CHAPTER 1

INTRODUCTION TO COMPUTERS

1.1 Introduction

A computer (or digital computer) is a general purpose electronic device that can be programmed to process the input information in order to generate a useful result. A computer accepts text/graphical information as input and process the input information based on a program and generates output text/graphical information. The program is software consisting of sequence of instructions and the computer is a hardware that executes the software.

The world's first programmable general-purpose electronic computer ENIAC (**E**lectronic **N**umerical **I**ntegrator and **C**omputer) was developed in the year 1945 in Moore School of Electrical Engineering at the University of Pennsylvania, USA. The system consumed 150 kilowatts of electricity, occupied 1800 square feet of floor space and weighed 27 tons. The design was based on vacuum tubes as switching elements. ENIAC contained over 17000 vacuum tubes, 70000 resistors, 10000 capacitors, 6000 switches and 1500 relays.

Invention of semiconductor transistors in the year 1947 led to development of digital computers using discrete semiconductor transistors as switches. The invention of ICs (**I**ntegrated **C**ircuits) in the year 1958 led to development of digital computers using ICs in which the CPU (**C**entral **P**rocessing **U**nit) is developed using multiple ICs. With advancement in IC fabrication technology the multi-chip CPU is integrated as single chip CPU and named as microprocessor. The first commercial microprocessor INTEL4004 was released by Intel Corporation, USA in the year 1971. The microprocessor is also commonly (or simply) referred to as a processor.

IBM (**I**nternational **B**usiness **M**achine), USA released a personal computer in the year 1981 with INTEL 8088 processor as CPU which is considered as bench mark for all digital computers of modern age. Subsequently, IBM released personal computers with Intel processors 80286, 80386, 80486 and PENTIUM as CPU. Apart from IBM many companies started manufacturing personal computers using Intel processors. Some of the prominent companies are HP (**H**ewlett-**P**ackard), Dell, Lenovo, Acer, ASUS, etc.

1.1.1 Computer Generations

The computers developed using vacuum tubes are referred to as first generation of computers and the computers developed using semiconductor discrete electronics components like transistors and diodes are referred to as second generation of computers.

Examples of the first generation computers: ENIAC, EDVAC, IBM-650, IBM-701, Manchester Mark-1, Mark-2, etc.

Examples of the second generation computers: IBM 1620, IBM 7090, IBM 7094, PDP-8, CDC 1604, CDC 3600, UNIVAC 1108.

The computers developed using **I**ntegrated **C**ircuits (ICs) manufactured using SSI/MSI (**S**mall **S**cale **I**ntegration/**M**edium **S**cale **I**ntegration) technology are referred to as third generation computers.

Examples of the third generation computers: IBM-360 series, Honeywell-6000 series, PDP (**P**ersonal **D**ata **P**rocessor) and IBM-370/168.

The fundamental electronic device in a digital IC is transistor. With advancement in IC fabrication technology, large numbers of transistors are packed in a single IC. One way of classification of digital ICs are based on density of fundamental transistors used to fabricate an IC. Based on number of transistors packed in a single IC, the digital ICs are classified as SSI, MSI, LSI, VLSI and ULSI integrated circuits.

The SSI/MSI refers to integration of 500 transistors on a single chip. LSI (**L**arge **S**cale **I**ntegration) refers to integration of 500 to 20000 transistors on a single chip. VLSI (**V**ery **L**arge **S**cale **I**ntegration) refers to integration of 20000 to 1 Million transistors on a single chip. ULSI (**U**ltra **V**ery **L**arge **S**cale **I**ntegration) refers to integration of more than 1 Million transistors on a single chip. The ULSI technology is currently used to fabricate modern processors with billions and trillions of transistors on a single chip.

The computers developed using processors manufactured using LSI, VLSI and ULSI technology are referred to as fourth, fifth and sixth generation computers respectively.

Examples of fourth generation computers: IBM PC (First **P**ersonal **C**omputer (PC) developed by IBM), STAR 1000, CRAY-X-MP (Super Computer), DEC 10, PDP 11, CRAY-1, STAR 1000, APPLE II, Apple Macintosh, Alter 8800.

The fifth and sixth generation computers or latest generation computers are mostly marketed by different companies with processor used in the computer rather than computer name/model. Intel processors are widely used in popular computers. Some of the processors by Intel that are used to develop fifth and sixth generation computers are,

- Pentium
- Pentium Pro
- Pentium II
- Celeron
- Pentium III
- Pentium 4
- Xeon
- Itanium
- Pentium M
- Pentium D
- Core 2 Duo
- Atom
- Intel Core-i series (i3, i5, i7, i9)
- Intel Core ultra series 1, 2

Table 1.1: Computer Generations

Generation	Period	Technology	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 (SSI/MSI)	1,000,000
4	1972-1977	Large scale integration (LSI)	10,000,000
5	1978-1991	Very large scale integration (VLSI)	100,000,000
6	1991- Till date	Ultra large scale integration (ULSI)	>1,000,000,000

1.1.2 Computer Classification based on Current Market Requirements

Computer technology has made incredible progress in the roughly 79 years (1945 to 1924) since the first general-purpose electronic computer ENIAC was created and have set the stage for a dramatic change in how we view computing, computing applications and the computer markets in the modern era. The following five different computing markets have emerged, each characterized by different applications, requirements and computing technologies.

Personal Mobile Devices (PMD): The collection of wireless devices such as cell phones, tablet computers, play-stations, etc., can be grouped under PMD. These devices are developed using high performance processors with only semiconductor memory RAM and ROM. They are provided with interface to external magnetic memories for additional storage and low power battery operated hand held devices.

Laptops and Desktops: These are high performance personal computers widely used by individuals and in almost all work places. They are general purpose computers with a single powerful processor and large capacity of different types of memory. The laptops are powered by battery and desktops are powered by AC to DC power converters so that they can operate directly on AC mains.

Servers: These are very high performance, high speed computers developed with a single or multiple processors with very large capacity of RAM. They are provided with large capacity of magnetic memories for data back. They are mainly used to monitor and store information of multiple desktop computers connected as a network of computers.

Clusters and Warehouse-Scale Computers: Clusters are collections of desktop computers or servers connected by local area networks to act as a single larger computer. Each node runs its own operating system, and nodes communicate using a networking protocol. Anetwork of clusters is called **W**arehouse-**S**cale **C**omputer (WSC), in which tens of thousands of clusters can act as one system. The WSC is the foundation of Internet services many people use every day for search, social networking, online maps, video sharing, online shopping, email services, etc.

Embedded systems: They are small computers developed for dedicated application, like washing machine, printer, process control systems in industries, etc. They are designed with simple microcontrollers or processors as CPU and loaded with simplified operating system with limited features for specific applications.

1.2 Functional Units of a Digital Computer

The basic functional units of a digital computer are CPU (**C**entral **P**rocessing **U**nit), Memory, Input devices and Output devices. The basic block diagram of the functional units of a digital computer is shown in Fig. 1.1. From fourth generation of computer onwards the microprocessors are used as CPU.

The CPU is microprocessor which is a programmable IC. The microprocessor has an ALU (**A**rithmetic and **L**ogic **U**nit), set of registers and control and timing unit. The timing of various operations of the microprocessor are synchronized or governed by a clock. The microprocessor is also commonly (or simply) referred to as a processor.

The memories are semiconductor memories RAM and ROM that can directly work with microprocessor and SSD (**S**olid **S**tate **D**rive) and magnetic memories like hard disk for permanent storage.

A computer need an initialization program called boot program that has to be run whenever the computer is turned ON. The boot program is permanently stored in ROM memory. The program or software are permanently stored in hard disk or SSD and whenever needed the CPU will copy the software from hard disk or SSD to RAM memory and run the software only from RAM. The results that need permanent storage are stored in hard disk or SSD.

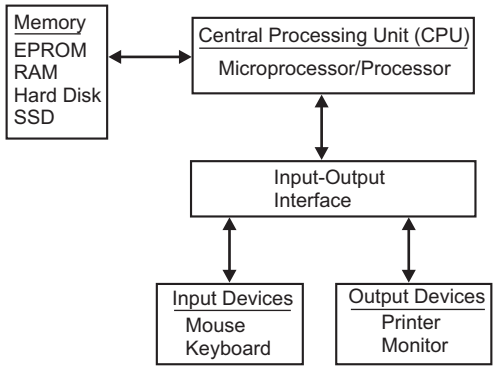

Fig. 1.1: *Basic functional block diagram of a digital computer.*

The input and output devices are provided for human communication and also for the external devices to communicate with CPU. The popular input devices are keyboard and mouse. The popular output devices are printer and monitor.

A computer needs an **O**perating **S**ystem (OS) for coordination of all activities of a computer. The operating system is software that is permanently stored (or installed) in hard disk or SSD and loaded in RAM memory whenever the computer is turned ON. The popular operating systems are Windows and Linux.

1.2.1 Computer Architecture and Organization

In describing computers, a distinction is made between ***computer architecture*** and ***computer organization***. The development of computer from software perspective is commonly known as ***computer architecture*** and from hardware perspective is called ***computer organization***.

Computer architecture refers to the features of a system as viewed by a programmer or the features that have a direct impact on the logical execution of a program. Examples of architectural features include the instruction set, the number of bits used to represent various data types (e.g., numbers, characters), the techniques used to access data stored in memory (memory addressing modes) and IO mechanisms.

Computer organization refers to the operational units and their interconnections that realize the architectural specifications. Organizational features include the hardware details transparent to the programmer, such as control signals, interfaces between the computer and peripherals and the memory technology used.

For example consider the implementation of multiplication operation in computers, it is an architectural design issue whether a computer will have a multiply instruction and it is an organizational issue whether that instruction will be implemented by a special multiply unit or by a mechanism that makes repeated use of the add unit of the system. The organizational decision may be based the computational speed and the cost of a special multiply unit.

Many computer manufacturers offer a family of computer models, all with the same architecture but with differences in organization. Consequently, the different models in the family have different price and performance characteristics.

In microcomputers, the relationship between architecture and organization is very close. Changes in technology influences organization and result in introduction of more powerful and complex architectures.

1.2.2 Instruction Set Architecture (ISA)

A term that is often used in synonymous with computer architecture is **I**nstruction **S**et **A**rchitecture (ISA). The ISA defines instruction formats, instruction opcodes, registers, instruction and data memory, the effect of executed instructions on the registers and memory and an algorithm for controlling instruction execution.

The processor (or microprocessor) that is used as a CPU of a computer will have a set of instructions designed by the manufacturer of the processor. Any computer program for a specific task or operation has to be written only using these instructions. The instruction set of a processor depends on the registers, memory, addressing methods and data types supported by the processor.

The development of instruction set of a processor is called ***instruction set architecture (ISA)***. Alternatively, the instruction set of a processor along with supporting data types and addressing modes is called ***Instruction Set Architecture (ISA***). The ISA completely utilize the hardware capability of a processor so that efficient software can be written for various applications.

Usually, software for various application programs are written using processor independent high level languages like C, C++, JAVA, Python, etc. These high level languages will have compilers to convert the software to machine language that can be understood by the processor.

1.2.3 RISC and CISC Instruction Set Architecture

RISC (**R**educed **I**nstruction **S**et **C**omputing/**C**omputer) and CISC (**C**omplex **I**nstruction **S**et **C**omputing/**C**omputer) are two types of processor instruction set architecture. They differ in the way of framing instructions to execute various tasks of computers.

RISC focuses on a smaller and more efficient set of instructions that can be executed quickly. CISC focuses on a broader range of instructions to perform more complex operations in fewer lines of code.

Features of RISC Architecture

- RISC architectures use a small, highly optimized set of instructions. Each instruction is designed to execute in a single clock cycle. This allows for a more efficient pipeline design.
- The simplicity of the instruction set allows for faster execution and easier hardware implementation.
- RISC architectures employ a load/store architecture in which the operations are performed on data in registers rather than directly in memory.
- The ARM architecture of embedded processors is example of RISC architectures.

Features of CISC Architecture

- CISC architectures have a larger set of instructions. CISC instructions can perform complex tasks in a single instruction. This can reduce the number of instructions per task/operation but may require more clock cycles to execute.

- CISC instructions can vary in length and complexity, which can complicate the hardware design of the processor.
- CISC architectures allow instructions to operate directly on memory, which can reduce the number of instructions needed for certain operations/tasks.
- The x86 architectures used in most personal computers is example of CISC architectures.

1.3 Basic Organization of a Computer

The digital computers from the year 1981 are personal computers (PC) or desktop and laptop computers developed using single chip processor as CPU. The organization of a typical PC using single chip processor as CPU is shown in Fig. 1.2.

The PC is built using a modern single chip processor as main processing unit, DRAM based semiconductor memory as main memory and an IO processor for interfacing IO devices. The CPU is a single chip processor consisting of ALU (**A**rithmetic **L**ogic **U**nit), Registers, Cache memory and Bus interface unit. A magnetic memory like hard disc is connected as secondary memory via DMA (**D**irect **M**emory **A**ccess). A system bus connects processor, DRAM main memory and IO processor. The IO processor consists of separate controllers for each IO device and the controllers are connected to a common local IO bus. The IO devices are directly connected to the dedicated controller in the IO processor.

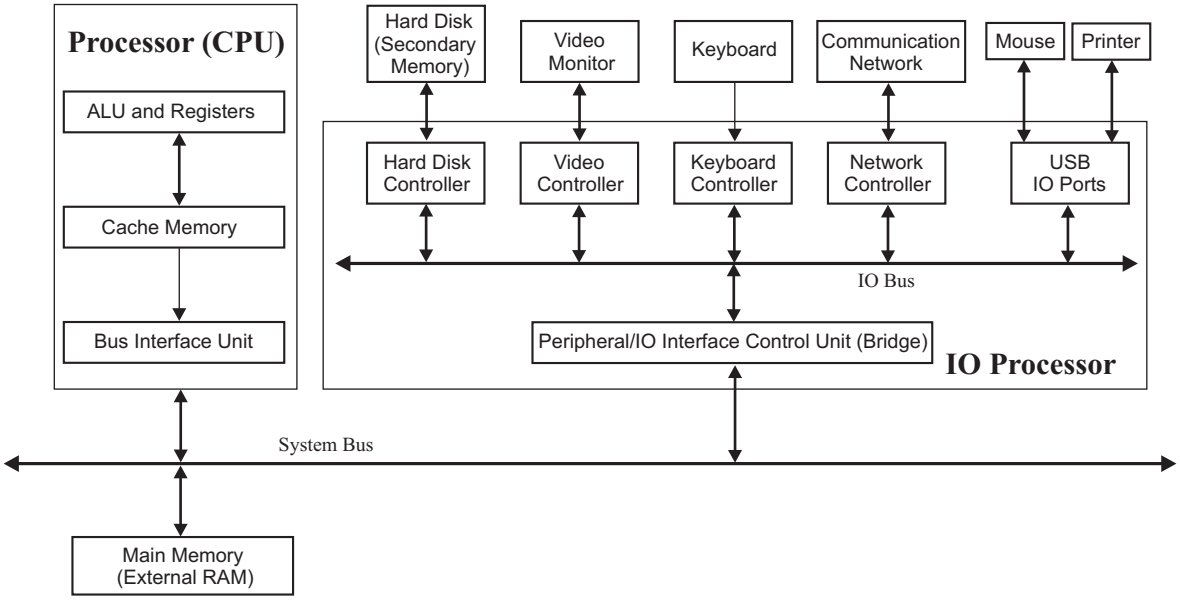

Fig. 1.2: *Basic organization of a personal computer.*

The IO bus is connected to system bus via a special bus-to-bus control unit referred to as bridge. The IO devices of a personal computer include a hard disk, keyboard, mouse, video monitor, printer, etc. The multimedia or audio-visual IO devices in PC are microphone, loudspeaker, camera, etc. The PC will have parallel and serial IO ports, which are replaced by USB ports in modern PC and ports for connecting computers to computer networks and internet, etc.

1.4 Basic Computer Architecture

The connection of various functional unit of a computer to perform the necessary operations from software point of view is called ***architecture***.

Three basic architectures have been proposed for computers.

- Von Neumann Architecture
- Harvard Architecture
- Modified Harvard Architecture

Basically the computer work on the concept of stored-program which means that the program is stored in a storage device (memory) and computer will execute the stored program. The various architecture basically defer in the way how the program and data are stored and retrieved.

1.4.1 Von Neumann Architecture

The Von Neumann architecture was proposed by John von Neumann in the year 1945. This architecture proposes a single storage (memory) for both program and data. At any one time either program or data can be fetched by CPU from storage device (memory). The basic functional block diagram of von Neumann architecture is shown in Fig. 1.3.

The Von Neumann architecture proposes a common address and data bus for accessing program, data and IO devices.

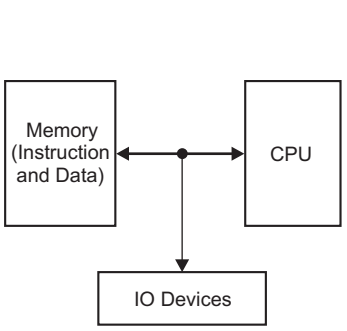

Fig. a: *Simplified block diagram.*

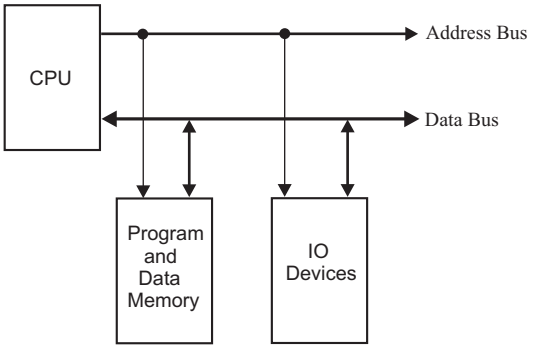

Fig. b: *Bus structure.*

Fig. 1.3: *Block diagram of Von Neumann architecture.*

1.4.2 Harvard Architecture

Harvard architecture was developed by Harvard University, USA in the year 1944. This architecture proposes separate storage (memory) for program and data. In this architecture, program and data can be fetched by CPU from the storage device (memory) simultaneously using two different paths (buses). The basic functional block diagram of Harvard architecture is shown in Fig. 1.4.

The Harvard architecture has separate buses for program memory and data memory facilitating the simultaneous access to instruction and data.

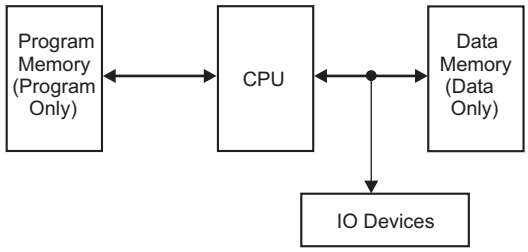

Fig. a: *Simplified block diagram.*

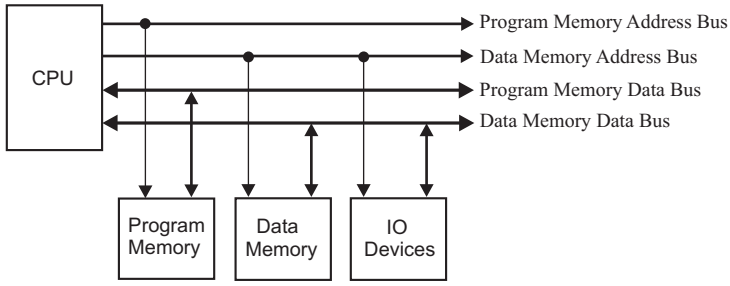

Fig. b: *Bus structure.*

Fig. 1.4: *Block diagram of Harvard architecture.*

1.4.3 Modified Harvard Architecture

The modified Harvard architecture evolved over time to suit the needs of microcontrollers and digital signal processors. The modified Harvard architecture proposes two (or more) storage devices (memories) with no strict separation between program and data storage. In the popular model, two storages are proposed, one for program and data and the other for only data. The basic functional block diagram of modified Harvard architecture is shown in Fig. 1.5.

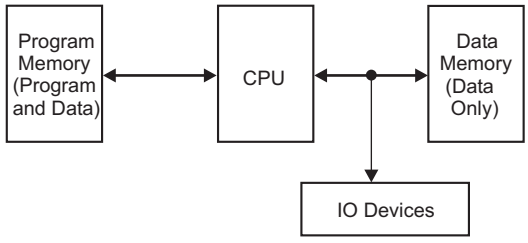

Fig. a: *Simplified block diagram.*

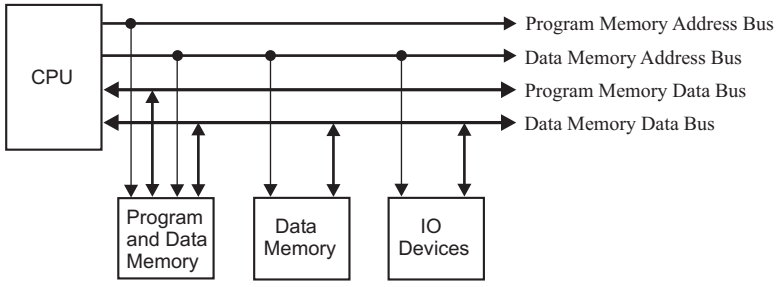

Fig. b: *Bus structure.*

Fig. 1.5: *Block diagram of modified Harvard architecture.*

1.5 Advanced Computer Architectures

The microprocessors (processors) are used as CPU from fourth generation computers onwards and hence the architecture of advanced computers are influenced by processor architecture and so some the advanced computer architecture names will be same as processor architectures. Some of the advanced computer architectures are,

- **Pipelined architecture:** A pipelined architecture divides the instruction execution into multiple stages. This allows multiple instructions to be processed concurrently by overlapping various stages of instruction execution.
- **Superscalar architecture:** A superscalar architecture will have multiple execution units and hence the superscalar processor can execute multiple instructions simultaneously within a single clock cycle.
- **VLIW (Very Long Instruction Word) architecture:** The VLIW architecture is a processor design that aims to achieve instruction level parallelism by encoding multiple instruction codes within a single instruction. This is achieved by grouping independent operations into a very long instruction word. Such instructions can be executed concurrently on multiple functional units within the processor.
- **Multicore architecture:** In multicore architecture the computers will employ processors with multiple cores within a single processor chip. Each core is a processor itself with its own cache memory and all the cores will share the common resources of the computer. Each core can execute instructions/programs independently and in parallel, allowing for enhanced performance and multitasking capabilities.
- **Multiprocessor architecture:** In multiprocessor architecture the computers will have multiple processors connected in a particular bus configuration that share access to a common memory and peripherals. This allows the system to execute multiple tasks or instructions simultaneously, enhancing processing speed and overall performance.

1.5.1 Flynn's Classification of Computer architectures

Flynn's Classification is a method used to categorize parallel processing computer architectures based on the number of instruction streams and data streams that can be handled by computers simultaneously. Flynn's Classification was introduced by Michael J. Flynn in 1966 to classify different types of computer systems according to their parallel processing capabilities and it is still valid for modern multiprocessor and multicore processor based architectures. Flynn's Classification includes the following four primary categories.

- SISD (Single Instruction stream, Single Data stream)
- SIMD (Single Instruction stream, Multiple Data streams)
- MISD (Multiple Instruction streams, Single Data stream)
- MIMD (Multiple Instruction streams, Multiple Data streams)

SISD (Single Instruction stream, Single Data stream)

In SISD architecture, a single processor executes one instruction at a time on a single data stream. There is no parallelism, meaning the CPU processes instructions sequentially. The early computers and single-core processors are examples of SISD architecture. The SISD architecture supports sequential tasks like legacy software, word processors, and basic applications that don't require parallel processing.

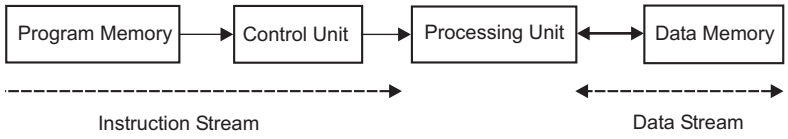

Fig. a: *SISD.*

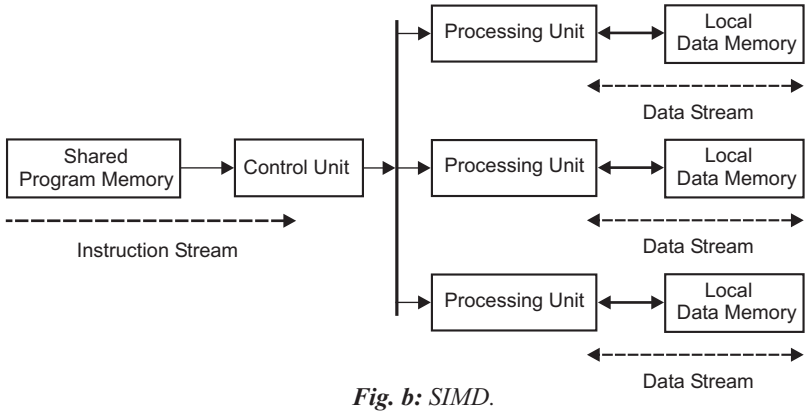

Fig. b: *SIMD.*

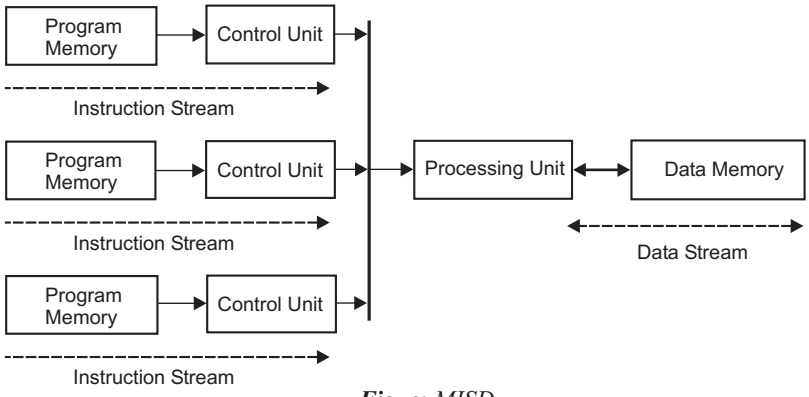

Fig. c: *MISD.*

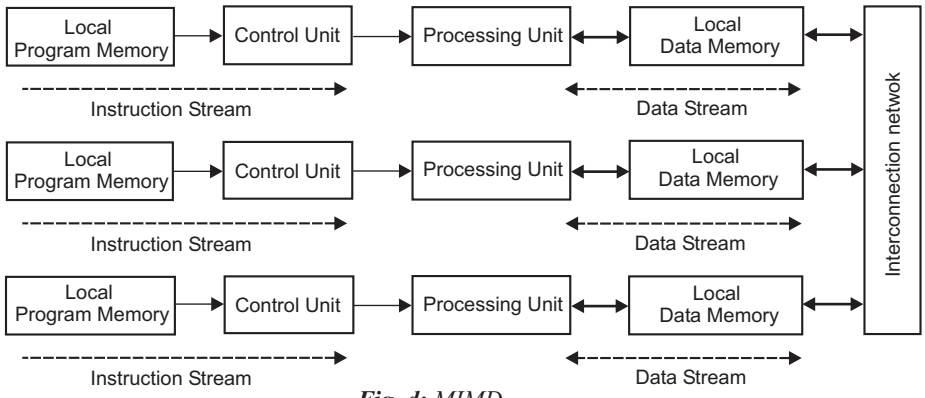

Fig. d: *MIMD.*

Fig. 1.6: *Flynn's classification of computer architectures.*

SIMD (Single Instruction stream, Multiple Data streams)

SIMD architectures allow a single instruction to be executed on multiple data elements simultaneously. This is a form of data-level parallelism, where the same operation is performed on different pieces of data at the same time. Examples of SIMD architectures are vector processors, modern GPUs (**G**raphics **P**rocessing **U**nits) and some specialized processors used for scientific computing, graphics rendering and machine learning. The SIMD architecture supports image processing, signal processing, simulations and tasks where the same operation is performed on large datasets (e.g., matrix multiplication, video encoding/decoding).

MISD (Multiple Instruction streams, Single Data stream)

MISD systems execute multiple instruction streams, but only one data stream is processed. This type of architecture is rare and not commonly used in practice. MISD systems are used in fault-tolerant applications where multiple processing units perform different computations on the same data for redundancy and hence high reliability (e.g., space missions).

MIMD (Multiple Instruction streams, Multiple Data streams)

MIMD architectures involve multiple processors, each executing its own instruction stream on different data streams. MIMD is a form of task-level parallelism, where different processors work independently on different parts of the problem. MIMD architectures include modern multi-core processors, supercomputers, and distributed systems like clusters or grids. MIMD architecture is useful for complex computations, server environments, web hosting, scientific simulations, big data processing and any application requiring true multitasking or distributed processing.

1.6 Computer Performance Benchmarks

Benchmarks are used to evaluate the performance of the computers by running a set of standard programs or workloads on a computer system. The benchmarks are used to compare different architectures and assess real-world performance. Benchmarks are also developed to test specific components like CPU, GPU(**G**raphics **P**rocessing **U**nit), memory, etc.

Whetstone and ***Dhrystone*** are two computer performance benchmarks developed in the 1980s to evaluate the performance of processors and their ability to handle integer and floating-point operations.

Whetstone and ***Dhrystone*** benchmarks are now considered out dated for assessing the performance of modern systems, especially those with complex processors that include features like vectorization, multi-core processing, and hardware acceleration. However, they remain historically important. In different period of time, modern benchmarks, such as ***SPEC CPU, Geekbench***, LINPACK, PCmark & 3Dmark and PassMark are developed by different companies to evaluate modern computers more effectively.

1.6.1 Whetstone Benchmark

The ***Whetstone*** benchmark is designed to test a system's floating-point arithmetic performance. It was created in 1974 and is primarily used to measure the performance of the CPU when executing floating-point operations, which are important in scientific, engineering and graphics calculations.

Key Features

- **Focus on Floating-Point Operations:** Whetstone evaluates a system's ability to perform operations like addition, subtraction, multiplication and division using floating-point numbers. This is significant for tasks that require high precision in calculations, such as simulations and scientific computations.
- **Test Program:** Whetstone consists of a set of 14 subroutines, each designed to execute a specific set of floating-point operations (e.g., trigonometric functions, array manipulations).
- **Score:** The benchmark produces a score based on the number of iterations the system can perform per second, known as "Whetstone MIPS" (**M**illion **I**nstructions **P**er **S**econd). A higher score indicates better floating-point performance.

1.6.2 Dhrystone Benchmark

The ***Dhrystone*** benchmark is another classic performance benchmark, but it focuses on testing integer performance. It was introduced in 1984 by Reinhold Weicker and is widely used to measure the performance of computers when performing operations like string manipulation, loops and simple arithmetic.

Key Features

- **Focus on Integer Operations:** Dhrystone emphasizes integer operations like comparisons, additions and multiplications on integer data. This makes it more representative of workloads typical in office applications, operating systems and database processing, which often rely on integer operations.
- **Test Program:** Dhrystone runs a series of tasks such as conditional branches, function calls and arithmetic operations using integer data types. The code is designed to mimic the general-purpose workload found in business applications.
- **Score:** The benchmark produces a result in terms of "Dhrystone MIPS" (**M**illion **I**nstructions **P**er **S**econd), which reflects how many iterations of the benchmark the system can complete per second. A higher Dhrystone MIPS value indicates better integer performance.

1.6.3 SPEC (Standard Performance Evaluation Corporation) Benchmark

The **S**tandard **P**erformance **E**valuation **C**orporation (SPEC) is a non-profit consortium that establishes and maintains standardized benchmarks and performance evaluation tools for new generations of computing systems. SPEC was founded in 1988 and its membership comprises computer hardware and software vendors, educational institutions, research organizations and government agencies internationally. SPEC benchmarks and tools are widely used to evaluate the performance of computer systems and the test results are published on the SPEC website.

1.6.4 Geekbench Benchmark

Geekbench is a proprietary utility program for benchmarking the **C**entral **P**rocessing **U**nit (CPU) and **G**raphics **P**rocessing **U**nit (GPU) of computers, laptops, tablets and Cell phones, founded in 2007, by Primate Labs.

The Geekbench benchmark tests the GPU's ability to perform complex computations, particularly those relevant to graphics processing, AI/ML and other GPU-accelerated tasks. It provides a score based on the GPU's performance in these workloads, allowing users to compare the performance of different GPUs across various platforms.

1.7 Computer Performance Measurement

The computer Performance measurements are time-based metrics defined to evaluate how quickly and efficiently a computer system or processor completes tasks or executes programs. These metrics help in comparing systems and guiding design decisions to improve overall system efficiency.

The various performance measures are,

- Clock Rate (Frequency) and Clock Cycle Time.
- CPI (Cycles Per Instruction) and Execution Time (CPU Time).
- MIPS (Millions of Instructions Per Second).
- FLOPS (Floating Point Operations Per Second).
- Throughput and Latency.
- Power and Efficiency.
- Amdahl's Law.

Clock Rate (Frequency) and Clock Cycle Time

Every computer/processor is operated by a clock, which is a periodic square pulse train with a fixed frequency called clock rate. The inverse of clock rate is clock cycle time.

$$\text{Clock Cycle Time} = \frac{1}{\text{Clock Rate}}$$

The basic unit of clock rate (frequency) is Hertz (Hz) but the modern processors clock rate will be in the order of GHz (Giga-Hz or 10^9 Hz) and hence the clock cycle time will be in the order of ns (nano-seconds or 10^{-9} seconds).

The clock rate is dictated by manufacturing process followed for making processor IC. While comparing various computers clock rate, the higher clock rate indicates fastness in computation.

CPI (Cycles Per Instruction) and Execution Time (CPU Time)

Every processor instruction will be executed in a definite number of clock cycles and it is specified as cycles per instruction.

Execution Time (CPU Time) refers to total time required by a computer to execute a computer program. While comparing various computers execution time, the smaller CPU time indicates better performance.

$$\text{CPU Time} = \text{Instruction Count} \times \text{CPI} \times \text{Clock Cycle Time}$$

$$= \frac{\text{Instruction Count} \times \text{CPI}}{\text{Clock Rate}}$$

MIPS (Millions of Instructions per Second)

Since processor clock cycle time is in the order of nano-seconds, it is possible to calculate how many millions of instructions are executed in one second. It's a measure of speed of program execution. While comparing various computers MIPS rate, the higher MIPS rate indicates fastness in computation.

$$\text{MIPS} = \frac{\text{Instruction Count}}{\text{Execution Time} \times 10^6}$$

FLOPS (Floating Point Operations per Second)

The scientific calculation by processor involve number of floating point operations. Hence a measure of number of floating point operations is specified to compare the fastness of performing floating point operations.

$$\text{FLOPS} = \frac{\text{Number of Floating Point Operations}}{\text{Execution Time}}$$

Throughput and Latency

The throughput is a performance measure defined to measure number of programs or operations completed per unit time. It's a useful measure in comparison of various parallel computing computers. Higher throughput is a performance measure of better computer architecture.

The latency refers to the time from start to finish of a single task. Lower latency implies faster response time.

Power and Efficiency

Power refers to the amount of electrical energy consumed by the processor measured in ***watts (W)***. Efficiency refers to power efficiency and it is the work done by a processor for every unit of power it consumes. Efficiency is a measure of ***performance-per-watt***. The performance refers to MIPS or FLOPS.

Total power in CMOS processors can be divided into following three components,

$$P_{total} = P_{dynamic} + P_{short-circuit} + P_{static}$$

The Dynamic Power is defined as,

$$P_{dynamic} = \alpha C_L V^2 f$$

α = Activity factor (fraction of transistors switching per clock cycle, $0 < \alpha \leq 1$)

C_L = Load capacitance

V = Supply voltage

f = Clock frequency

The short-circuit power is due to very small conducting period of NMOS and PMOS during switching and static power is due to leakage current. Short-circuit and static powers are negligible when compared to dynamic power. Hence, the dynamic power can be considered as total power by neglecting short-circuit and static power.

$$\therefore P_{total} = P_{dynamic} = \alpha C_L V^2 f$$

The processor power efficiency is typically defined as,

$$\text{Efficiency} = \frac{\text{Performance}}{\text{Power}}$$

When performance is measured in MIPS,

$$\text{Efficiency} = \frac{\text{MIPS}}{P_{total}}$$

When performance is measured in Flops,

$$\text{Efficiency} = \frac{\text{FLOPS}}{P_{total}}$$

Amdahl's Law:

Amdahl's Law defines the speed up factor in multi-processor systems. Amdahl's law was first proposed by Gene Amdahl in 1967 to estimate the speedup of a program using multiple processors compared to a single processor. Consider a program running on a single processor such that a fraction "(1 – f)" of the execution time involves code that is sequential, and a fraction "f" that involves code that is parallelizable. Let T, be the total execution time of the program using a single processor.

The speedup using a parallel processor with N processors that fully exploits the parallel portion of the program is given by,

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

$$= \frac{T[(1-f)+f]}{T\left[(1-f)+\frac{f}{N}\right]} = \frac{1}{(1-f)+\frac{f}{N}}$$

The following two conclusions can be drawn from Amdahl's law,

- When f is small, the use of parallel processors has little effect.
- When N is large, speedup is given by 1/(1 - f).

1.8 Summary of Important Concepts

1. A computer accepts text/graphical information as input and process the input information based on a program and generates output text/graphical information.
2. IBM (International Business Machine), USA released a personal computer in 1981 with INTEL 8088 processor as CPU which is considered as bench mark for all digital computers of modern age.
3. The digital computers from the year 1981 are Personal Computers (PC) or desktop and laptop computers developed using single chip processor as CPU.
4. The basic functional units of a digital computer are CPU (Central Processing Unit), Memory, Input devices and Output devices.
5. Computer organization refers to the operational units and their interconnections that realize the architectural specifications.
6. The instruction set of a processor along with supporting data types and addressing modes is called Instruction Set Architecture (ISA).
7. The connection of various functional unit of a computer to perform the necessary operations from software point of view is called architecture.
8. Three basic architectures of computers are Von Neumann architecture, Harvard architecture and modified Harvard architecture.

9. The Von Neumann architecture was proposed by John von Neumann in the year 1945.
10. Von Neumann architecture proposes a common address and data bus for program, data and IO devices.
11. Harvard architecture was developed by Harvard University, USA in the year 1944.
12. Harvard architecture proposes separate storage (memory) for program and data.
13. Harvard architecture has separate buses for program memory and data memory facilitating the simultaneous access to instruction and data.
14. The modified Harvard architecture proposes two (or more) storage devices (memories) with no strict separation between program and data storage.
15. RISC (Reduced Instruction Set Computing/Computer) and CISC (Complex Instruction Set Computing/Computer) are two types of processor instruction set architecture.
16. RISC focuses on a smaller and more efficient set of instructions that can be executed quickly.
17. CISC focuses on a broader range of instructions to perform more complex operations in fewer lines of code.
18. A pipelined architecture divides the instruction execution into multiple stages.
19. A superscalar architecture will have multiple execution units and hence the superscalar processor can execute multiple instructions simultaneously within a single clock cycle.
20. The VLIW architecture is a processor design that aims to achieve instruction level parallelism by encoding multiple instruction codes within a single instruction.
21. In multicore architecture the computers will employ processors with multiple cores within a single processor chip.
22. In multiprocessor architecture the computers will have multiple processors connected in a particular bus configuration that share access to a common memory and peripherals.
23. Flynn's Classification is a method used to categorize parallel processing computer architectures based on the number of instruction streams and data streams that can be handled by computers simultaneously.
24. Flynn's Classification includes four primary categories: SISD (Single Instruction stream, Single Data stream), SIMD (Single Instruction stream, Multiple Data streams), MISD (Multiple Instruction streams, Single Data stream) and MIMD (Multiple Instruction streams, Multiple Data streams).
25. In SISD architecture, a single processor executes one instruction at a time on a single data stream.
26. SIMD architectures allow a single instruction to be executed on multiple data elements simultaneously.
27. MISD systems execute multiple instruction streams, but only one data stream is processed.
28. MIMD architectures involve multiple processors, each executing its own instruction stream on different data streams.
29. Benchmarks are used to evaluate the performance of the computers by running a set of standard programs or workloads on a computer system.
30. The Whetstone benchmark is designed to test a system's floating-point arithmetic performance.
31. The Dhrystone benchmark is focuses on testing integer performance.
32. The Standard Performance Evaluation Corporation (SPEC) is a non-profit consortium that establishes and maintains standardized benchmarks and performance evaluation tools for new generations of computing systems.

33. The computer Performance measurements are time-based metrics defined to evaluate how quickly and efficiently a computer system or processor completes tasks or executes programs.
34. A clock is a periodic square pulse train with a fixed frequency called clock rate. The inverse of clock rate is clock cycle time.
35. Every processor instruction will be executed in a definite number of clock cycles and it is specified as cycles per instruction.
36. Execution time (CPU time) refers to total time required by a computer to execute a computer program.
37. MIPS (Millions of Instructions per Second) and FLOPS (Floating Point Operations per Second) rate indicates fastness in computation.
38. The throughput is a performance measure defined to measure number of programs or operations completed per unit time.
39. The latency refers to the time from start to finish of a single task. Lower latency implies faster response time.
40. Amdahl's law defines the speed up factor in multi-processor systems.

1.9 Short-Answer Questions

Q1.1 List the functional units of a digital computer.

The functional units of a digital computer are,

- CPU (Central Processing Unit)
- Memory
- Input devices
- Output devices

Q1.2 Differentiate computer architecture and organization.

The development of computer from software perspective is commonly known as computer architecture and from hardware perspective is called computer organization.

Q1.3 What is instruction set architecture.

The instruction set of a processor along with supporting data types and addressing modes is called Instruction Set Architecture (ISA). The ISA defines instruction formats, instruction opcode, registers, instruction and data memory, the effect of executed instructions on the registers and memory and an algorithm for controlling instruction execution.

Q1.4 What is software and hardware?

Software is a set of instructions or commands needed for performing a specific task by a programmable device or a computing machine.

The hardware refers to the components or devices used to form computing machine in which the software can be run and tested. Without software the hardware is an idle machine.

Q1.5 What is meant by a program?

A program is a set of instructions written to perform a certain task.

Q1.6 List the features of RISC architecture.

- RISC architectures use a small, highly optimized set of instructions. Each instruction is designed to execute in a single clock cycle.
- The simplicity of the instruction set allows for faster execution and easier hardware implementation.
- RISC architectures employ a load/store architecture in which the operations are performed on data in registers rather than directly in memory.

Q1.7 List the features of CISC architecture.

- CISC architectures have a larger set of instructions. CISC instructions can perform complex tasks in a single instruction.
- CISC instructions can vary in length and complexity, which can complicate the hardware design of the processor.
- CISC architectures allow instructions to operate directly on memory, which can reduce the number of instructions needed for certain operations/tasks.

Q1.8 List the features of Whetstone benchmark.

- **Focus on Floating-Point Operations:** Whetstone evaluates a system's ability to perform operations like addition, subtraction, multiplication, and division using floating-point numbers.
- **Test Program:** Whetstone consists of a set of 14 subroutines, each designed to execute a specific set of floating-point operations (e.g., trigonometric functions, array manipulations).
- **Score:** The benchmark produces a score based on the number of iterations the system can perform per second, known as "Whetstone MIPS" (Million Instructions per Second).

Q1.9 List the features of Dhrystone benchmark.

- **Focus on Integer Operations**: Dhrystone emphasizes integer operations like comparisons, additions and multiplications on integer data.
- **Test Program:** Dhrystone runs a series of tasks such as conditional branches, function calls, and arithmetic operations using integer data types.
- **Score:** The benchmark produces a result in terms of "Dhrystone MIPS" (Million Instructions per Second).

Q1.10 Write a short note on Geekbench benchmark.

Geekbench is a proprietary utility program for benchmarking the Central Processing Unit (CPU) and Graphics Processing Unit (GPU) of computers, laptops, tablets, and cell phones, founded in 2007, by Primate Labs. The Geekbench benchmark tests the GPU's ability to perform complex computations, particularly those relevant to graphics processing, AI/ML and other GPU-accelerated tasks.

Q1.11 What are clock rate and clock cycle time?

Every computer/processor is operated by a clock, which is a periodic square pulse train with a fixed frequency called clock rate. The inverse of clock rate is clock cycle time.

Q1.12 What are CPI and CPU time?

Every processor instruction will be executed in a definite number of clock cycles and it is specified as CPI (Cycles per Instruction).

1.10 Exercises

I. Fill in the blanks

1. The world's first programmable general-purpose electronic computer is _____.
2. The first commercial microprocessor _____ was released by Intel Corporation, USA in the year 1971.
3. The _____ is permanently stored in ROM memory.
4. The popular input devices are _____ and _____.
5. The popular output devices are _____ and _____.
6. The development of computer from software perspective is commonly known as _____.
7. The development of computer from hardware perspective is called _____.
8. A magnetic memory like hard disc is connected as secondary memory via _____.
9. _____ architecture proposes a single storage (memory) for both program and data.
10. _____ architecture proposes separate storage (memory) for program and data.
11. _____ architectures involve multiple processors, each executing its own instruction stream on different data streams.
12. _____ defines the speed up factor in multi-processor systems.

Answers

1. ENIAC	7. computer organization
2. INTEL4004	8. DMA
3. boot program	9. Von Neumann
4. keyboard, mouse	10. Harvard
5. printer, monitor	11. MIMD
6. computer architecture	12. Amdahl's law

II. State whether the following statements are True or False

1. First commercial processor INTEL4004 was released by Intel Corporation, USA in the year 1971.
2. The computers developed using Integrated Circuits manufactured using SSI/MSI technology are referred to as second generation computers.
3. The boot program is permanently stored in RAM memory.
4. The popular input devices are printer and monitor.
5. In Harvard architecture, program and data can be fetched by CPU from the storage device (memory) simultaneously using two different paths (buses).
6. RISC focuses on a smaller and more efficient set of instructions that can be executed quickly.
7. CISC focuses on a small set of instructions to perform more complex operations in fewer lines of code.
8. The Whetstone benchmark is designed to test a system's floating-point arithmetic performance.
9. While comparing various computers execution time, the larger CPU time indicates better performance.
10. The throughput is a performance measure defined to measure number of operations completed per unit time.
11. The latency refers to the time from start to finish of a single task. Higher latency implies faster response time.
12. Amdahl's law defines the speed up factor in single processor systems.

Answers

1. True
2. False
3. False
4. False
5. True
6. True
7. False
8. True
9. False
10. True
11. False
12. False

III. Answer the following questions

1. Write a description about computer generations.
2. Explain the functional units of a digital computer with a simple sketch.
3. Draw the block diagram of basic organization of a personal computer and explain.
4. Draw the block diagram of different computer architectures and explain.
5. List some of the advanced computer architectures.
6. What is meant by RISC and CISC computer architecture?
7. How the computers can be classified based on market requirements?
8. Explain Flynn's classification of computers.
9. What are Whetstone and Dhrystone benchmark?
10. List the various performance measures of computers.
11. What are MIPS and FLOPS?
12. What is Amdahl's law?