# 1

# Computer-History, Classification and Basic Anatomy

## 1.1 GENERATIONS OF COMPUTER

The first electronic digital computer, called the Atanasoff–Berry Computer (ABC), was built by Dr John V. Atanasoff and Clifford Berry in 1937. An electronic computer called the Colossus was built in 1943 for the US army. Around the same time, many others were also trying to develop computers. The first general-purpose digital computer, the Electronic Numerical Integrator and Computer (ENIAC), was built in 1946.

Computers since 1946 are categorized in five generations:

- First Generation: Vacuum Tubes
- Second Generation: Transistors
- Third Generation: Integrated Circuits
- Fourth Generation: Microprocessors
- Fifth Generation: Artificial Intelligence

Follows a brief description of each generation:

### 1.1.1 First Generation (1946–1956) Vacuum Tubes

Vacuum tubes were used to make circuits of first generation computers. For building memory, magnetic drums were used that were huge in size and weight. First generation computers were so large in size that they often took an entire room. They were also very prone to error. They were too expensive to operate and in addition consumed huge electricity. It is worth mentioning the amount of heat they generated. Despite using liquid based cooling system, they often got damaged due to heat.

Programs for first generation computers were written in machine language, the lowest-level programming language understood by computers, to perform operations. They were designed to solve only one problem at a time. Punched cards and paper tapes were used to feed input, and output was displayed on printouts.

The ENIAC (**Electronic Numerical Integrator and Computer)** and **UNIVAC** (**Electronic Discrete Variable Automatic Computer)** computers are examples of first-generation computing devices.

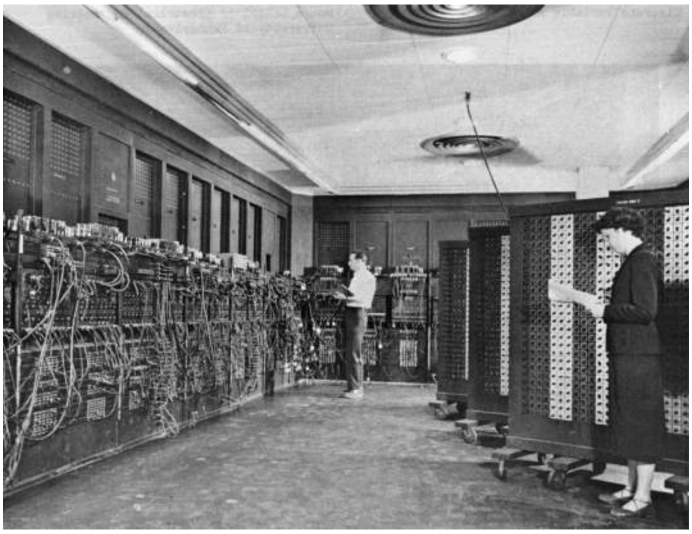

**Fig. 1.1 ENIAC**

ENIAC was the first electronic general-purpose computer. It was capable of being reprogrammed to solve various numerical problems. ENIAC was primarily designed to calculate artillery firing tables. It was mainly used in the United States Army's Ballistic Research Laboratory. ENIAC was introduced to the public at the University of Pennsylvania in 1946 as "Giant Brain." ENIAC' was funded by the United States Army.

ENIAC had a modular design. It had individual panels to perform separate functions. Twenty modules among them were accumulators that could add, subtract and hold a ten-digit decimal number in memory. Numbers were passed between these modules through several general-purpose buses. The modules were able to send and receive numbers, compute, save the answer and trigger the next operation without any moving component. That is why it could achieve high speed. Key to its versatility was the ability for branching. It could switch to different operations, depending on the sign of a computed result.

ENIAC contained 17,468 vacuum tubes, 1500 relays, 70,000 resistors, 7200 crystal diodes, 10,000 capacitors. It had a whopping 5,000,000 hand-soldered joints. It weighed more than 27 tons, was roughly 8 × 3 × 100 feet in size, occupied 1800 $ft^2$ and consumed 150 kW of electricity.

After ENIAC, a much improved computer named EDVAC (Electronic Discrete Variable Automatic Computer) was designed. EDVAC was a stored program computer. EDVAC was the first computer to work in binary number system. This is a major difference with ENIAC that used decimal number system.

### 1.1.2 Second Generation (1956–1963) Transistors

Transistors replaced vacuum tubes in the second generation of computers. The transistor was far superior to the vacuum tube in terms of size, generated heat and energy consumption. So computers made using transistors became smaller, faster, cheaper, more energy-efficient and more reliable.

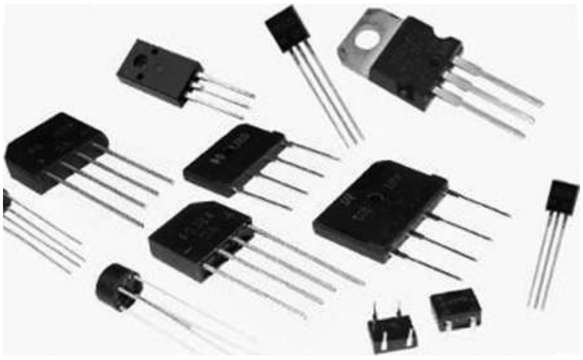

**Fig. 1.2** **Transistor**

The transistors also generated a lot of heat that subjected the computer to damage. But it was much better than the vacuum tubes in terms of size and heat.

Second-generation computers used symbolic or assembly languages instead of binary machine language that allowed programmers to specify instructions in words instead of machine code or binary.

Second-generation computers were still using punched cards for input and print-outs for output. At this time, high-level programming languages, like early versions of COBOL and FORTRAN were being developed. The first computers that were developed around this time were used in the atomic energy industry.

### 1.1.3 Third Generation (1964–1971) Integrated Circuits

Third generation computers were based on integrated circuits (IC circuits). Transistors were much smaller. They were placed on silicon chips, called semiconductors. This invention dramatically increased the speed and efficiency of computers.

In these computers, users interacted through keyboards and monitors and interfaced with an operating system. Many different applications were possible to run at one time. A central program usually resided in memory to monitor others. Computers for the first time became accessible to a mass audience because they became cheaper and smaller.

### 1.1.4 Fourth Generation (1971–Present) Microprocessors

Brain of the fourth generation of computers is **microprocessors.** Thousands of integrated circuits were built on a single silicon chip. Computers of the size of an entire room in first generation could now fit in palm. The Intel 4004 chip, developed in 1971, contained all the components of the computer like the central processing unit, memory and input/output controls on a single chip.

In 1981 IBM introduced its first computer for the home user, and in 1984 Apple introduced the Macintosh. Microprocessors also moved out of the realm of desktop computers and into many areas of life as more and more everyday products began to use microprocessors.

As these small computers became more powerful, they could now be linked together to form networks, which eventually led to the development of the Internet. Fourth generation computers were also equipped with the mouse and other handheld devices. Graphical user interface (GUI) was also designed for these computers.

### 1.1.5 Fifth generation (Present and Beyond) Artificial Intelligence

Fifth generation computing devices are based on artificial intelligence and are still being developed, while some applications, like voice and handwriting recognisers, are in use today. Fifth-generation computing aims to develop devices that will be responsive to natural language input and will also be able to learn and self-organize.

### 1.1.6 Evolution of Intel Processors

As Intel processors or compatible processors are most popular for desktop systems, it is worth mentioning the various microprocessors introduced by Intel. The first processor available to public was 8085 having 40 pin, 2 MHz clock frequency, 6500 transistors, 8 bit data bus and 16 bit address bus. But first personal computer made by IBM with Intel microprocessor was based on 8086 having 1 MB addressable memory and 30000 transistors. Then came 80286 followed by 386SX, 386DX, 486 SX and 486 DX.

The 486 processors contained up to 1.4 million transistors, reached 100 MHz frequency and can address up to 4 Gigabyte of memory (32 bit address bus so $2^{32}$ = 4 GB). After 486, came the decade of Pentium, Pentium with MMX (multimedia extended) and Pentium II processors. Then the Pentium III processor touched the 1 GHz clock frequency mark. From mid of year 2000 the market leader was various versions of Pentium IV for steady six years, achieved clock frequency up to 3.8 GHz and packed up to whopping 180 million transistors in it. Presently core-i3, core-i5 and core-i7 are ruling the market though Intel dual core processors are also available. (In this span, Intel marketed few other processors but they did not get good response.)

## 1.2 CLASSIFICATION OF COMPUTERS

Computers can be classified in many ways depending on various features and criteria. Many of them will be mentioned in this section. Though readers of this are being introduced to computer and do not have an in-depth knowledge of computer architecture, some classifications based on various architectural difference are also being listed. This will build interest of the readers in the subjects such as ‘MICROPROCESSOR’, ‘ARCHITECTURE’, ‘COMPUTER ORGANIZATION’, ‘PARALLEL PROCESSING’, ‘ASSEMBLY LANGUAGE PROGRAMMING’, and ‘PIPELINING’.

**Note** A computer's **architecture** is its abstract model and is the programmer's view in terms of instructions, addressing modes and registers. A computer's **organization** expresses the realization of the architecture. Architecture describes **what** the computer does and organization describes **how** it does it.

Here in this topic, following types of classifications will be discussed:

i. Based on computation power.
ii. Based on number of operands.
iii. Scalar VS Vector processor.

iv. Flynn's taxonomy
v. von Neumann vs. Harvard Architecture
vi. Big-endian vs. Little-endian

### (i) Based on Size, Computing Power and Price

Based on their size, computing power and price the computers are broadly classified into four categories:

1. Microcomputers
2. Minicomputers
3. Mainframe computers
4. Supercomputer.

As the list descends, things will become large, expensive, complex and fast. Now follows a small introduction to all of them:

**(1) Microcomputers** are the most common kind of computers used by people today, whether in a workplace, at school or on the desk at home. These are microprocessor based systems. Microcomputers are small, low-cost and single-user digital computer. They consist of CPU, input unit, output unit, storage unit and the software. Microcomputers are stand-alone machines but they can be connected together to create a network of computers that can serve more than one user. Microcomputers include desktop computers, notebook computers or laptop, tablet computer, handheld computer, smart phones, etc. The most popular microprocessors in world are made by Intel and Apple for their Pentium based PCs and Apple Macintosh.

Here the evolution of Intel microprocessors will be worth mentioning. Intel first introduced the 8085 microprocessor having 8 bit data bus and 16 bit address bus but no personal computer was made based on it. The journey started with 8086 when IBM made IBM-AT based on it. Gradually came 80286, 386, 486 SX and 486 DX. Next processor was named Pentium with clock frequency 60 MHz to 300 MHz and introduced in the period 1993–1997. It had a 32 bit data bus and 32 bit address bus. The speed and internal cache memory size increased with time. Pentium II, Pentium III and Pentium IV were the successors. Then came Pentium DUAL CORE, Core2-DUO, and Quad-core processors having frequency in GHz level. Now the latest processors from Intel's stable are core-I3, core-I5 and core-I7.

**(2) Minicomputer** is a class of multi-user computers that lies in the middle range of the computing spectrum, in between the top end single-user systems (microcomputers or personal computers) and low end mainframe computers. They have high processing speed and high storage capacity than the microcomputers. Minicomputers can support up to 200 users simultaneously. The users can access the minicomputer through their PCs or terminal. They are used for various processor hungry applications in industries and research centers.

**(3) Mainframe computers** are multi-user, multi-programming and high performance computers. They operate at a very high speed, have huge storage capacity and can handle the workload of many users. The user accesses the mainframe computer via a terminal that may be a dumb terminal, an intelligent terminal or a PC. The terminals and PCs utilize the processing power and the storage facility of the mainframe computer. Mainframe computers are used primarily by corporate and governmental organizations for critical applications, bulk data processing such as census, industry and consumer statistics, enterprise resource planning and transaction processing.

**(4) Supercomputers** are the fastest and the most expensive machines. They have high processing speed compared to other computers. The speed of a supercomputer is generally measured in FLOPS (Floating point Operations Per Second). Some of the faster supercomputers can perform trillions of calculations per second. Supercomputers are built by interconnecting thousands of processors that can work in parallel. Supercomputers are used for highly calculation-intensive tasks, such as, weather forecasting, climate research, molecular research, biological research, nuclear research and aircraft design. Some examples of supercomputers are IBM Roadrunner, IBM Blue gene. The supercomputer assembled in India by C-DAC (Center for Development of Advanced Computing) is PARAM. PARAM Padma is the latest machine in this series. The peak computing power of PARAM Padma is One Tera FLOP.

### *(ii) Based on Number of Operands*

The instruction-set of a computer is designed first, and then accordingly the architecture is designed. Just like a KEY is designed first then corresponding LOCK is assembled. A computer instruction should contain one operation code and zero or more operands on which the operation will be performed or where the result will be stored. Depending on maximum number of operands allowed in an instruction, computers can be classified:

1. Zero address machines (stack machine)
2. One address machine (accumulator machine)
3. Two address machine (Intel processors are of this type)
4. Three address machine

### *(iii) Scalar Processor vs Vector Processor*

Each instruction executed by a scalar processor generally manipulates one or two data items at a time. On the contrary, instructions of a vector processor operate simultaneously on many data items.

### *(iv) By Flynn's Taxonomy*

Flynn's taxonomy is a classification of computer architectures, proposed by Michael J. Flynn in 1966. There are four categories defined by Flynn. His classification is based upon the number of concurrent instruction (or control) and data streams available in the architecture:

**1. Single Instruction, Single Data stream (SISD)** A sequential computer which exploits no parallelism in either the instruction or data streams. Single control unit (CU) fetches single Instruction Stream (IS) from memory. The CU then generates appropriate control signals to direct single processing element to operate on single Data Stream.

**2. Single Instruction, Multiple Data streams (SIMD)** A computer which exploits multiple data streams against a single instruction stream to perform operations which may be naturally parallelized. For example, an array processor or GPU.

**3. Multiple Instruction, Single Data stream (MISD)** Multiple instructions operate on a single data stream. Uncommon architecture which is generally used for fault tolerance.

**4. Multiple Instruction, Multiple Data streams (MIMD)** Multiple autonomous processors simultaneously executing different instructions on different data. Distributed systems are generally recognized to be MIMD architectures; either exploiting a single shared memory space or a distributed memory space.

### (v) von Neumann vs. Harvard Architecture

While talking about computers, there should be a topic on John von Neumann. This Hungary-born American mathematician gave the concept of stored-program architecture first but unfortunately in his life span no fully functional computer was made. His ideas will be discussed later on this chapter.

| von Neumann architecture | Harvard architecture |
|---|---|
| 1. Same memory for data & program | 1. Separate data and code memory |
| 2. Instructions are executed one by one. Execution requires at least two clock cycles. One for fetch and others for execution and pipelining is not possible. | 2. Pipelining is possible. |

### (vi) Big-endian vs Little-endian

This is based on "How a large binary number is stored in 8 bit (byte) wide memory" (remember – each memory address can store 1 byte of data). Suppose the following 32 bit number have to be stored:

11111111 00000000 10101010 11110000

In case of **big-endian**, the most significant byte is stored in the smallest address. Here's how it would look:

| Address | Data value |
|---|---|
| 400 | 11111111 |
| 401 | 00000000 |
| 402 | 10101010 |
| 403 | 11110000 |

In case of **little-endian**, the least significant byte is stored in the smallest address. Here's how it would look:

| Address | Data value |
|---|---|
| 400 | 11110000 |
| 401 | 10101010 |
| 402 | 00000000 |
| 403 | 11111111 |

It can be easily understood by a little thought that "performing arithmetic operations are easier in little-endian computers where as comparing strings chronologically is easier in big-endian computers".

## 1.3 BASIC ANATOMY OF A COMPUTER

A digital computer is an electronic device that receives data, performs arithmetic and logical operations and produces results according to a predetermined program. It receives data from input devices and gives results to output devices. The central processing unit, also known as processor processes the data. Memory (primary and secondary) is used to store data and instructions. Follows, block diagram of a digital computer identifying the key components and their interconnection.

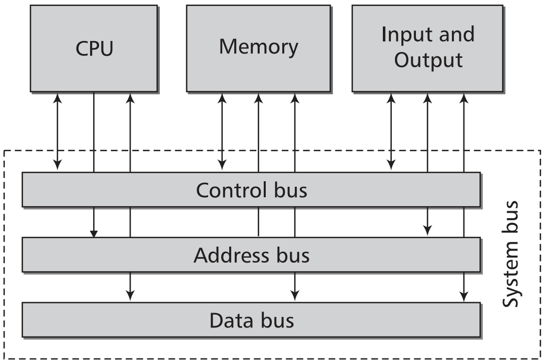


**Fig. 1.3 Components of a computer**

The Central Processing Unit (CPU) is like the brain of the computer. It is responsible for executing instructions. It controls the sequence of execution of instructions. It comprises a Control Unit (CU), an Arithmetic & Logic Unit (ALU) and huge number of registers. The CU controls the execution of instructions. First it decodes the instruction and then generates micro-operations in a particular order with the help of control memory. The ALU is responsible for performing arithmetic and logic operations.

The interconnections are referred as BUS. Buses are nothing but bunch of wires used to carry digital signals. There are three kinds of bus:

1. Address bus
2. Data bus
3. Control bus

Address bus carries address of memory from where to read/to where to write data. Size of address bus of a processor defines the amount of memory addressable by it. For example a processor with 16 bit address bus can access $2^{16}$ = 64 KB memory ($2^6$ = 64 and $2^{10}$ = 1024 = 1 K) and a processor with 32 bit address bus can access $2^{32}$=4 GB memory ($2^2$ = 4 and $2^{30}$ = 1024 × 1024 × 1024 = 1024 × 1024 × 1 K = 1024 × 1 Mega = 1 Giga.). Address bus is unidirectional, i.e., it carries signal from CPU to other components (only CPU is intelligent enough to generate address).

Data bus is bidirectional. It carries data read from/to be written to a device. Its size and width of registers signify the size of data that can be crunched by the processor in one go.

Control bus carries control signals that activate/deactivate various circuits.

Block diagram of the common bus architecture follows:

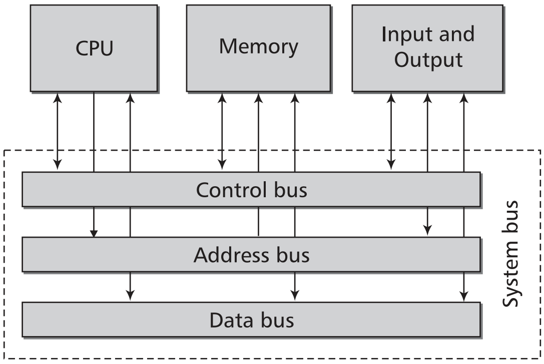


**Fig. 1.4 Common bus architecture**

## 1.4 VON NEUMANN ARCHITECTURE

While studying computer architecture, the name of Von Neumann comes first. Though he could not see any working model based on his proposal during his life span, design of all modern computers is based on the stored program model proposed by him. In 1945, the mathematician and physicist John von Neumann, along with others, had described computer architecture in the First Draft Report on the EDVAC. The Von Neumann architecture, which is also known as the Von Neumann model and Princeton architecture, is based on that. In this report a digital computer has been proposed that will contain the following parts:

- A processing unit containing an arithmetic logic unit and processor registers.
- A control unit containing an instruction register and program counter.
- Memory to store both data and instructions.
- External mass storage.
- Input and output mechanisms.

Stored-program computers were much advanced as compared to the program-controlled computers of the 1940s, like the Colossus and the ENIAC. These were programmed by switches and patches, which led to route data and could control the signals between the various functional units.

Some key points to remember about this architecture:

- A program should totally reside in the main memory prior to execution.*
- Data and code will reside in the same memory and will be indistinguishable.†

*As the programs became larger and larger, it was not possible to put the total program in the relatively smaller main memory. So **virtual memory** was introduced in operating systems. It is a technique that shows the free part of secondary storage as main memory. It keeps the total program in the secondary storage in blocks and fetches the required blocks to primary when necessary.

†Here **Von Neumann** design differs from **HAVARD ARCHITECTURE** that uses separate code and data memory on separate bus.

- Instructions will be fetched from memory and executed one at a time in a linear fashion.‡ $$

## 1.5 MEMORY CLASSIFICATION AND HIERARCHY

Broadly classified, a computer system has two types of memory – Primary and Secondary.

RAM (Random Access Memory), ROM (Read Only Memory) and Cache memory (one kind of very fast random access memory) falls in the primary category where as Magnetic tape, CD-ROM, DVD, Hard disk, pen drive (flash memory) falls in the secondary mass storage category.

Computer system uses memory hierarchy to optimize hardware cost. For example, in a system where 4 megabyte of cache memory is being used, the size of RAM can be 2 GB or 4 GB. If a system is build with total 2 GB cache instead, the system cost will go beyond imagination but 'speed boost' will be only 10% of the system mentioned earlier.

If memory is classified in terms of access strategy then there are three categories:

1. Sequential access.
2. Random access.
3. Direct access.

In **sequential access memor**y, access time is directly proportionate to address. Magnetic tapes falls in this category.

In **random access memory**, access time is constant, i.e., to access content at 1st location or at millionth location, same time will be required.

All sorts of circular and rotating memories like hard disk, CD, DVD fall in **direct access** category where access time can be expressed as a function T = Ax + By where A, B are constants and x, y are variables. These sorts of memory are divided in tracks and sectors. Time required by the read/write head to reach a track is known as **seek time**. Time required by a sector to reach under read/write head is known as **latency time.**

## 1.6 INPUT AND OUTPUT DEVICES

In this section various input and output devices commonly used in computer systems will be discussed.

### 1.6.1 Input Devices

#### *1.6.1.1 Keyboard*

Text information is entered in the computer by typing on the computer keyboard. Most keyboards, for example – the 101-key US traditional keyboard or the 104-key Windows keyboard, have alphabetic and numeric characters, punctuation marks and function keys. Keyboards are generally connected to the computer by a PS/2 connector or USB port.

‡ To speed up execution, multiple instructions are fetched in a queue in a pipelined computer.