CSSE - 503

Computer Architecture

Date 24 Feb - 2021

Recommended Book

- Computer Organization & Architecture by William Stallings (10th Edition)

Course Outline :- (Part I)

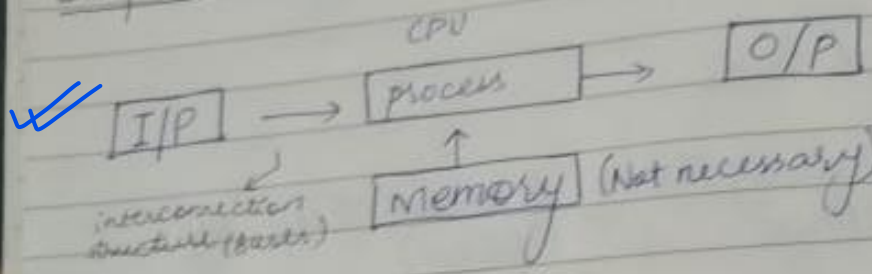
* Overview
  - Organization & Architecture
  - Evolution
* Computer System
  - Interconnection Structures
    - Bus interconnection
    - Point-to-point interconnection
  - Memory Systems
    - Cache Memory
    - Internal Memory
* CPU
  - Computer Arithmetics
  - Instruction set
  - Addressing mode
  - Instruction Pipelining

Part II :-

* Instruction level Parrallelism
* Micro - operations
* Parallel Processing

Yeh hm abi parenge kyun ke ye advance hai

Computer Architecture :-



Stored Program :-

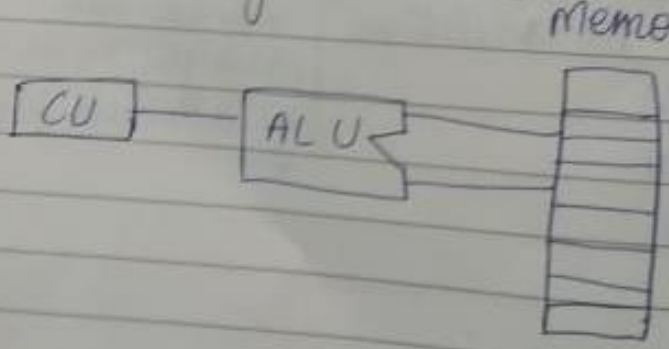
Memory is required It is special type of computer that require memory

+ Normally memory is not compulsory in computer.

Interconnection Structure :-

Is this ~~situation~~ ko join krne k liye hardware technique use kren

Computer Organization :-



Date

→ koi bhi system bnana ho toh bi Requirement Analysis se shuru hogi.

→ jisse computer use krna hai use jo Programmer cheezen effect krte wo computer architecture hai.

→ computer architecture nhi bnaege tb tk hardware/design ki traf nhi jana.
→ jo sari cheezen programmer ko visible hon wo computer architecture hai.

problem

Computer Architecture :-

- It refer to those attributes of a system which are visible to the programmer

- Those attributes which have a direct impact on the logical execution of the program

For example :-
Instruction set, number of bits used to represent various data types, I/O mechanism etc.

solution

Computer Organization

- It refers to the operational units & their interconnections that realize the architectural specifications.
- The organizational attributes are those hardware features or details that are transparent to the programmer.

For example :-
control signal, Interface between the computer & peripherals, the memory technology use data.

- It is architectural issue whether a computer will have a multiply instruction or not. - An architecture may survive for long.	- It is an organizational issue whether a multiply operation is performed by a special multiply unit or by repeatative use of adder circuit. - An organization may change more frequently.

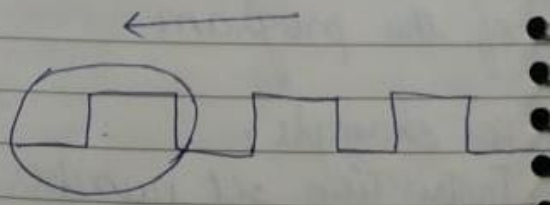
→ Ek pe chalne wala program dusre pe chal rha tou architecture same hai warna different hai.

Flip Flop: (Latch)
- NAND / NOR gates
- semi-conductor Memory

It is made up of gates.

→ Is magsoos

Clock rate:-
Instruction / second



- Clock ki kitni cycles guzrengi.
- signals ko valid tasawur krne k lye kitna time chahye.

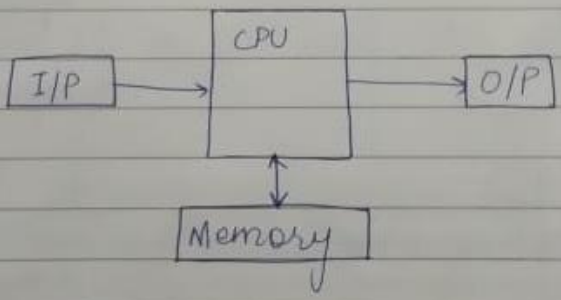
Date

Clock rate is the basic feature to improve performance

Lecture # 02 3/March/21

Evolution :-

- Increase in processor performance
- Decrease in component size
- Increase in memory capacity
- Increase in I/O performance and memory



* Most of the time system is idle b/c it has processing capability but less or no. software stream

* main memory works in:
* Clock frequency determines the performance
* Objective of designer is to keep a balance in performance of different units.

How processor's performance is increased in recent days?

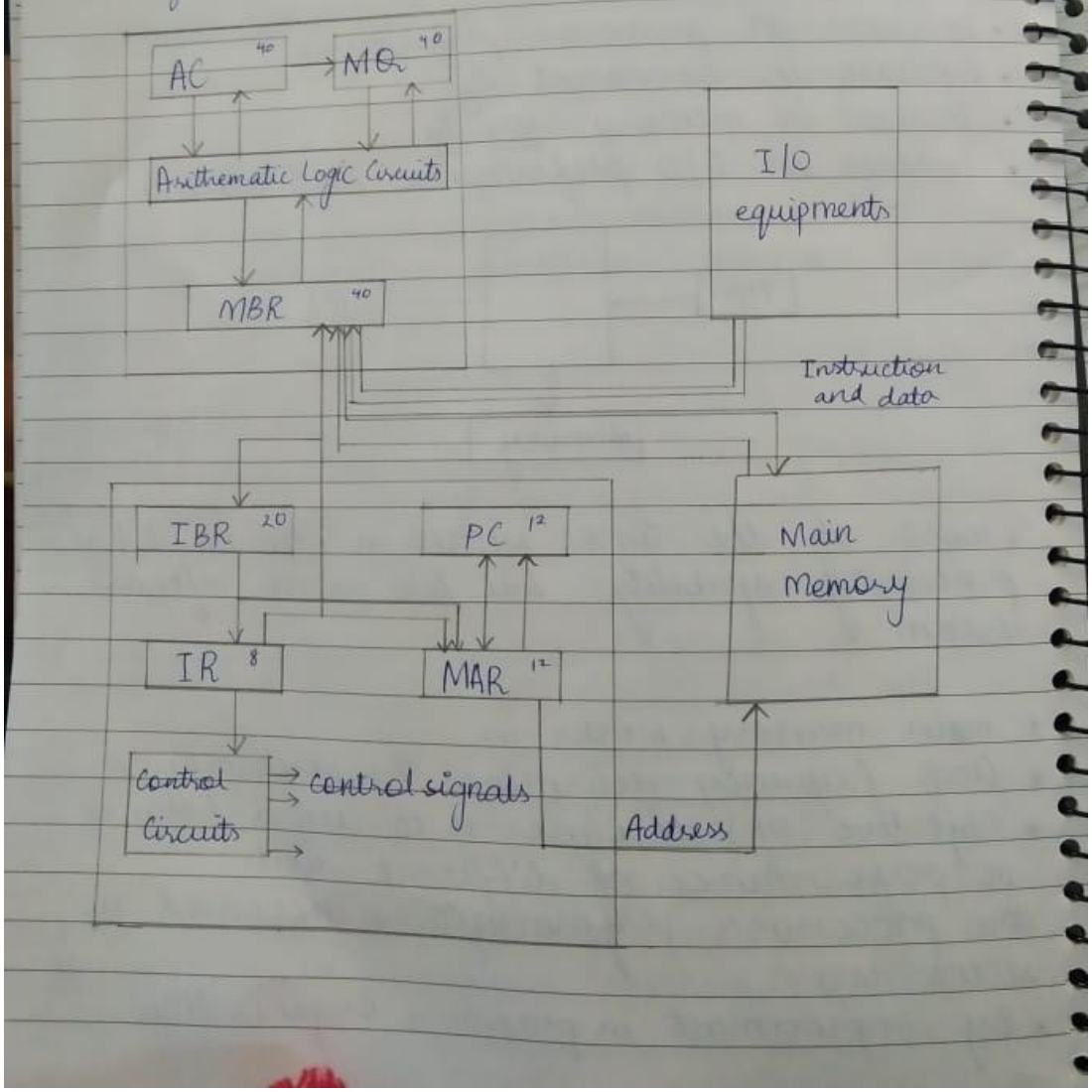
- By improvement in processor's Organization.

speculative execution :- prediction of next instruction next instruction konsi fetch hogi pehle se pta krna

IR

Date:

By heavy use of pipelining, use of parallel execution techniques, speculative execution technique, etc.
Now, the focus is to keep the processor busy as much of the time as possible.



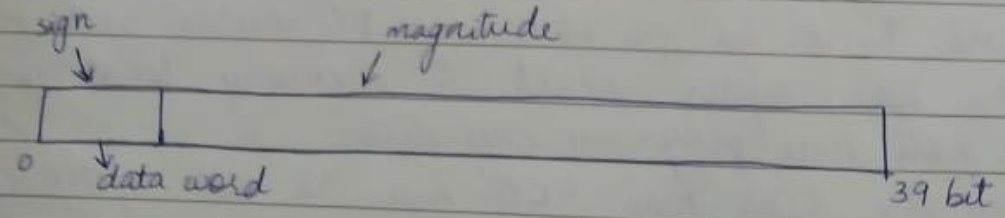
IAS (Von-Neumann) Architecture (1945 - 52)

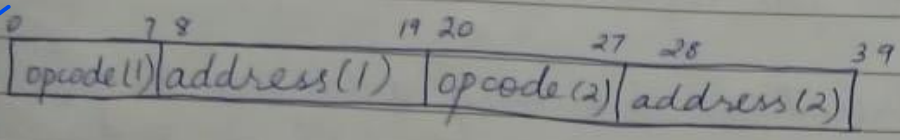
IAS computer had 21 total instructions, divided into 5 groups

i) data transfer instruction
ii) Unconditional branch instruction
iii) Conditional
iv) Arithmatic instruction
v) Address modify instruction

Refer to table 11 in the book for 21 instructions (Pg 16) 10th edition

20 / instruction data transfer ki instruction hoti hai or 20 / arithmatic





instruction word

IAS provides flexibility to programmers. Warna rigid programs bante hain.

Date:

operands ko identify karna comes in addressing mode

i) Immediate Addressing mode

ii) direct addressing mode

iii) register addressing mode

etc.

Instruction ko btana hota hai k result ko kahan store karna hai (result)

System start karte hi pc btata hai PC shuru se shure hota hai after reset. 0 memory location se shuru hoti hai program execution.
Yeh 0 memory ROM hai hoti hai. 32 bit system mai 4GB RAM jab lagate the tou 3.8 kuch show hoti hai.
Har ek instruction ka alag address hoga memory mai, 32 bit system mai 2^{32} se zada memory nahi lga sakte. 32 bit mein 8 GB bhi lga den to 3.8 hi accessible hogi jo thori bht remaining hai jese 3.8 hai 32 bit mai

$2^{32} \rightarrow$ 4 GB
$2^{8} \times 2^{8} \times 2^{8} \times 2^{8}$

$2^{48} \rightarrow$ 256 TB memory can be attached

Date:

toh wo dusri memory ko assign hojati hai jese ROM kyun address mojood nhi hota allocated memory 64 bit system mein data 64 ka hota hai magr address 48 ki hoti hai.

Von Neumann flow :- M. ± Reg

PC mai abhi 0 hai, PC ka size hai 2 bit PC se MAR mai jayega wahan se main memory mai. Control signals bataenge k yahan se read karni hai memory. Read/write control circuitry batati hai Read ko k MBR M. Buffer mai jayega Pehli memory (1) jab system shuru hota hai usko instruction word smjha jata hai. Instruction word ki pehli instruction ko opcode IR mai or address MAR mai chala jayega Pehli wali memory over write hojayegi Dusri instruction IBR mai jayegi jb tk instruction word ki pehli instruction execution complete naa karle IBR se dusri instruction 2 hisse mein divide ho k 8 bit ka opcode or 12 bit ka address. IR decodes instruction. Control circuitry interpret the opcode and executes the instruction by sending out appropriate control signals to cause data to be moved or an operation to be performed by ALU

Date

IAS Architecture:

All computers of today have Von-Neumann architecture which makes the following by concepts.

1) Data and instructions are stored in a single read/write memory
2) The contents of this memory are addressable by location without regard to the type of data contain there
3) Execution occurs in a sequential fashion (until explicitly modified) from one instruction to the next.

An IAS architecture includes:

i) A main memory which stores both data & instruction in the same form No separate area were reserved for the two This resulted in many advantages one of being that the machine could modify it just as it did for the data.

- An arithematic & logic unit (ALU) capable of operating on binary data.
- A control unit (CU), which interprets the instructions in memory and causes them to be executed
- Input & output equipments operated by control units & recording medium.

Date

→ Jo bhi instruction hm sbse phele run krna chahte hain tou hm assume krte hain k 0 memory location pe hai.

→ Agr 0 memory location pe agr

→ Ek instruction complete tb hoti hai -

→ Result ko accumulator register pe store kiya jayega. ye bt implied hai.

→ Programmer can modify the instructions during the execution of program.

Lecture #03

Date: 10- March -2021

Interconnection structures:

1) Its purpose is to connect different components of computer
2) It cannot be connected without hardware
3) Our computer is the network of sub-units that are integrated inside computer system
4) Communication is the contact between two computer components, the means of communication is not specific
5) Our computer is divided into 3 modules,

→ i) I/O modules

I/O is slow. It can't be connected to processors directly. I/O ports are used to connect them to processors I/O ports are used to provide buffer area.

→ ii) Processors (can be more than 1)

Processor's performance is more than I/O & memory modules.

→ iii) Memory

storage device / location

be equal in size

location should be uniq

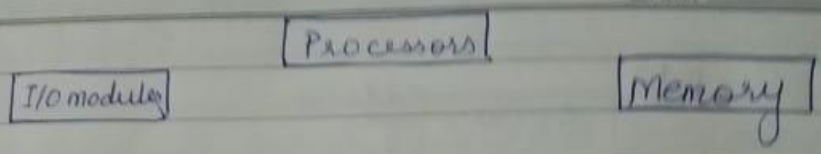
Buffers :- something b/w two things that facilitates both of them

Temporary memory

↳ cache's purpose is same as buffer

↳ Register's
Latch's

Date:



* Buffer/cache/register/latch are terms that are used for intermediate memory.
* The type of memory that is being used depends on natured of communication.
The type of data that is being transferred

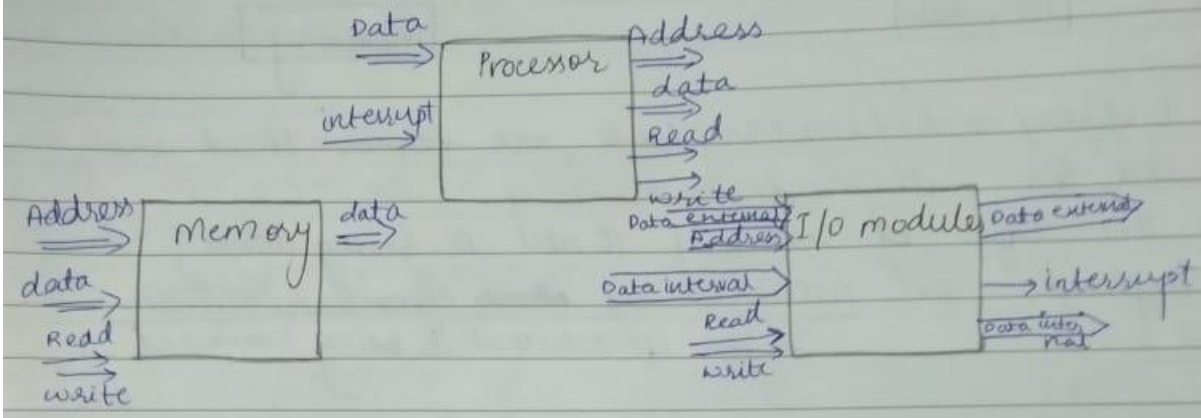
iii) Memory

•) location address should be unique because we need to tell the system even if we are sending a single bit that where we want to write the data.

•) We need address of location, where we want to perform operation i e read and write.

•) Memory should have lines that helps in reading/writing of data on the said address.

Date:



ii) Processor

- Address retrieval through fetch instruction.
- Microprocessor should have lines for sending data
- If microprocessor wants sends data then it should have lines for sending data
- If microprocessor wants to ~~send~~ receive data then it should have lines for receiving data.
- We need communication structure for bringing data such as buses or any other interconnection structures.

Lines — buses / interconnection structures

* Interrupt :- (alt + ctrl + del → basic interrupt)
To stop OS. Interrupt is performed by user through hardware
* I/O module transfers interrupt to processor

Date:

1) I/O Module.

•) If we need to input or output something then we need address of it.
•) Processor tells the address of location that from where to receive or send data.

1 Interrupt driven I/O operation

•) whenever a device wants to input/output operations then it sends an interrupt signal to processor.

•) I/O devices are very slow.

2 •) Another way of IO interruption is DMA

Third type of I/O operation is DMA I/O.

•) It performs input/output operation without interrupting processor.
•) The data that is received from the processor to the I/O module is called data internal.
•) The data that is received from the external device is called data external.

Date:

Bus Interconnection as Multiple Bus Interconnection :

- Bus is a communication mechanism / device
- Bus is a public transport.
- It is shared between whoever want to use it for travelling.
- Wait if overloaded
- Bus is a shared transmission medium that is used to connect different devices.
- The ~~public~~ shared transmission of bus is sometimes considered as a problem b/c it may be occupied.
- Ek bus pe ek communicate kregi bagi sb liten kerege.

- It should be able to transfer address/I/O operation and I/O data
- Every one will check if the address is related to them or not.
- We use multiple buses for interconnection so if two components are communicating and other two are not so we use a bus for them so they also communicate.

Date:

Buses:-
Collection of lines.

- It is use to transfer signals one line transfer one bit only.
- We should have lines for transferring address, control signals and data.
- Purpose of every line is already defined
- A 32 bit system will at least have 50 lines.
- One line can be assigned two functions

Multiplexing:
Defining multi-purpose of a single line

System Bus:
That connects processor and major components of a compiter system.

- Address bus, data bus and control bus are the functionalities of a bus
- Every bus have address lines, control lines and data lines
- Lines of data bus tell that how many bits of data can we transfer at a time
- The size of bus and register can be different but it is tried to make it equal

Date:

•) Example of a bus is USB, SATA, IDE cable, VGA cable, PH2, SCSI

•) use to connect external devices / hardware

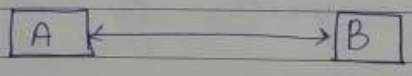
•) Every wire that is connecting two o components are considered as interconnection structure in computer system.

Timing signal – clock
command – I/O instruction

•) Slow devices are only held in expansion bus
•) Fast devices " " " high-speed bus

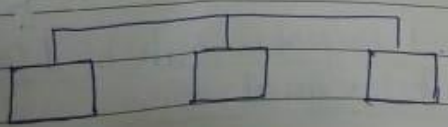
Dedication :-

i) Physical dedication.



It is physically dedicated b/w these two modules. It will face latency when if B is transferring, A cannot transfer, it can only receive

ii) Functional :-



Date:

Functional dedication i.e jo function set kia wa hai srf wahi perform hoga.

Another is multiplex.

Method of Arbitration :-

Arbitation :- جگڑے میں صلح کرانے کی کوشش کرنا

- Here fight is of transmission of data, who will transfer first.

- Centralized : single Hardware device is bus controller/ Master.
- Distributed : Each Module contain access control logic.
- Every bus should have equal access/sight.

Data Transfer Type :-

1) Simple Read / write operation

Advance Data Transfer Type :-

ii) Read - Modify - Write
iii) Read - After - Write
iv) Block Transfer.

- Here we are using bus as active component
- DMA ka kaam bus se kra rhe hain

• reference from given slide

Date:

Read - Modify - Write :-
100 instructions for Reading
" " " writing
" " " modifying

Lecture # 04

Summary of previous class:-

Bus :- function of a specific line

Control line depend kiti hain specific type of bus pe

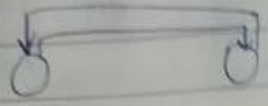
- To external devices connect kr rhe hain wo bhi bus hain or jo internally connect kr rhe hain wo bhi bus hain

- Bus ko access krne k lye har component k ps bus ki logic honi chahye.

New lecture :-

Point to Point Interconnection with packetized Data Transfer :-

Two components will be connected and both will transmit and receive simultaneously. Every interconnection structure where both components can transmit and receive at the same time



Computer :- Network of processors, memories, I/O modules
↳ system

Packetized Data Transfer :-
Introduced in 2002/2003 → 1st point-to-point interconnection

Network :- Connected components

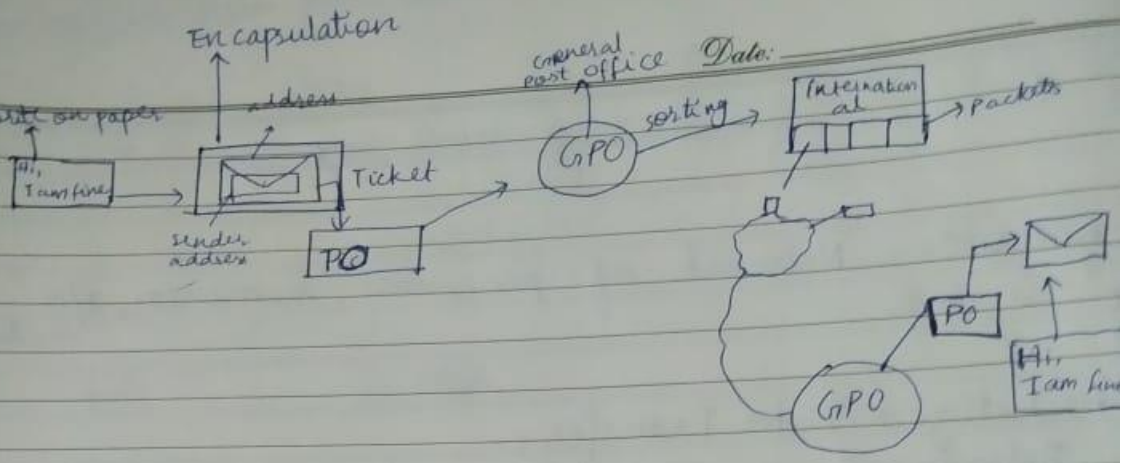
Computer Network :- i) Establishing • communication b/w components

OSI Model :-

* Taking communication b/w two devices as task
* This task is ~~broke~~ divided into small boxes.

Postal Mail Network :-

* Write letter on papers.
* Then put it into an envelope
* Put address on envelope
* & also put ticket



Date:

* It is a task of communication.
* We divide this task into small sub-task so that communication won't be affected by the changes happening in other sub-tasks.

* The purpose of breaking tasks of communication into sub-tasks (layers) is that every sub-task has their own protocols. And these protocols won't affect the initial process and final process but only throughout the transfer of communication, it will affect. Protocols of one sub-task won't affect other sub-tasks.

* Network layer pe packets ka transfer hota hai
* Data link layer pr frame ki information hoti hai.
* raw bit string mai data transfer krna.
* Here packetized data will be transferred.
* Data key (Address key)

raw data

Date:

Package :-
raw additional information of address with raw data

* Data is encapsulated.

Point to Point Interconnection :-
* Compared to the shared bus, the point to point interconnection has lower latency & higher data rate.

Quick Path Interconnect (QPI) :-
→ Processors ki cores ko apas mai connect krna ← use
2008 → new point to point structure → ultra path (UPI) Interconnection
* QPI k baad UPI aya tha.
PCI → interconnection of I/O devices.

Multi direct connection. Multiple components har direct connection to other components. No need for arbitation.

Layered Protocol Architecture :- (Explanation)
QPI processor level interconnect use a layered protocol architecture, rather than use of simple control signals. layers were.
Protocol: Set of rules for exchanging packets of data between devices.
Routing: Directing the packets through the fabric
link :- Unit of transfer in 80 bit flit (flow control unit)

Routing → upar se aaney wale packet ki control header ko dekhe k aagey wale packet ko ks API layer ki traf bhejna hai ussk hawale ksdegi

Link → divide into 80-80 bits packets ko.

Flowcontrol → Maintain sequence arrangements of packets

~~Physical → divide flits into 20 20bits unit (phit)~~

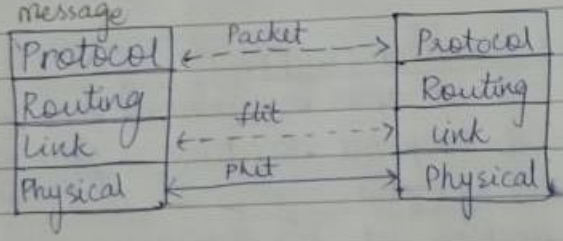
* Transference from one API interface to another.

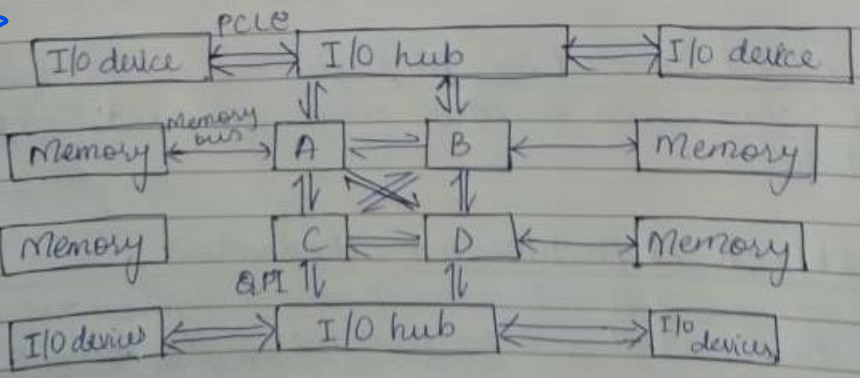
Proto col → create packets

Physical:- Unit of transfer in 20 bits phits (physical units)

iii) Packetized Data Transfer:-

Data are not sent as a raw bit stream, rather sent as a sequence of packets which include control header & error control code.

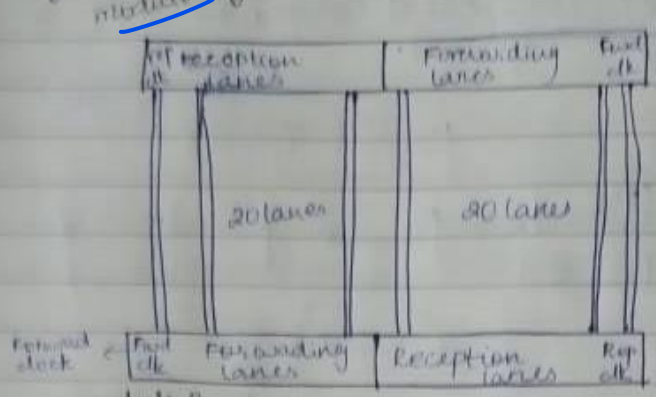




* Sari cores (A, B, C, D) as cores I/O hub k sth bhi 84 lines k zarye connected hain
Total lines = 84

Lane :- pair of two lines

Physical layer :



Total lanes = 42

* We are not using lines because when we are sending a bit and it is received then it will come from different line.
* If the sequence in which we are sending is received then it is valid otherwise it is not.

Why we donot connect memory with interconnection lanes? Why we use lines?

* Because we donot want to allow multiple acess to memory. The reason of not allowing is if we allow both accesses (write, read), data inconsistency will occur.

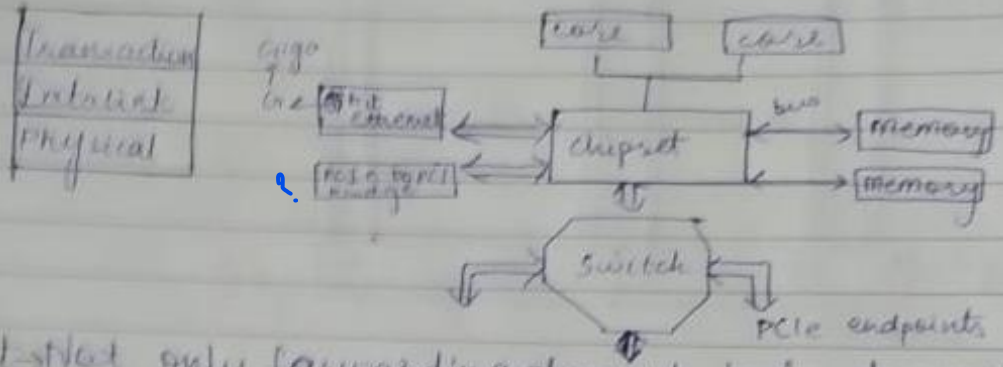
* under certain conditions we allow multiple access for memory.

* PCI ko fsb (front side bus) se connect kiya gya tha

Date

PCI express (PCIe): (Peripheral component Interconnect Express)

* Introduced in 2004



Chipset → Not only forwarding element but also perform some functionality

* core ko bhi information deg chipset ko degi tou wo raw bit stream ki form mai hoga.
* Packet decapsulation krta hai

Lecture #05 Date: 24 March 21

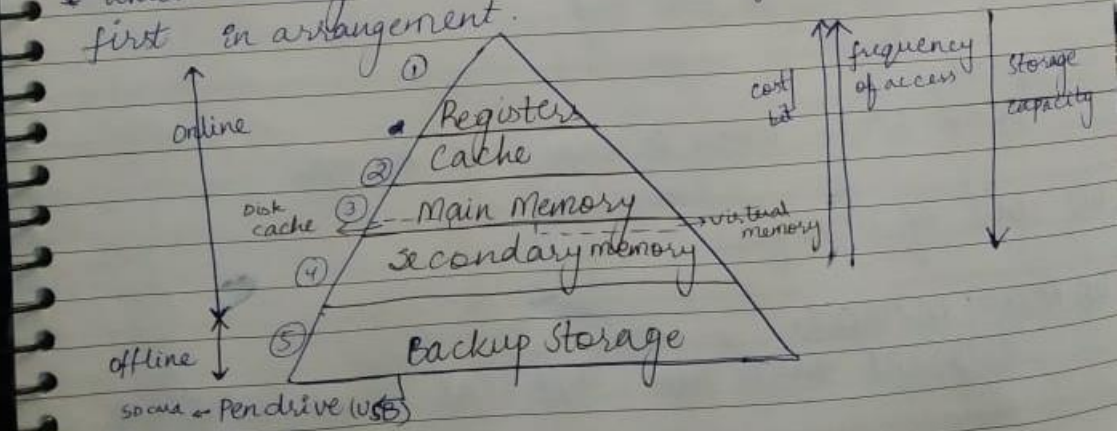
Computer Memory system:-

* stored program introduced the concept of memory.
* Instruction and Data will be stored together in Vonn Neumann architecture.
* Memory location is of equal size.
* Every storage location has unique address

Types of Memory in a system:-

- Virtual Memory
- RAM
- Cache
- ROM
- Register
- secondary memory → Hard disk
- Buffer

* Which memory is closer to memory will be first in arrangement.



* Harddisk is accessed by input/output module

Frequently Date

* Registers are the most accessed memory than any other
* Cache is less frequently accessed memory than Registers but greater than others.

* we use this hierarchy to lower the cost of the system.
* Cache are of different types and on different locations
* cost jese move krte hain upar ki traf toh cost brhti jati hai
* We use tape drive as backup to save cost.
* jese jese hm neeche move krte jate hain access time barhta jata hai

Key Characteristics of Computer Memory System:-

i) Location: Internal Memory
External Memory

ii) Capacity:-
Some times word is expressed in bytes

iii) Unit of transfer:-

iv) Access Method:-
Different devices have different access method

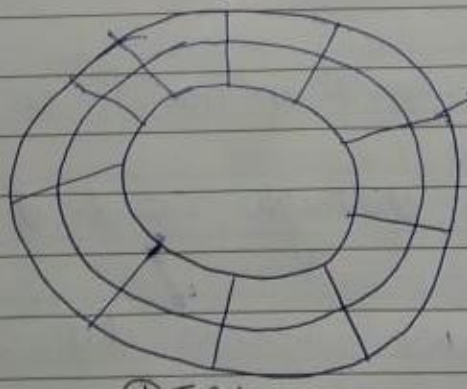
a) Linear Sequential Method
b) Direct Access Method
c) Random Access Method
d) Associative access method

c) Random Access Method:-

- Is access method jis mai ksi bhi storage location ko same time mai ksi bhi waqt access kiya jaa ske
- Mechanism of selection is created in such a way that we can access every storage location in same time
- Every storage should have read & write mechanism.
- This method is expensive.

b) Direct Access Method:-

- It is a combination of random and linear method
- This method is used in every disk storage (hard disk). Disk shape ki storage mai apply hoga.
- Format → ڈسک کو ترتیب دینا / arrange karna.



Date

* sector and tracks are used to format data.

storage location	sector

* It will maintain the record of on which storage location is saved in which sector.

* Allocation table mai sector and tracks will be stored.

* Floppy mai ek read write mechanism hota hai
* hard drive mai multiple read/write mechanism hote hain.

d) Associative Method:

* we are focusing on Von Neumann's second key content.
* It will access/search through content or a part of content.
* In cache memory, we access data with the help of content

Date:

Performance:-

1st parament:- location pe pohanchne (travel time) ka time.

* Access time can measure performance.

* Access time of internal and external memory is defined differently.

* External is accessed in the form of blocks.

[RAM system → Main memory, cache, Register
[Non RAM → secondary, backup

Non RAM:-

Access time is only the time of reaching to the location. It cannot fetch data.

RAM:-

Access time includes fetching data.

* data tk pohanchna or wapis le k ana.

* 2nd Parameter:-

Cycle time

* we can also measure performance from cycle time.

[cycle time = access time (t_a) + transcant time

* jska yeh time km hoga uss ki performance achi hogi.

3rd parameter :-
transfer rate

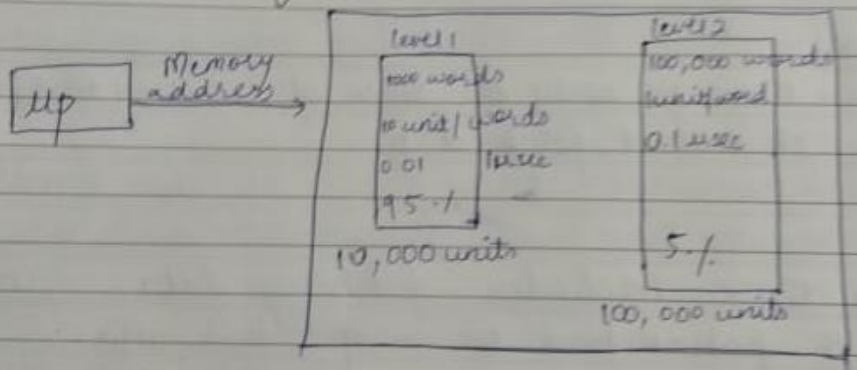
transfer rate = $\frac{1}{\text{cycle time}}$ or $\frac{1}{\text{access time}}$

* how much data access.

For Non-RAM system :-

$R = \frac{N}{T_n - T_a}$ → N random accesses (no arrangement)

Computer Memory system :-



* To top k level hote hain unki cost/bit ziada hoti hai

2 cheezen ki bunyadd pe compare krenge.

① Performance
② Cost

Total capacity = 101,000 words
Total cost = 110,000 units

Average Access time:-

Date

per unit cost (or cost/unit or cost/word) = Total cost / Total capacity

= 1.089 units/word

Average Access time = $\frac{95(0.01) + 5(0.1 + 0.01)}{100}$

= 0.015 μsec

* ye level 1 k qareeb hojaega. 0.01
frequency of access change krenge.

Average A.T = $\frac{50(0.01) + 50(0.1 + 0.01)}{100}$

= 0.06 μsec

* yeh zyada level 2 k qareeb agaya.

* Agr capacity vise bt karen tou yeh hona chahye

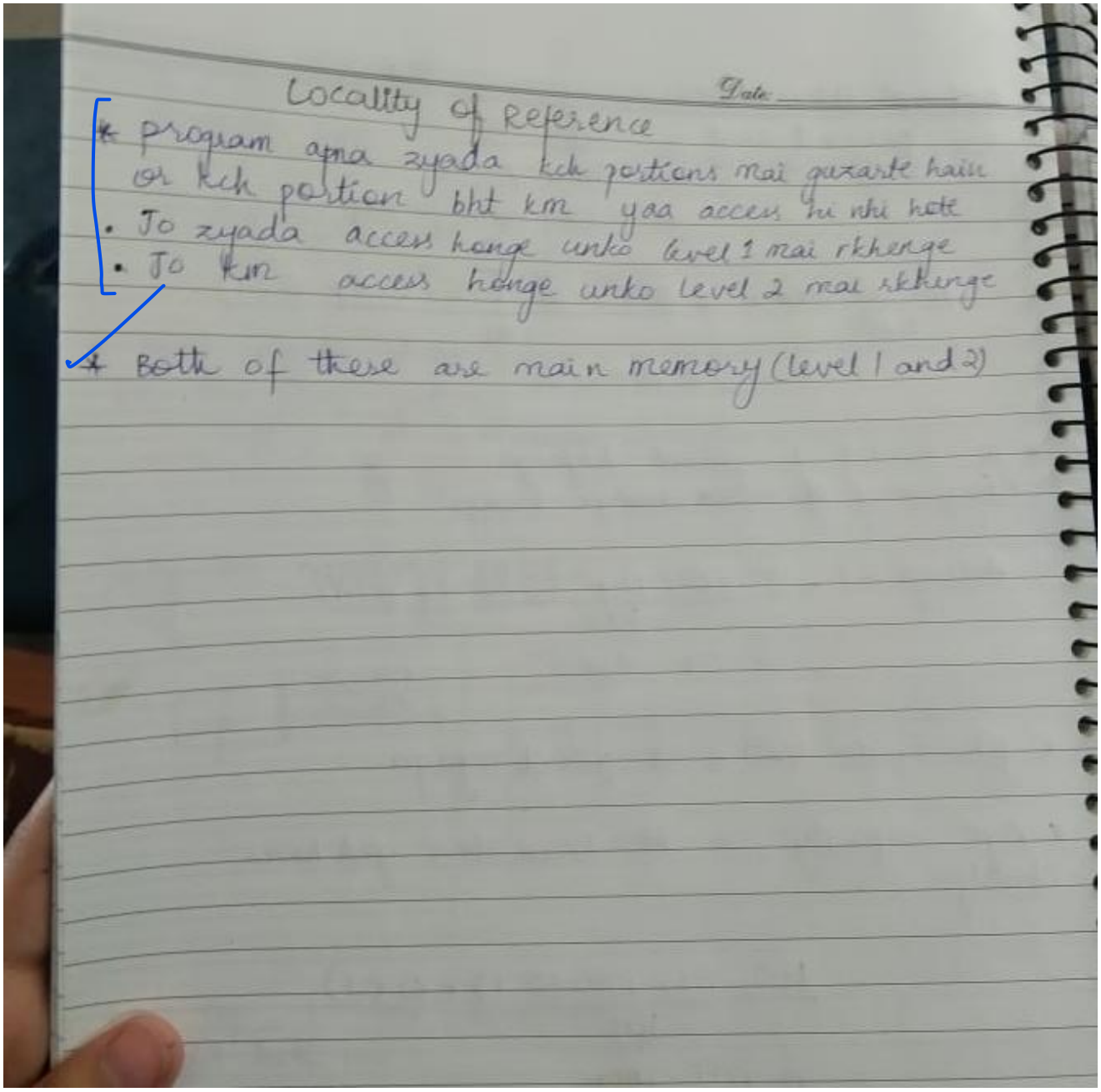
= $\frac{1(0.01) + 100(0.1 + 0.01)}{101}$

= 0.109 μsec

access time is increased from level 2.

* dono alag alag access pe honge tou pehle le pe jaenge

* Isa system ki cost km hogi or performance zyada hogi.



Locality of Reference

Date

* program apna zyada kch portions mai guzarte hain or kch portion bht km yaa access hi nhi hote
  - Jo zyada access honge unko level 1 mai rkhenge
  - Jo km access honge unko level 2 mai rkhenge

* Both of these are main memory (level 1 and 2)