whole of the computer including the instruction set, computer organization and hardware.

~~Exam~~
* Von Neumann Stored Program Concept
This states that all instructions and data must first reside in the memory before they are acted upon.

NOTES

Historically there have been 2 types of computers
① Fixed Program Computers: Their function is very specific and they couldn't be reprogrammed e.g calculator
② Stored Program computers: These can be programmed to carry out many different tasks, applications are stored on them, hence the name.

In stored-program concept, programs and data are stored in a separate storage unit called memories and are treated the same. Computers built with this architecture would be much easier to reprogram



Source:
geeksforgeeks.org
computer organization

HP Pavilion g series

DEFINITIONS

(1) Register: Keeps track of every operation that takes place in the C.P.U

(2) Control Unit: Responsible for fetching instruction from the memory and stores it in the register, when the job is finished, it carries it back to the main memory.

Exam

* Conversion from one Base To Another.

PIPELINING. [A technique for improving processor performance.]

* Pipelining allows a processor to overlap the execution of several instructions so that more instructions can be executed in the same period of time.

Pipelining is a technique for overlapping the execution of several instructions to reduce the execution time of a set of instructions.

Terms in Pipelining: (1) Latency (2) Throughput.

i) Latency: The amount of time that a single operation takes to execute. expressed in seconds

ii) Throughput: The rate at which operations get executed, expressed as operations/second or operations/cycle.

In a nonpipelined processor throughput = 1/Latency.

In a piplined processor throughput > 1/Latency

Bandwidth = Throughput X amount of data transferred

SOLVED PROBLEMS

① Why does Pipelining improve performance?

In an unpipelined processor, each instruction is executed completely before execution of the next instruction begins. In a pipelined processor, instruction execution is divided into stages, and execution of the next instruction starts as soon as the current instruction has completed the first stage. This increases the rate at which execution of instruction is being performed thereby improving performance.

② What are the limits on how much a processor's performance can be improved using pipelining

There are two main limitations.

① As the number of pipeline stages increases, the fraction of the latency of each stage that is due to the pipeline latch increases.

ii) Data dependencies and branch delays. Instructions that depend on the results of other instructions have to wait for those instructions to complete, creating pipeline stalls (bubbles), and instructions that follow branches have to wait for the branch to complete so that the processor knows which instruction to execute next.

Page 135 ⟶ Computer Architecture Textbook.

HP Pavilion g series

MEMORY SYSTEMS

Latency, Throughput, and Bandwidth.

Latency: The time taken to complete an individual operate

Throughput: The rate at which operations can be completed.

Bandwidth: The total rate at which data can be moved between the processor and the memory system.

Bandwidth = throughput * Amount of data referenced by each memory operation

* PIPELINING, PARALLELISM, AND PRECHARGING

Example 1: What is the bandwidth of a memory system with a latency of 40ns, that transfers 1 byte operation, and is pipelined to allow 4 operations to overlap execution (assume no pipeline overhead)!

Solution

$$\text{Latency} = \frac{40ns}{4} = 10ns$$

$$\text{Throughput} = \frac{1}{\text{Latency}} = \frac{1}{10ns} = \frac{1}{10 \times 10^{-9}} = 10^8$$

Bandwidth = throughput × Amount of data transferred

$$= 10^8 \times 1 = 10^8 \text{ bytes/s}$$

Example 2). What is the bandwidth of a memory system that has a latency of 20ns, a precharge time of 5ns, and transfer 2 bytes of data per access

Solution

Latency = 20ns, Precharge time = 5ns

Latency (new) = 20 + 5 = 25ns

$$\text{throughput} = \frac{1}{25 \times 10^{-9}} = \frac{10^9}{25} = 4 \times 10^7 \text{ operation/s}$$

$$\text{Bandwidth} = 4 \times 10^7 \times 2$$
$$= 8 \times 10^7 \text{ bytes/s}$$

MEMORY HIERARCHY

A way of keeping the data that will be referenced most by a program in the top level of the hierarchy, so that most memory requests can be handled by the top level or levels.

* Memory hierarchy is arranging different kinds of storage present on a computing device based on speed of access

SRAM AND DRAM

SRAM Advantage	DRAM Advantages
- fast read & write speed	- High density
- Low Power	- fast read & write speed
SRAM Disadvantage	DRAM Disadvantage
- Low density	- volatile
- volatile	- High Power

TERMINOLOGIES IN MEMORY HIERARCHY

Hit:- When the address that an operation references is found in a level of the memory hierarchy

Miss: When the address is not found.

Hit Rate: This is the percentage of references that reach the level that result in hits

Miss Rate: Percentage of references that reach the level that results in misses

Hit Rate + Miss Rate = 100%

In Probability Hit Rate + Miss Rate = 1

Access Time: Time to complete a request that hits

Average Access Time (AAT) $= (T_{hit} \times P_{hit}) + (T_{miss} \times P_{miss})$

$T_{hit} \Rightarrow$ Time to resolve requests that hit in the level

$P_{hit} \Rightarrow$ Hit Rate of the level (expressed as Probability)

$T_{miss} \Rightarrow$ AAT of the levels below the one in the hierarchy

$P_{miss} \Rightarrow$ Miss Rate of the level

Example: If a level of the memory hierarchy has a hit rate of 75%, memory requests take 12ns to complete if they hit in the level, and memory requests that miss in the level take 100ns to complete, what is the average access time of the level?

Solution

$P_{hit} = 75\% = 0.75$

$P_{miss} = 1 - P_{hit} = 1 - 0.75 = 0.25$

$T_{hit} = 12ns$
$T_{miss} = 100ns$

Average Access Time $= (P_{hit} \times T_{hit}) + (P_{miss} \times T_{miss})$
$= (0.75 \times 12) + (0.25 \times 100$
$9 + 25$
$= 34ns$

Example:- In a two-level memory hierarchy, if the top level has an access time of 9ns and the bottom level has an access time of 60ns, what is the hit rate in the top level required to give an average access time of 10ns.

Solution

Average Access Time $= (P_{hit} \times T_{hit}) + (P_{miss} \times T_{miss}) = 10ns$

$P_{hit} = ?$ recall $P_{hit} + P_{miss} = 1$

$P_{miss} = 1 - P_{hit}$

$T_{hit} = 9ns$ $T_{miss} = 60ns$

Average Access Time $= (P_{hit} \times 9ns) + ((1 - P_{hit}) \times 60ns)$

$10 = 9P_{hit} + 60 - 60P_{hit}$

$10 - 60 = -51P_{hit}$

$-50 = -51P_{hit}$

$P_{hit} = 50/51 = 0.98$

hit rate $= 0.98 \times 100$

$= 98\%$

Example: A memory system contains a cache, a main memory, and a virtual memory.

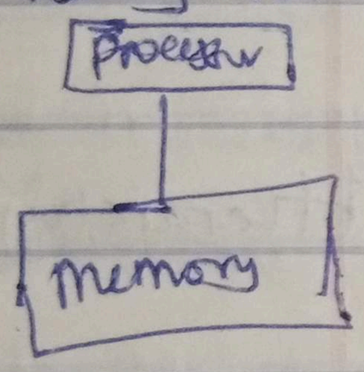
	Access Time	Hit rate	
Cache	5ns	80%	= 0.80
main memory	100ns	99.5%	= 0.995
Virtual memory	10ms	100%	= 1

What is the average access time of the hierarchy?

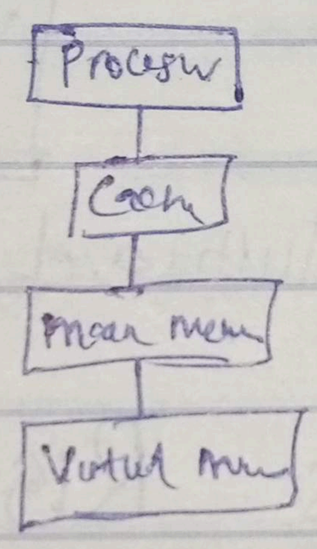
Solution

Starting from the bottom of the hierarchy and going up.

single level memory hierarchy



multilevel memory hierarchy



Average Access time for refuest that reach the main memory $= (T_{hit} \times P_{hit}) + (T_{miss} \times P_{miss})$

$= (100ns \times 0.995) + (10ms \times (1-0.995))$

$= 9.95 \times 10^{-8} + 5 \times 10^{-5}$

$= 5.00995 \times 10^{-5}$

Converting to ns $= 5.00995 \times 10^{-5} / 10^{-9}$

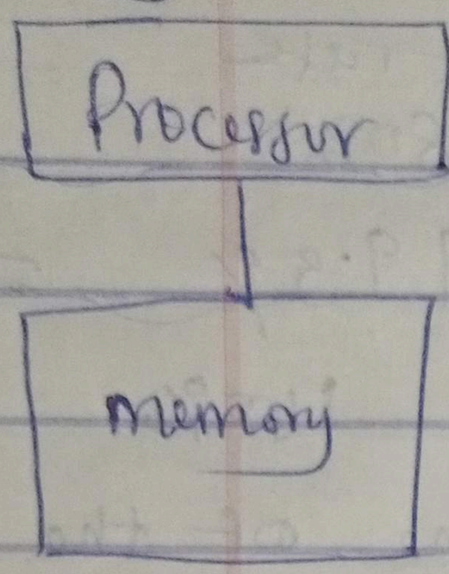
$= 50099.5 ns$

Average Access time for refuest that reach the Cache (which is all refuests) $= (5 \times 0.80) + (50099.5 \times 0.20)$

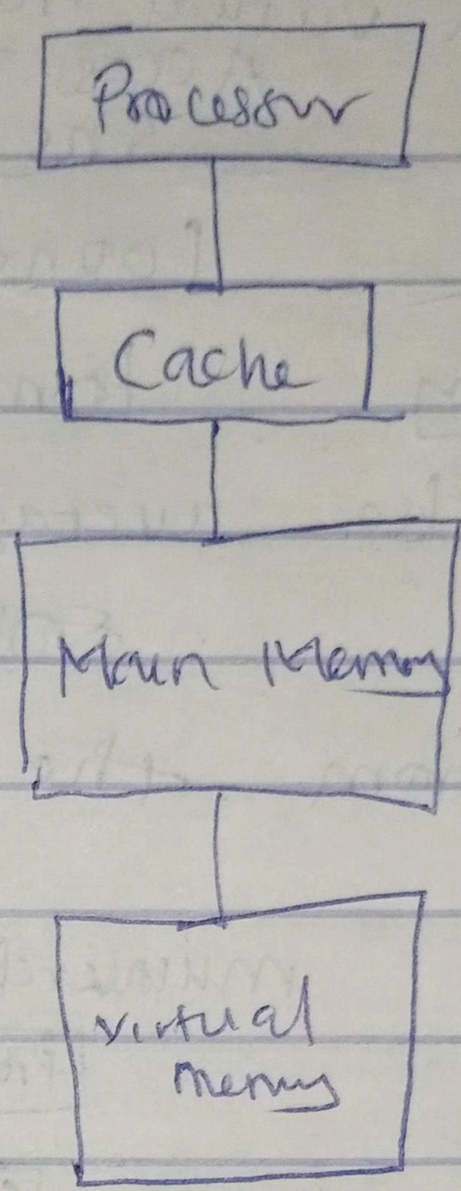
$= 10023.9$

$\simeq 10.024 ns$

Memory Hierarchies



Single-level memory sys



Multilevel memory Hierarchy

CISC Vs RISC

CISC → Complex instruction set Computer
↳ developed to make Compiler development easier & simpler
They are chips that are easy to program that makes efficient use of memory.

RISC → Reduced Instruction Set Computers
It is a microprocessor that is designed to perform smaller number of computer instruction so that it can operate at a higher speed.

* Characteristics of CISC

i) A large number of instruction
ii) Instruction-decoding logic will be complex
iii) A large variety of addressing modes
iv) Instruction may take more than a single clock cycle to get executed.

Characteristics of RISC

i) Simpler instruction decoding
ii) One instruction per cycle
iii) Simple Addressing mode
iv) Fewer Data types
v) Register-to-register operation
vi) Instruction execution would be faster
vii) Simple instruction format
viii) Smaller programs.

DIFFERENCES BETWEEN CISC AND RISC

CISC	RISC
i) It has a microprogramming unit	i) It has a hardwired unit of programming
ii) Only single register set	ii) Multiple register set are present
iii) They are mostly less or not pipelines	iii) Highly pipelined
iv) Execution time is very high	iv) Execution time is very less
v) Code Expansion is not a problem	v) May create a problem
vi) It requires external memory for calculations	vi) Doesn't require external memory for calculations.
Example: AMD, Intel X86 CPUs	Example: ARM, SPARC
vii) Limited addressing modes	vii) Compound addressing modes
viii) Decoding of instructions is complex	viii) The decoding of instructions is simple

SRAM Vs DRAM

SRAM	DRAM
i) It does not need to be refreshed periodically	ii) It needs to be refreshed periodically when in use
ii) More expensive	ii) Less expensive
iii) It structure is complex than DRAM	iii) It structure is simple

iv) SRAM are faster than DRAM	iv) DRAM are slower than SRAM
v) Used in cache memory	v) uses in main memory
vi) Low density devices	vi) high density
vii) It has less storage capacity	viii) Large storage capacity
viii) Consume less power than DRAM	vii) Consume more power than SRAM

(non-volatile)

ROM ⟶ Read only Memory

↳ Stores crucial information essential to operate the system, like the program essential to boot the computer

Types ⟶ MROM, PROM, EPROM and EEPROM

*MROM :- Mask ROM : A kind of ROM, that is masked off at the time of production. This means users cannot change the data stored in it.

*PROM :- Programmable ROM : It can be programmed by the user. Once programmed, the the data and the instructions in it cannot be changed.

*EPROM : Erasable Programmable ROM :- It can be reprogrammed. To erase data from it, expose it to ultraviolet light. To reprogram it, erase all the previous data.

*EEPROM : Electrically Erasable Programmable ROM :- The data can be erased by applying an electric field, with no need for ultraviolet light. We can erase only portions of the chip.