BASIC CONCEPTS AND COMPUTER EVOLUTION

Date
Page

Computer Architecture	Computer Organisation
(I) Refers to those attributes of a system that are visible to the programmer and have a direct impact on logical execution of a program.	(I) Refers to the operation units and the interaction between them that achieve architectural specification.
(II) Deals with low level concepts.	(II) Deals with high level concepts
(III) Defines the system in abstract manner like what does the system do. "WHAT"	(III) Deals with realisation of abstract model like how to implement. "HOW"
Ex- types of instructions addresses modes memory.	Ex- Physical components like circuits, adder, substractor etc

- Computer architecture refers to the way that a computer system is designed. - It is concerned with the overall structure of a computer and how the components are arranged to provide the necessary functionality.
- Computer architecture is often described using terms such as instruction set, memory hierarchy and input/output system.
- For example: the architecture of a computer processor is concerned with the designed of the instruction set, the way that instructions are executed and the performance characteristic of the process.

On the other hand computer organisation is concerned with how the components of the computer system are implemented and how they interact with each other to provide the required personality. It deals with the specific details of how the components are arranged and how they communicate

with each other to perform the required operation.
For example - Computer organisation is concerned with the design of CPU which is responsible for executing instructions. It involves the way that the CPU is designed to fate report and execute instructions as well as how it communicates with other components of the system.

In summary, Computer architecture deals with the overall structure and designed of a computer system, while computer organisation deals with the specific details of how the components are implemented and interact with each other.

Structure and function of major components of a computer

Structure : The way in which the components are interrelated.

Function : The operation of each individual component as part of the structure.

Function :
There are 4 basic functions that a computer can perform.

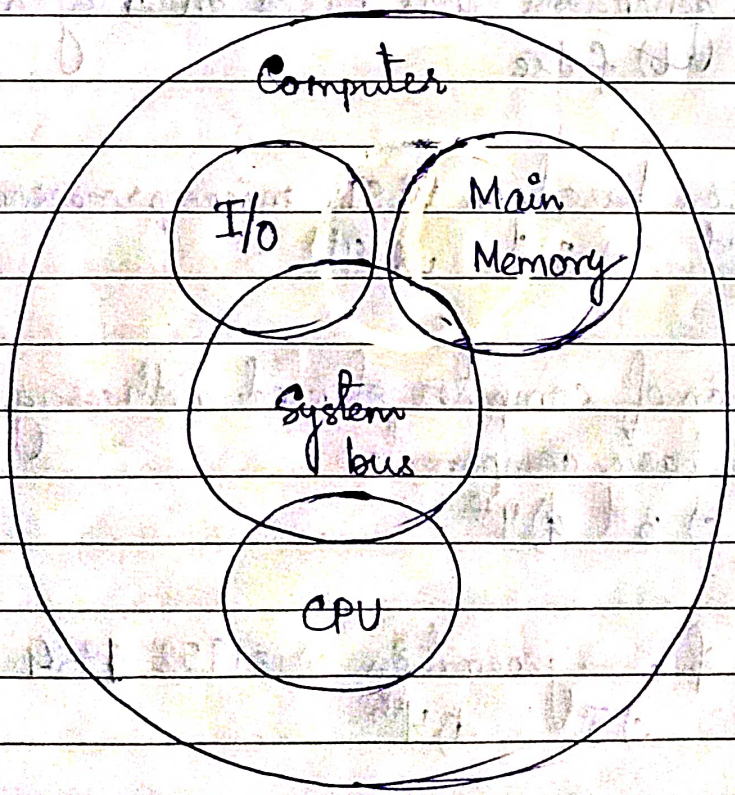
(i) Data Processing → - Data are of variety of forms.
- Data processing is manipulation of data by computers which includes the conversion of raw data to machine readable form, flow of data through CPU and memory to output device.

(ii) Data Storage → It is of two types -
Short term data storage function
Long term data storage function

(iii) Data movement → The receiving and delivering of data is called data movement.
- When data are received from and delivered to a device ie directly connected to the computer, the process is called Input-Output (I/O) and the device is called as Peripheral.
- When data are moved over a long distance to or from a remote device, the process is called data communication.

(iv) Control → A control unit manages the computers resources and arranges the performance of the computer functional parts.

Structure : The internal structure of a simple single processor.



Later Generations : (Microprocessors)

Two important developments of later generations are :

1) Semiconductor Memory : First application of IC is processor. It is faster, smaller in size, memory cost decreased with corresponding increase in physical memory density.

2) Microprocessors : It started in 1971 with the development of first chip 4004 to contain all the components of CPU on a single chip.

QUIZ

1) The fourth generation was based on Integrated circuits.
a) true (b) false

2) The generation based on VLSI microprocessor.
a) 1st b) 2nd c) 3rd d) 4th

3) ____ generation of computer started with using vacuum tubes as the basic components.
a) 1st b) 2nd c) 3rd d) 4th.

4) The period of ____ generation was 1952-1964.
a) 1st b) 2nd c) 3rd d) 4th

5) Select the technology that is used in the 1st generation of computer.
a) Transistor b) LSI c) Vacuum tube d) VLSI e) None of these

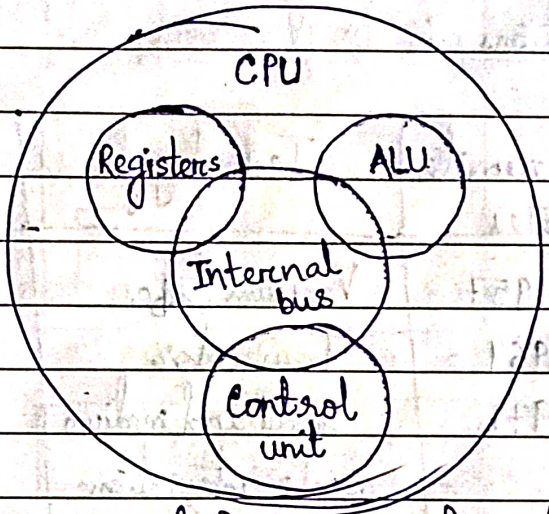
5) Select the name of generation in which ~~introduced~~ Microprocessor ~~was~~ introduced.
a) 4th b) 2nd c) Both (A) & (B) d) 3rd e) all of these.

CPU : It performs data processing and called as a processor.

Main memory : It stores data.

I/O : Moves data between computer and its external environment.

System Interconnection or System Bus :
System bus provides System interconnection.
System interconnection is a mechanism to communicate among CPU, main memory and I/O.



Control unit : It controls the operation of CPU

ALU (Arithematic and Logic Unit) : It performs the computers data processing functions.

Registers : Provides storage.

CPU Interconnection / Internal bus :
Internal bus provides CPU interconnection by communicating among Register, ALU and control unit.

Multiprocessor :

MultiCore Computer Structure : The computers with multiple processors present on a single chip is called a multicore computer and each processing unit consists of control unit, ALU, Registers and Cache Memory.
- An important feature of this is the use of multiple layers of memory called cache memory between the

processor and the main memory.

Brief History of Computers

Computers generation are classified based on the fundamental hardware technologies used. Each new generation is characterised greater processing performance, lower cost, smaller size and larger memory ~~specification~~ capacity than the previous one.

Computer Generations

Generation	Approximate Dates	Technology	Typical speed (operations per second)
1	1946-1957	Vacuum tube	40,000
2	1957-1964	Transistor	200,000
3	1965-1971	Small and medium scale integration	1,000,000
4	1972-1977	Large scale integration	10,000,000
5	1978-1991	Very large scale integration	100,000,000
6	1991-	Ultra large scale integration	>1,000,000,000

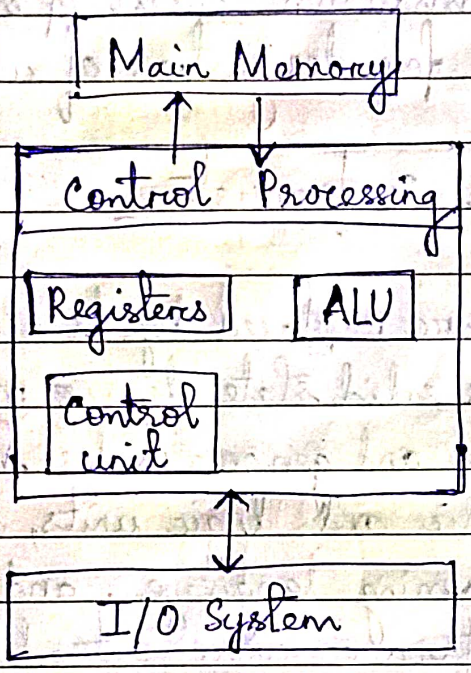
1st Generation : Vacuum tubes

- The first generation computers used vacuum tubes for digital logic elements and memory - It is known as IAS (Institute of Advance Studies).

- Basic Design approach is the stored program concept
  - The idea was proposed by Von Neumann.
  - It consists of (i) main memory (which stores both data and instructions)

(ii) ALU – (capable of operating on binary data)
(iii) Control unit (which interpretes the instructions in memory and causes them to be executed.
(iv) Input/Output (the equipments operated by control unit)

IAS Computer Structure



Von Neumann's Proposal

→ As the device is primarily a computer, it has to perform the elementary arithmetic operations.
→ The proper sequencing of operations can be carried out by the central control unit.
→ The device must have a memory unit to carry out a long and complicated sequences of operation.
→ The device must have interconnections to transform informations from outside of the device to the main memory.
→ The device must have interconnections to transform information from its memory to the outside of the device.

Second

→ The memory of IAS consist of 4096 storage locations. It stores both data and instructions.
Each number is represented by 40 bits.
1st bit as sign bit then and rest 39 bits indicate the value of numbers.
Similarly each instruction consist of 8 bits for opcode and 12 bits for designation of address.

Second Generation : Transistors

- Vacuum tubes were replaced by transistors.
- Transistor is a solid state device made from silicon. smaller, cheaper and generates less heat the vacuum tubes.
- Complex arithmetic and logic units, control units, high level programming language, and the provision of system software were introduced.
- Multiplexers were used which are the central termination point for data channels, CPU and memory.

Third Generation : Integrated Circuits (IC)

- The IC consists of discrete components like transistors, resistors and capacitors etc.
- The two fundamental components that are required are gates and memory cells.
- Gates control the data flow.
- Memory cells store 1 bit data.

Later Generation : (Previous page (after 4th generation))

Date
Page

QUIZ.

7) 2nd generation computers are manufactured of
a) Vacuum tubes b) LSI c) Transistors d) VLSI e) None of these

8) A computer contains
a) A CPU b) A memory c) Input and Output unit d) All of the above

Answers:

1. b) false
2. d) 4th
3. a) 1st
4. b) 2nd
5. c) Vacuum tube
6. a) 4th
7. c) transistors
8. d) all of the above

LAB

Microprocessor - It is CPU on a chip.
- A single chip is called microprocessor. It is also called CPU.

Evolution of microprocessors

4 bit : 1971 by Intel Corp.
. Intel 4004 then 4040.

8 bit : 1973 by Intel → Intel 8008 then 8088.
→ this microprocessor has 40 pins.

16 bit : Intel 8086 and 80286. (stores 1MB)

32 bit : Intel 80386.

Pentium Processor : New processor instead of 80586.

Features of 8086

- It is a 40 pin dual in line package IC.
- It is a 16 bit microprocessor.
- 8086 has a 20-bit address bus and can access upto 2^{20} (1 MB) memory locations.
- It can support upto 64 K I/O ports.

no. of
- It provides 14 ^ 16-bit registers.
- Word size is 16 bits and double word size is 4 bytes.
- It has multiplexed address and data bus AD0 - AD15 and A16 - A19.
- It requires +5 V power supply.
- It can pre-fetch upto 6 instruction bytes from memory and queues them in order to speed up instruction execution.
- ~~It has multiplexed address and data bus~~
- Address ranges from 00000H to FFFFFH.
- Memory is byte addressable — every byte has a separate address.
- 8086 is designed to operate in two modes : Minimum (8 bit) and Maximum. (16 bit)

n-bit address bus → 2^n storage
1 byte = 8 bits

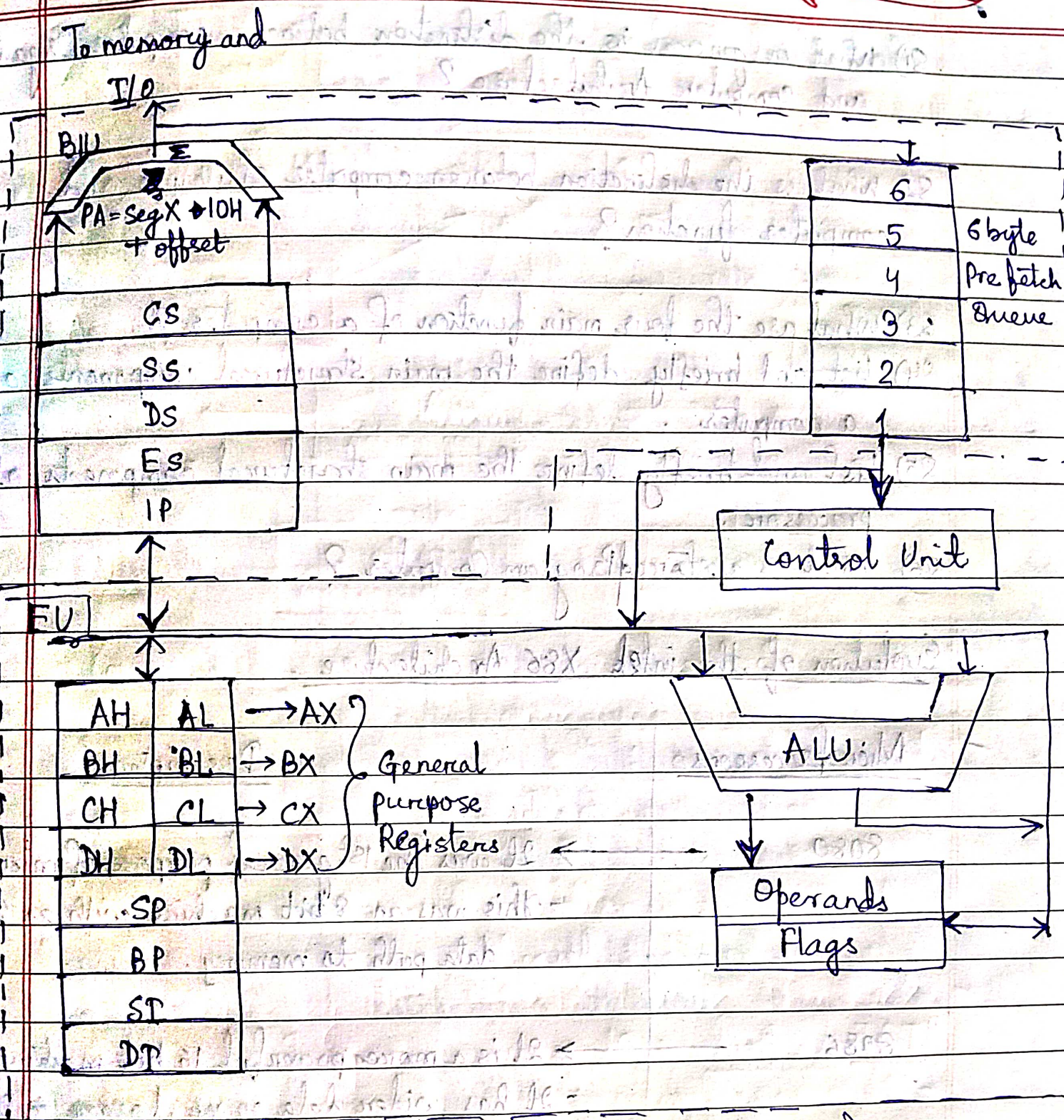
AD0 (zero)

What to Covered in Ch-1

(Pg 31)

1.1, 1.2 — fig 1.2, 1.3 (first generation) — fig (1.7) pg-38, 2nd generation, later generation, table 1.3, 1.4, 1.5, 1.6 — arm products, 1.7

(address line, memory, virtual memory, cache, no. of course)



(Block Diagram of 8086 Microprocessor)

Q1) What in general is the distinction between computer Organization and Computer Architecture?

Q2) What is the distinction between computer structure and computer function?

Q3) What are the four main function of a computer?

Q4) List and briefly define the main structural components of a computer.

Q5) List and briefly define the main structural components of a processore.

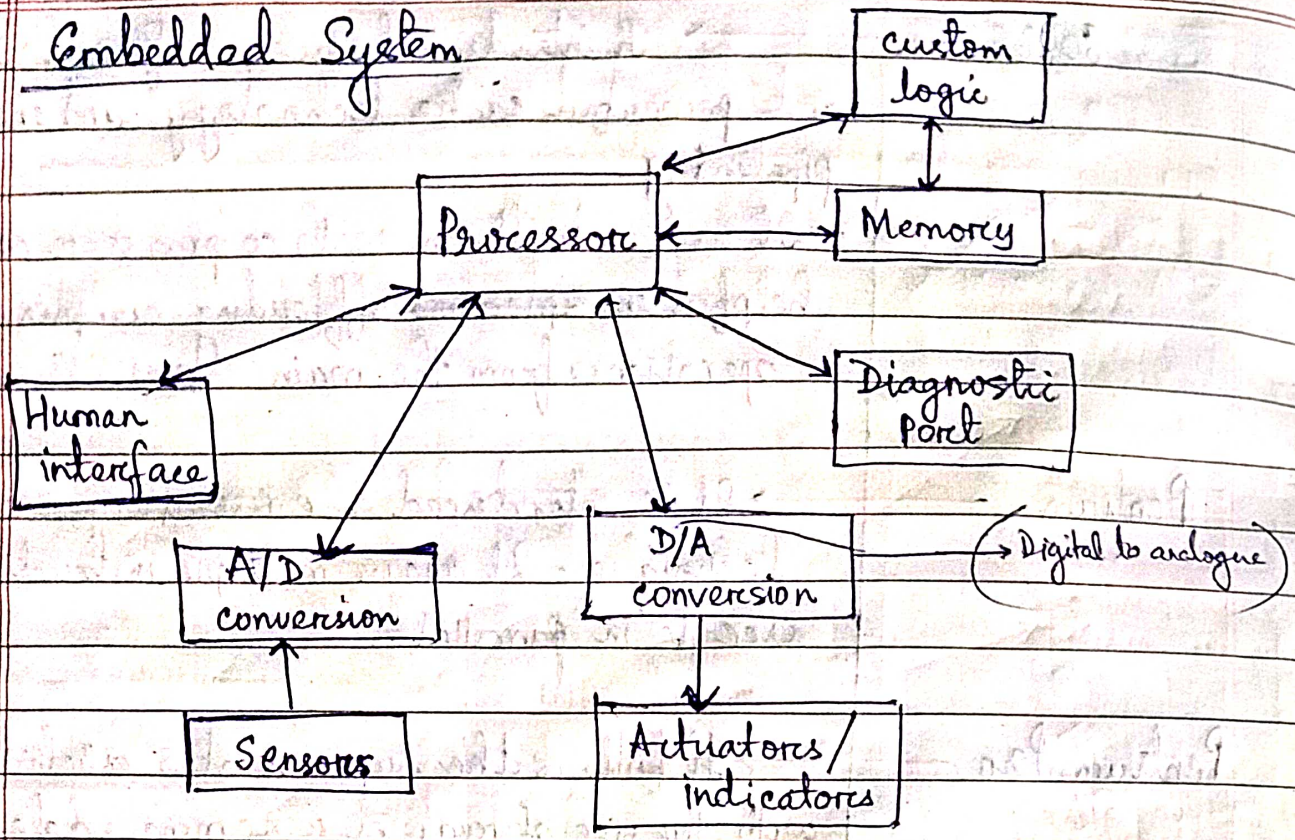
Q6) What is a stored Program Computer?

Evolution of the intel X86 Architecture.

Microprocessors	Description
8080	→ It was the 1st general purpose of microprocessor - This was an 8 bit machine with an 8 bit data path to memory.
8086	→ It is a more powerful 16 bit machine. - It has wider data power, larger registers and an instruction queue. - It is the first use of X86 architecture
80286	→ It is an extension of 8086. - It enabled addressing of 16 MB memory instead of 1 MB.
80386	→ It is intel's first 32 bit machine - It was the first intel processor to support multitasking.

80486	→ It introduced the use of sophisticated and powerful cache technology and instruction pipelining. - It uses a built-in math co-processor which is helpful in ~~uploading~~ offloading complex math operations from the main CPU.
Pentium	→ It is introduced the use of super scalar technique. It allows multiple instructions to execute in parallel.
Pentium Pro	→ It follows the super-scalar architecture with the use of register renaming, data flow analysis etc.
Pentium 2	→ It used intel MMX technology which is designed specifically to process video, audio, and graphics data efficient.
Pentium 3	→ It uses additional ~~protein~~ floating point instruction. It added 70 new instruction.
Pentium 4	→ It had other enhancement for multimedia.
Core	→ It is the first intel X86 microprocessors with dual code, referring to the implementation of 2 cores on a single chip.
Core 2	→ It extends the core architecture to 64 bits. The core 2 Quad provides four course on a single chip.

Embedded System



(Possible Organization of an Embedded System)

- Human Interface : The human interface in an embedded system is important for enabling interaction between user and the system.
  - It allow ~~all~~ users to input commands and settings and to recieve feedback from the system in a user friendly way.

Diagnostic Port : It provides a way of communicating with systems firmware and hardware.
- Firmware is a piece of programming code i.e. embedded in a specific hardware.
- The use of diagnostic port can greatly simplify the process of debugging an embedded system.
- It allows developers and technicians to monitor system behaviour in real time and to quickly identify and isolate problems.
- In some cases, the diagnostic port can be used to update firmware or reprogram the system to rectify issue.

Custom Logic - It can be used to tailor an embedded system with specific application or to add functionality ie. not available in standard components.
- It can be used to improve the performance of an embedded system by offloading certain task from the main processor.
- This can lead to faster response time and reduce power consumption.

- Elements that are different in an embedded system from typical desktop / laptop.

Similarity between Embedded Systems and general purpose Computers:

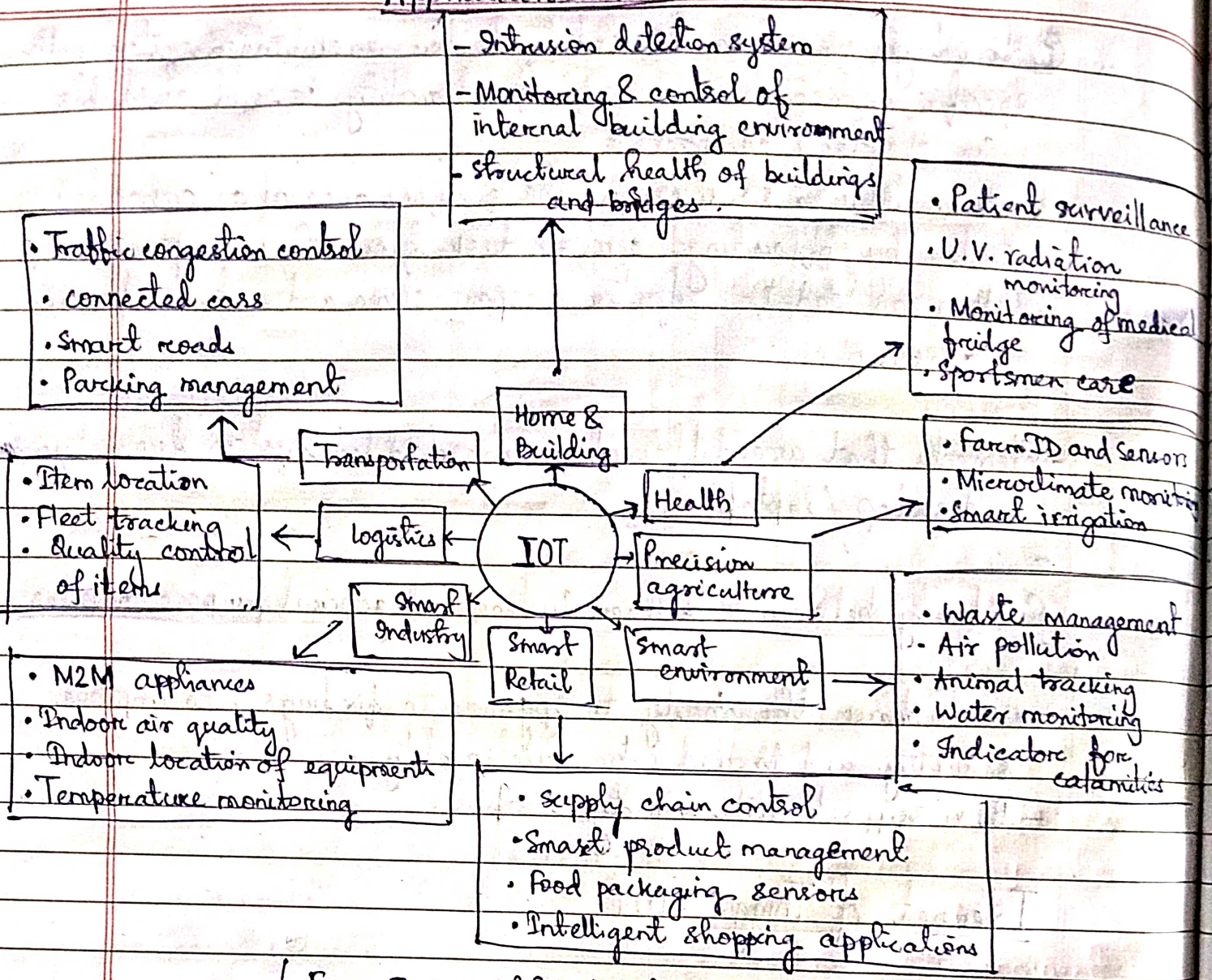
- Both have the ability to upgrade to fix bugs, to improve security and to add extra functionality.
- Both support wide variety of apps.

Internet of Things (IOT)

IOT Applications and Use Cases

1. Smart home
2. Wearables
3. Smart city
4. Smart grids
5. Industrial internet
6. Connected car
7. Connected health
8. Smart Retail
9. Smart supply chain
10. Smart farming

Application Domains



(Fig: IOT applications)

IOT

- It is a system of interrelated computing devices, with unique identifiers (UIDs) and they have the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction.
  Eg:- Embedded systems, wireless sensor networks, automation (home & building), smart home, home security system etc.
- These devices are low-bandwidth, low-repetition data capture, low-bandwidth data-usage appliances that communicate with each other.

- With reference to the products, the internet has gone through four generations.

1) Information Technology (IT) :- PCs, servers, routers, using wired connectivity.

2) Operational technology (OT) :- Medical machinery, SCADA, process control using wired connectivity.

3) Personal technology :- Smartphones, tablets and eBook readers using wireless connectivity.

4) Sensor/Actuator technology :- Single-purpose devices bought by consumers.

Embedded Operating Systems

There are two approaches to develop an embedded operating system (OS).

1) The 1st approach is to take an existing OS and adapt it for the embedded applications.
Eg:- Embedded versions of LINUX, Windows, MAC

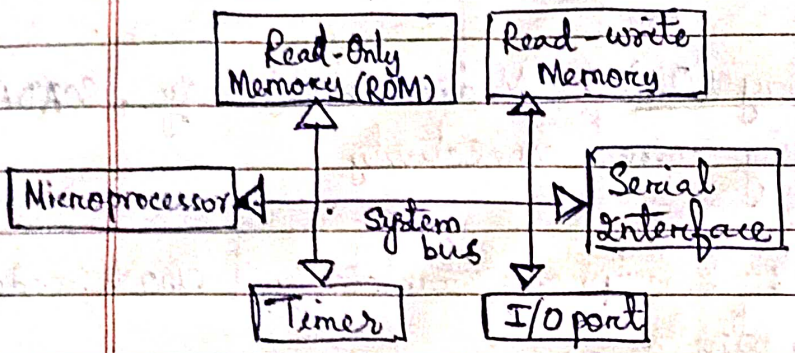
2) The other approach is to design and implement an OS intended solely for embedded use.
Eg: TinyOS (used in wireless sensore networks).

Application Processors Versus Dedicated Processors

- Application Processors are defined by the processors ability to execute complex OS such as LINUX, Android & Chrome.
- They are general-purpose in nature.
- Eg:- Use of application processor is the Smartphones.

- Dedicated Processors are dedicated to one or small no. of tasks.

Microprocessors Vs MicroControllers



Microcontroller	ROM	Read write memory
Timer	I/O port	Serial interface

- It is heart of Computer system
- It is just a processor, Memory and I/O components have to be connected externally.
- It is CPU on a chip.
- Since memory and I/O has to be connected externally, the circuit becomes large.
- Cannot be used in compact systems and hence inefficient.
- Cost of the entire system ↑ses.
- Due to external components, the entire power consumption is high. Hence it is not suitable to used with devices running on stored power like batteries
- Most of the microprocessors do not have power saving features.

- It is a heart of embedded system
- It has external processor along with internal memory and I/O components.
- It is a computer on a chip.
- Since memory and I/O are present internally, the circuit is small.
- Cannot be used in compact systems and hence it is an efficient technique.
- Cost of the entire system is low
- Since external components are low, total power consumption is less and can be used with devices running on stored power like batteries.
- Most of the microcontrollers have power saving modes like idle mode & power saving mode. This helps to reduce power consumption even further.

- Since memory and I/O components are all external, each instruction will need external operation, hence it is relatively slower.	- Since components are internal, most of the operations are internal instruction, hence speed is fast.
- These have less no. of registers, hence more operations are memory based.	- These have more no. of registers, hence the programs are easier to write.
- These are based on Von-Neumann model/architecture where program and data are stored in same memory module.	- These are based on Harvard architecture where program memory and data memory are separate.
- Mainly used in personal computers.	- Used mainly in Washing machine, MP3 players.

- Since memory and I/O components are all external, each instruction will need external operation, hence it is relatively slower.	- Since components are internal, most of the operations are internal instruction, hence speed is fast.
- These have less no. of registers, hence more operations are memory based.	- These have more no. of registers, hence the programs are easier to code.
- These are based on Von-Neumann model/architecture where program and data are stored in same memory module.	- These are based on Harvard architecture where program memory and data memory are separate.
- Mainly used in personal computers.	- Used mainly in Washing machine, MP3 players.

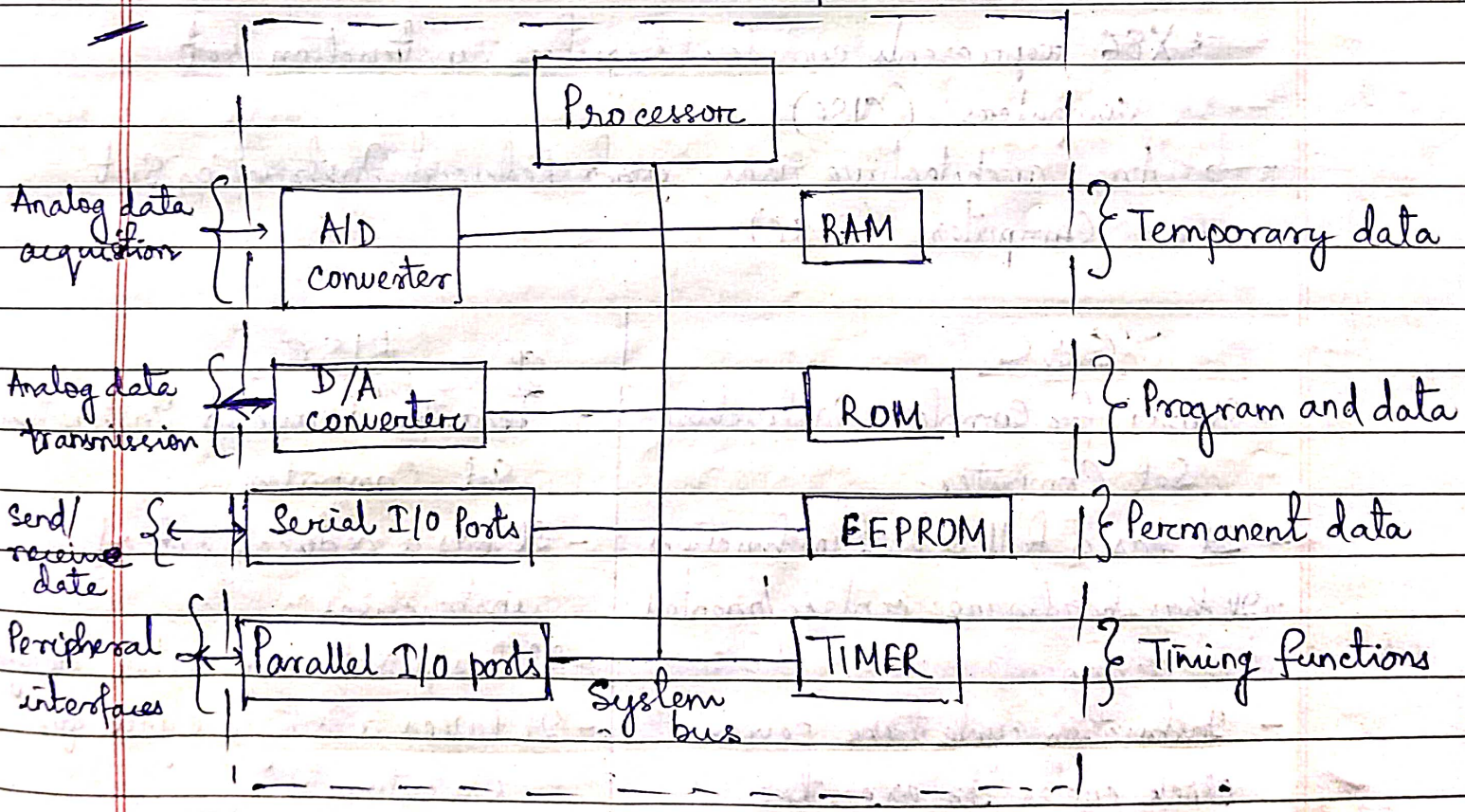


Fig: Typical Microcontroller chip Elements

EEPROM - Electrically Erasable Programmable ROM.

Embedded Vs. Deeply Embedded Systems

- Deeply Embedded Systems has a processor whose behaviours is difficult to observe
- It uses micro controller rather than microprocessors.
- Once it is designed it can't be programmable.
- It is dedicated, single purpose devices.

(Acorn) → (Advanced RISC Machines)

ARM Architecture

- There are two important processors family
  (1) X86 family
  (2) Arm architecture.

→ X86 represents complex instruction set computers (CISC)

→ Arm architecture uses Reduced Instruction Set Computer (RISC)

CISC	RISC
- Stands for Complex Instruction Set Computer.	- Stands for Reduced Instruction Set Computer.
- It has a full set of instructions	- It has a reduced set of instructions.
- It has hardware centric/oriented design.	- It has a software centric design
- Instruction cycle take several clock cycles to execute.	- It takes a single cycle for execution.
- Pipelining is difficult.	- Pipelining is easy.
- Complex and variable length instructions.	- Simple & standardized instructions
- large no. of instructions.	- Small no. of instructions.

- Compound addressing modes.	- Limited addressing modes.
- It uses less no. of registers.	- It uses more no. of registers.
- Require less RAM.	- Requires more RAM.
- Desktop, Computer & Laptops.	- Mobile phone & tablets.

ARM Revolution

- ARM is a family of RISC based microcontrollers.
- ARM chips are high speed processors.
- They are of very small size and required very less power.
- They are widely used in Smartphones and other hand held devices.
- It is most widely used embedded processor architecture.
- Acron RISC Machine / ARM was the first to develop commercial RISC Processor.

Instruction Set Architecture (ISA)

- All instructions are 32 bits long and follow a regular format.
- Augmentation of basic ARM ISA is Thumb Instruction Set. which is a re-encoded subset of the ARM instruction set.
- Thumb is designed to increase the performance of ARM implementations that use a 16 bit memory data bus.

ARM Products

(ARM- Advanced Risc Machine)

- There are three CORTEX architecture named with initials A, R & M.
  1. CORTEX - A / CORTEX-A50
  2. CORTEX - R
  3. CORTEX - M

1) CORTEX-A / CORTEX-A50

- They are application processors.
- They are used for mobile devices such as smart phones and ebooks readers.
- These processors run at higher clock frequency.
- They support a MMU (Memory Management Unit).
- The two architectures use both the ARM and Thumb-2 instruction sets.
- CORTEX-A is a 32-bit machine and CORTEX-A50 is a 64-bit machine.

2) CORTEX-R

- It is designed to support real time applications.
- They run at a higher clock frequency and have a very low response latency (delay).
- Most of these processors don't have MMU.
- It does not have a MPU (Memory Protection unit), cache and other memory features for industrial applications.
- Ex: Automotive breaking systems

3) CORTEX-M

- They have been developed for the microcontroller domain.
- They have MPU, but no MMU.
- It uses the Thumb-2 instruction set.
- Ex: IoT devices.
- They are of 4 versions:
  (i) CORTEX-M0
  (ii) Cortex-M0+
  (iii) CORTEX-M3
  (iv) CORTEX-M4

Cloud Computing -

Basic concepts

- It is a technology to use a network of remote services hosted on the internet to store, manage and process data rather than a local server or personal computer.
- These features are attractive to companies, govt. agencies and mobile users.
- The individual or company only needs to pay for the storage capacity and services they need.
- The cloud also takes care of the data security.

Cloud Networking -

- It refers to the networks and network management functionality that must be in place to enable cloud computing.

- Cloud Storage can be thought of one subset of cloud computing. It consists of database storage and database applications hosted remotely on cloud servers.

Cloud Services

- It is of 3 types.
- A cloud service provider (CSP) maintains computing and data storage resources that are available over the internet.
- Customers can rent a portion of these resources as needed.
- All cloud services are provided using one of the three models;
  (i) SaaS
  (ii) PaaS
  (iii) IaaS

(i) SaaS (Software as a Service)

- It is a way of delivering applications over the internet as a service.
- These are also known as web-based software, on-demand software or hosted software.
- Ex: Netflix, Robox, google workspace

(ii) PaaS (Platform as a Service)

- It is a cloud computing model where a third party provider delivers hardware and software tools to users over the internet.
- PaaS is an operating system in the cloud.
- It is useful for an organisation that wants to develop a new application while paying for the needed computing resources.
- Ex: Google appengine

(iii) IaaS (Infrastructure as a Service)

- In which IT infrastructure is provided to user through internet.
- Ex: Amazon Web Services (AWS), Microsoft Azure and google compute engine.

- Moore's Law

- In 1965, Gordon Moore predicted that the no. of transistors in a single chip would be doubled in every years and cost of the chip will remain virtually unchanged.

CHAPTER-2 PERFORMANCE ISSUES

In this chapter, we will learn different performance issues of the computer.
- How computers are designed to enhance the performances.
- How different processors are used for ~~differ~~ better performance.
- Basic measures of computer performance.
- Benchmark of performances.

Designing for Performance

① Microprocessor speed -

— Pipelining : while one instruction is being executed the computer is decoding the next instruction, this principle is called pipelining.

— Branch Prediction :

The processor looks ahead in the instruction code to predict which branches or group of instructions are likely to be processed next. In this way branch prediction potentially increases the ~~arrows~~ speed of the processor.

— Superscalar execution : It is a process of execution in which more no. of instructions can be executed by using a single clock cycle.

— Dataflow analysis : The processor analyses which instructions are dependent on each others data/result to create an optimized schedule of instructions. In fact, instructions are scheduled to be executed when they are ready and independent of the original program orders. This prevents unnecessary ~~dead~~ delay.

— Speculative execution : Using Branch Prediction and Dataflow analysis, some processors speculatively execute instructions ahead of their actual appearance in the program to speed up the processor.

② Performance Balance :-

The performance balance is an adjustment / tuning of the organization and architecture to compensate for the mismatch among the capabilities of the various components.

The followings are the methods for performance balance :-

- By increasing the no. of bits that are retrieved at one time by making DRAMs "wider" rather than "deeper" and by using wide bus data paths.
- ~~By changing the DRAM interface to make it more efficient~~
- By including more cache memory.
- By reducing the frequency of memory access by using efficient cache structures.
- By ~~changing~~ using high speed buses.

③ Improvements in Chip Organisation and Architecture

(i) By increase the hardware speed of the processor.
(ii) By increasing the size and speed of caches.
(iii) By making changes to the processor organisation and architecture to increase the effective speed of the instruction execution.

- Generally to increase the speed of the processor we should increase the clock speed and logic density.
- However, following facts are the obstacles to the increment of clock speed and logic density.

Power: As the logic density and clock speed on a chip increase, so does the power density. (Watts/cm^2).

RC delay: The speed at which electrons can flow on a chip between transistors is limited by the resistance and capacitance of the metal wires connecting them; As the RC products will increase the delay will also increase.

Memory latency and throughput: Memory access speed (latency) and transfer speed (throughput) lag processor speeds.

Multicore, MICS, GPGPUS

→ The designers used a new approach to improve the performance of the computer by placing multiple processors on the same chip with a large shared cache.
- The processors is called multicore and it provides better performance without increasing the clock speed.

→ Later designers tried to put as much as 50 crores in a chip and introduced a new term: Many Integrated Core (MIC).
→ After multicore and MIC the designers used Graphics Processing Units (GPUs) for vedio processing.

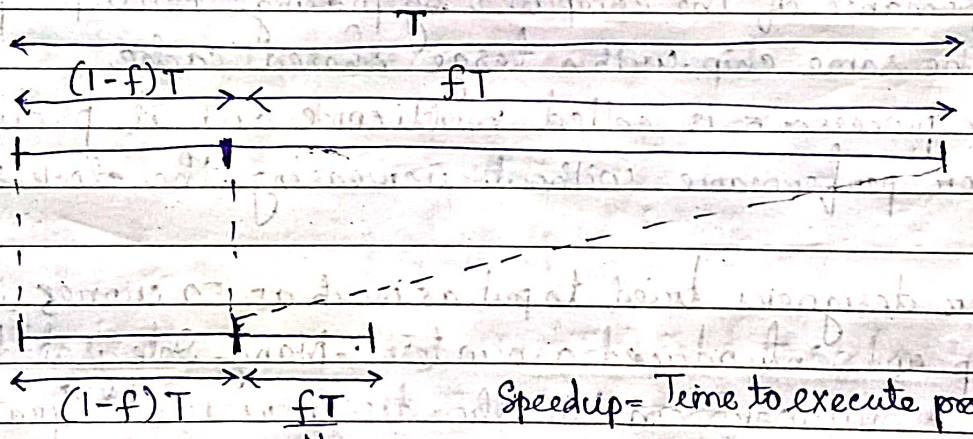
→ Then GPGPUs (General-Purpose Computing on GPUs) is used to enhance the performance of computer.
In GPGPUs, the GPUs is used to perform the general purpose computing of CPU especially to deal with graphic data.

Amdahl's law : It is a mathematical formula which is used to determine the maximum improvement possible by improving a particular path of a system.

- In parallel computing it is mainly used to predict the theoretical maximum speed up for program processing using multiple processors.
- When a program run on a computer with parallel processing computation,
- It may use serial processing or parallel processing or both.

Let f is the fraction of execution time when parallel processing is used.

(1-f) is the fraction of execution time when serial processing is used.



$$\text{Speedup} = \frac{\text{Time to execute program on a single processor}}{\text{Time to execute program on N parallel processors.}}$$

$$= \frac{(1-f)T + fT}{T(1-f) + \frac{Tf}{N}} = \frac{1-f+f}{1-f+\frac{f}{N}}$$

$$\text{Speedup} = \frac{1}{(1-f) + \left(\frac{f}{N}\right)}$$

where N = no of processors.

From this eq, two important conclusions can be drawn:

1. When f is small, the use of parallel processors has little effect.
2. As N approaches infinity, speedup is bound by $1/(1-f)$, so that there are diminishing returns for using more processors.

Q) Findout the diminishing returns.
50% of time for parallel processing. What is the speed up and with infinite no. of processor.

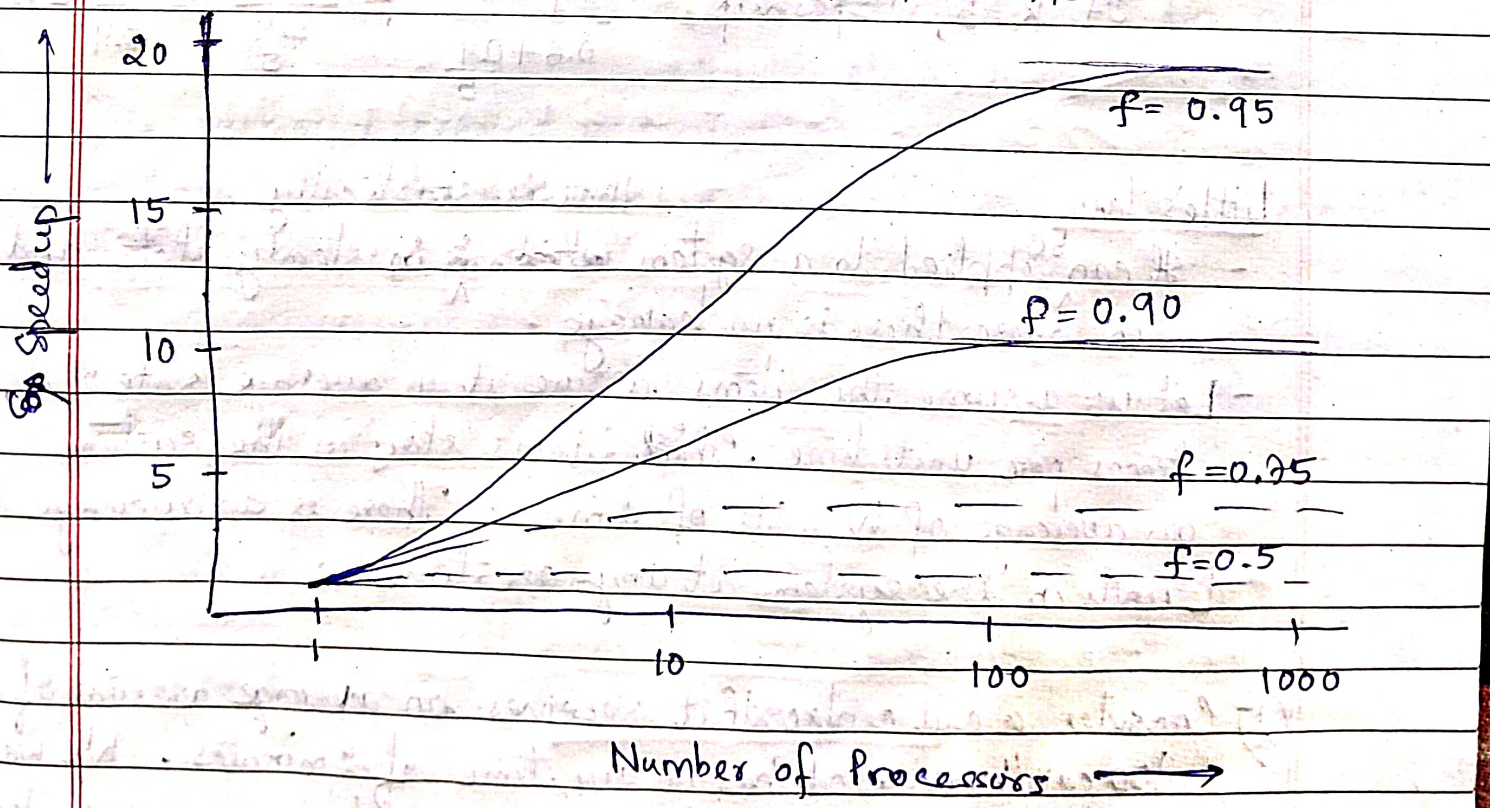
Ans) $f = 50\% = 0.5$

$$\text{Speedup} = \frac{1}{1-0.5} = \frac{1}{0.5} = 2$$

If $f = 75\% = 0.75$, $\text{Speedup} = \frac{1}{1-0.75} = \frac{1}{0.25} = 4$.

If $f = 90\% = 0.9$, $\text{Speedup} = \frac{1}{1-0.9} = \frac{1}{0.1} = 10$

If $f = 95\% = 0.95$, $\text{Speedup} = \frac{1}{1-0.95} = 20$



Overall speedup -

Suppose that a feature of the system is used during execution a fraction of the time f, before enhancement, and that the speedup of that feature after enhancement is SU_f. Then the overall speedup of the system is

$$\text{Speedup} = \frac{1}{(1-f) + \frac{f}{SU_f}}$$

Example 2.1 → Suppose that a task makes extensive use of floating-point operations with 40% of the time consumed by floating-point operations. With a new hardware design, the floating-point module is speed up by a factor of K. Then the overall speedup is as follows.

Ans → $\text{Speedup} = \frac{1}{(1-f) \quad 0.6 + \frac{0.4}{K}}$ $\left(\frac{1}{(1-f) + \frac{f}{SU_f}}\right)$

If $K=5$, $\text{Speedup} = \frac{1}{0.6 + \frac{0.4}{5}} = \frac{1}{\frac{3.4}{5}} = \frac{5}{3.4} = 1.47$

Little's law

- It can be applied to a system that is stastically in steady state and in where there is no leakage.
- Let us assume the items arrive at an average rate of λ times per unit time. The items stay in the system an average of W units of time. There is an average of L units in the system at any one time.

Q) Consider a call centre that receives an average arrival of 100 call/hour and has an avg. time of 2 minutes. We want to calculate the avg no. of customers (L) at a given time by using Little's law.

($W = 2$ mins/call)
($\lambda = 100$ calls/hr)

Ans) $L = \lambda W = 100 \times 2 = 200$ mins/hr.

- Consider a multicore system, with each core supporting multiple threads of execution. At some level, the cores share a common memory. ~~The cores share a common main memory and typically share a~~ common cachememory ~~as well~~.

- For this purpose, each user request is broken down into subtasks that are implemented as threads. We then have, λ = the avg. rate of total thread processing required after all members requests have been broken down into whatever detailed subtasks are required. Define L as the avg. no. of stopped threads waiting during some relevant time. Then,

  W = avg. response time.

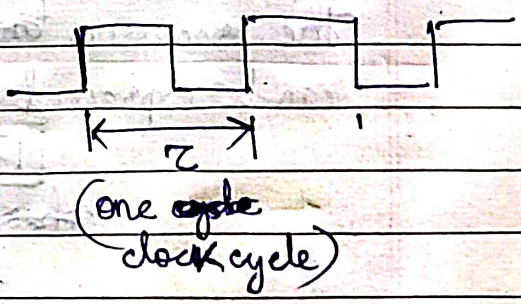
Basic Measures of a Computer Performance

(1) Clock Speed - The speed of a processor is defined by the pulse frequency produced by the clock, which is measured in cycles per second, or Hertz (Hz).

- The rate of pulses is known as clock rate, or clock speed.
- One increment, or pulse, of the clock is called as a clock cycle or clock tick. The time between pulses is the cycle time.

Clock rate / clock frequency (f)

$f = \frac{1}{\tau}$ in Hz (= cycles/sec)



Cycle time, $\tau = \frac{1}{f}$

(2) Instruction Execution Rate

- A processor is driven by a clock with a constant frequency f and with a constant cycle time τ, where $\tau = \frac{1}{f}$.

Instruction Count : (I_c)

- For a program, the no. of machine instructions executed until it runs to completion or for some defined intervals.

- Avg. Cycles per Instruction (CPI)

$$CPI = \frac{\sum_{i=1}^{n} (CPI_i \times I_i)}{I_c}$$

(There are different types of instruction which takes different no. of clock cycles)

Instruction type	No. of instructions	CPI (cycles per instruction)
ALU	40	2
Load and store	15	3
Branch	35	4
Other	10	5
	$I_c = 100$	

Avg. / Effective / Overall CPI $= \frac{(2\times40) + (3\times15) + (4\times35) + (5\times10)}{40+15+35+10}$

$$= \frac{(CPI_1 \times I_1) + (CPI_2 \times I_2) + (CPI_3 \times I_3) + (CPI_4 \times I_4)}{I_c} = 3.15$$

$$= \sum_{i=1}^{n} \frac{(CPI_i \times I_i)}{I_c} \quad \text{(here, } n=4\text{)}$$

The processor time, T needed to execute a given program can be expressed as

$$\boxed{T = I_c \times CPI \times \tau}$$

T = execution time = second (unit)

$$T = I_c \times [P + (m \times k)] \times \tau$$

where 'P' is the no of processor cycles needed to decode and execute the instruction.

'm' is the no. of memory references needed.

'k' is the ratio between memory cycle time and processor cycle time.

Instruction cycle divided into processor cycle and memory cycle. Usually memory cycle is k times the processor cycle.

MIPS Rate : (Million of Instructions per Second Rate)

It is the rate at which instructions are executed.

$$\text{MIPS rate} = \frac{I_c}{T \times 10^6} = \frac{\frac{I_c}{T}}{10^6} = \frac{f}{CPI \times 10^6}$$

$$= \frac{I_c}{I_c \times CPI \times \tau \times 10^6} = \frac{1}{CPI \times \frac{1}{f} \times 10^6} = \frac{f}{CPI \times 10^6}$$

(2) Ex-22

Instruction type	CPI	Instruction Mix (%)
Arithmetic and Logic	1	60
Load/store with cache hit	2	18
Branch	4	12
Memory reference with cache miss	8	10

Consider the execution of a program that results in the execution of 2 million instructions on a 400-MHz processor. The program consists of four major types of instructions. The instruction mix and the CPI for each instruction

type are given below, based on the result of a program trace experiment :

$$\text{Avg. CPI} = \frac{(60\times1) + (18\times2) + (12\times4) + (10\times8)}{100} = 2.24$$

$$\text{MIPS rate} = \frac{f}{\text{CPI} \times 10^6} = \frac{400 \times 10^6}{2.24 \times 10^6} = 178.57$$

(Q2.1) Ex A benchmark program is run first on 200 mHz and then on a 300 mHz processor. The executed program consists of 1 million instruction executions with the following instruction mix and clock cycle count :

Instruction type	Instruction count	Cycles per instruction
Integer arithmetic	4,00,000	1
Data transfer	3,50,000	2
Floating point	2,00,000	3
Control transfer	50,000	2.

Determine the effective CPI and MIPS rate for both the cases.

Ans) Effective CPI $= \frac{(4,00,000\times1) + (3,50,000\times2) + (2,00,000\times2) + (50,000\times2)}{10,09000}$

$= 1.8$

$$\text{MIPS Rate} = \frac{300\times10^6}{1.8\times10^6} = 166.67$$

$$\text{MIPS rate} = \frac{200\times10^6}{1.8\times10^6} = 111.11$$

(Q2.2) Ex Consider two different machines, with instruction set of 100,000 instructions, both of which have a clock rate of 400 MHz. The following measurements are recorded on the two machines running a given set of benchmark programs.

Instruction type	Instructions mix (%)	Cycles per instruction
Machine A	-	-
Arithmetic and logic	50	2
Data transfer	15	3
Control transfer	15	4
Others	20	2
Machine B		
Arithmetic and logic	65	1
Data transfer	15	4
Control transfer	10	3
Others	10	2

Determine the effective CPI, MIPS rate and execution time for both the machines.

Ans) For A

$$\text{Effective CPI} = \frac{(50\times2) + (15\times3) + (15\times4) + (20\times2)}{100} = 2.45$$

$$\text{MIPS Rate} = \frac{400\times10^6}{2.45\times10^6} = 163.2\%.$$

$$\text{Execution time} = T = I_c \times CPI \times C = 100\times2.4\times\frac{1}{400\times10^6}$$

$$= 0.1625\times10^{-6}$$

For B

Effective CPI = $\frac{(65 \times 1) + (15 \times 4) + (10 \times 32) + (10 \times 2)}{100} = 1.75$

MIPS rate = $\frac{400 \times 10^6}{1.75 \times 10^6} = 228.57$

Execution time = T = $100 \times 1.75 \times \frac{1}{400 \times 10^6}$

$= 0.4375 \times 10^{-6}$

Calculating the Mean

In the field of benchmarking single no. are used for performance comparison by calculating the mean value of a set of data points related to execution time. But multiple alternative algorithms are used for calculating a mean value such as arithmetic mean, geometric mean, harmonic mean.

(1) Arithmetic mean – It is a good candidate for comparing the execution time performance of several system.

- It is given as $AM = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$

where $x_1 + x_2 + x_3 + \dots + x_n$ is a set of n real no.

(2) Harmonic mean (HM) = $\frac{n}{\left(\frac{1}{x_1}\right) + \left(\frac{1}{x_2}\right) + \dots + \left(\frac{1}{x_n}\right)} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$

Execution

Suppose we have a set of n benchmark program and the execution time of each program of a given system as $t_1, t_2, t_3, \dots, t_n$.

Let us assume that each program executes same no. of operations (Z)
- The execution rate for each individual program is
$R_i = Z/t_i$

Now let's use AM to calculate the average ~~cost~~ execution rate.

$$AM = \frac{R_1 + R_2 + R_3 + \cdots + R_n}{n} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\sum_{i=1}^{n} \frac{Z}{t_i} = \frac{Z}{n}\sum_{i=1}^{n} \frac{1}{t_i}$$

Here the AM execution rate is proportional to sum of the inverse execution rate.

- For simplicity, let us assume that each program executes the same no. of operations Z, ~~we could weight the individual programs and~~
- The execution rate for each individual program is $R_i = Z/t_i$
- we will use the AM to calculate the avg. execution rate.
- We will use the HM to calculate the avg. execution rate.

$$HM = \frac{n}{\sum_{i=1}^{n} \frac{1}{R_i}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{(Z/t_i)}} = \frac{nZ}{\sum_{i=1}^{n} t_i}$$

We observe that AM execution rate is proportional to the sum of the execution times.

$$R_i = \frac{Z}{t_i} \qquad AM = \frac{Z}{n}\sum_{i=1}^{n} \frac{1}{t_i}$$

The HM execution rate is inversely proportional to the total execution time.

NOTE: less execution time taken by Computer A so, C-A is the fastest.
(AM) Maximum execution rate ⟶ fastest
Maximum HM execution rate ⟶ fastest

Table 2.1 : Comparison of AM and HM for Rates

	Computer A-time (Secs)	Computer A rate (MFLOPS)	C-B $	C-B rate	C-C	C-C rate
Program-1	2.0	$\frac{100\times10^6}{2\times10^6} = 50$	1.0	100	0.75	133.33
P-2	0.75	$\frac{100\times10^6}{0.75\times10^6} = 133.33$	2.0	50	4.0	25
Total execution time	2.75	–	3.0	–	4.75	–
Arithmetic mean	$\frac{2+0.75}{2} = 1.38$	–	1.5	–	2.38	–
Inverse of total execution time (1/sec)	$\frac{1}{2.75} = 0.36$	–	0.33	–	0.21	–
Arithmetic mean of rates	–	$\frac{50+133.33}{2} = 91.67$	–	75.00	–	79.17
Harmonic mean of rates	–	$\frac{2}{\left(\frac{1}{50}+\frac{1}{133.33}\right)} = 72.72$	–	66.67	–	42.11

- The greatest value of AM is for computer A, which means Computer A is the fastest computer.
  B is also slower than C, where as B is faster than C.

- Interms of total execution time, A has minimum time, so it is the fastest computer out of the three.

- The HM values correctly reflect the speed ordering of the computers. This confirms that the HM is preferred when calculating the rates.

⟶ HM is preferred over AM.

WHM (Weighted HM)

$$WHM = \frac{\sum_{j=1}^{n} Z_j}{\sum_{i=1}^{n} t_i}$$

$\sum_{j=1}^{n} Z_j$ → sum of the instruction count

$\sum_{i=1}^{n} t_i$ → sum of the execution times

The numerator of the formula calculates the sum of the operations for all programs where as denominator calculates the sum of the execution times of all programs.

③ Geometric Mean (GM).

$$GM = \sqrt[n]{x_1 * x_2 .. x_3 . \ldots . x_n} = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = e^{\left(\frac{1}{n}\sum_{i=1}^{n} \ln(x_i)\right)}$$

GM of execution rate

$$= \left(\prod_{i=1}^{n} \frac{Z_i}{t_i}\right)^{1/n} = \frac{\left(\prod_{i=1}^{n} Z_i\right)^{1/n}}{\left(\prod_{i=1}^{n} t_i\right)^{1/n}}$$

Comparison of AM and GM for Normalised Results to Computer A :-

(GM, n=2)

	Computer A time	Computer B time	Computer C time
Program 1	2.0 (1.0) $= \frac{2}{2}$	1.0 (0.5) $= \frac{1}{2}$	0.75 (0.38) $= \frac{0.75}{2}$
Program 2	0.75 (1.0) $= \frac{0.75}{0.75}$	2.0 (2.67) $= \frac{2}{0.75}$	4.0 (5.33) $= \frac{4}{0.75}$
Total execution time	2.75	3.0	4.75
AM of normalized times	1.00 $= \frac{1+1}{2}$	1.58 $= \frac{0.5+2.67}{2}$	2.85 $= \frac{0.38+5.33}{2}$
GM of normalized times	1.00 $= \sqrt[2]{1 \times 1}$	1.15 $= \sqrt[2]{0.5 \times 2.67}$	1.41 $= \sqrt[2]{0.38 \times 5.33}$

Total execution time : fastest → A > B > C

G.M of normalized time : fastest → A > B > C

AM of normalized times : fastest → A > B > C

→ GM is preferred over AM.

Disadvantage of GM - If AM is monotonically increasing it does not mean GM will increase monotonically.

Advantages of GM - It gives consistent results.
- It is not biased by outliers than the HM or AM.
(less)
- Distributions of performance ratios are better modeled by lognormal distributions than by normal ones.

BenchMarks & SPEC

2) What are the desirable ~~qualities~~ characteristics of a benchmark program?

Ans) (i) It is written in a high level language, making it portable across different machines.
(ii) It is representative of a particular kind of programming domain or paradigm such as systems programming, numerical programming or commercial programming.
(iii) It can be measured easily.
(iv) It has wide distribution.

SPEC Benchmarks : A benchmark suite is a collection of programs defined in a high-level language, that together attempt to provide a representative test of a computer in a particular application or system programming area. The best known such collection of benchmark suites is defined and maintained by the Standard Performance Evaluation Corporation (SPEC), an industry consortium. This organisation defines several benchmark suites aimed at evaluating computer systems. SPEC performance measurements are widely used for comparison and research purposes.