**CSC2231 COMPUTER ARCHITECTURE**

* Basic concepts
  + Brief Historical Overview
  + Stored program (von Neumann) computer organization
  + Instruction Sets
  + Processor Cycle
  + Measuring, predicting, and improving performance
* Inside the CPU
  + Basic Device Physics
  + Boolean Algebra and Digital Logic
  + Implementing CPU elements with Digital Logic
  + Pipelining
  + Controlling the CPU
* Memory Architecture
  + Memory Design and Construction
  + The Memory Hierarchy
  + The Memory Map
  + Caches
* I/O and Peripheral Control
  + Buses
  + Memory-mapped I/O
  + Interrupts
  + Direct Memory Access
* Hardware/Software Interface
  + What happens during compilation and linking?
  + Machine Code and Assembly Language
  + Generating Code
  + Addressing Modes

|  |  |  |
| --- | --- | --- |
| **Computer Organization, Design, and**  **Architecture** | Sajjan G. Shiva | CRC Press, 2014 |
| **Computer organization and Design - The**  **Hardware/Software Interface** | John Hennessy and David  Patterson | Morgan Kaufmann, 2009 |
| **Structured Computer Organization** | Andrew Tanenbaum and  Tood Austin | Pearson Education, 2013 |
|  |  |  |
| **Schaum’s Outline: Computer Architecture** | Nicholas Carter | McGraw Hill, 2002 |
| **Computer Architecture: A Quantitative**  **Approach** | John Hennessy and David  Patterson | Morgan Kaufmann, 2011 |
| **Computer Organization and Architecture** | William Stallings | Pearson Education, 2013 |

Can be defined at different levels:

**Level 7**

* Can be defined at different levels:
* Translation (compiler)

**Level 6**

* Assembly Language level
* Translation (assembler)

**Level 5**

* Operating system machine level
* Partial interpretation (OS)

**Level 4**

* Instruction Set Architecture level
* Interpretation or Direct execution (Micro program)

**Level 3**

* Micro architecture level
* Hardware

**Level 2**

* Digital logic level
* Hardware

**Level 1**

* Device Level

**Levels 1, 3 and 4 will be our main preoccupation**

Why do we bother with all these?

* In order to do:
* Tell us how computer are built
* How the computer operates
* Knowing what's inside and how it works
* help you design, develop, and implement applications better, faster, cheaper, more efficient, and easier to use because you will be able to make informed decisions
* Be a good Computer Scientist
* Understands how computer processes information
* In describing computers, a distinction is often made between computer architecture and computer organization.
* Computer architecture refers to those attributes of a system visible to a programmer, those attributes that have a direct impact on the logical execution of a program. A term that is often used interchangeably with computer architecture is instruction set architecture (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.
* Computer organization refers to the operational units and their interconnections that realize the architectural specifications.

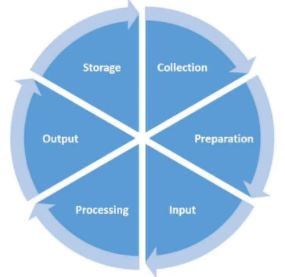
**STRUCTURE AND FUNCTION**

* A computer is a complex system; contemporary computers contain millions of elementary electronic components. How, then, can one clearly describe them?
* **Structure:** The way in which the components are interrelated.
* **Function:** The operation of each individual component as part of the structure.

**Function**

There are four (4) basic Function that a computer can perform

* **Data processing**: Data may take a wide variety of forms, and the range of processing requirements is broad. However, we shall see that there are only a few fundamental methods or types of data processing.
* **Data Processing Cycle:** The data processing cycle consists of a series of steps where raw data (input) is fed into a process (CPU) to produce actionable insights (output). Each step is taken in a specific order.



**Data Processing Cycle**

1. **Collection:** The collection of raw data is the first step of the data processing cycle. The type of raw data collected has a huge impact on the output produced.
2. **Preparation:** Data preparation or data cleaning is the process of sorting and filtering the raw data to remove unnecessary and inaccurate data. Raw data is checked for errors, duplication, miscalculations or missing data, and transformed into a suitable form for further analysis and processing
3. **Input:** In this step, the raw data is converted into machine readable form and fed into the processing unit. This can be in the form of data entry through a keyboard, scanner or any other input source.
4. **Data Processing:** In this step, the raw data is subjected to various data processing methods using machine learning and artificial intelligence algorithms to generate a desirable output.
5. **Output:** The data is finally transmitted and displayed to the user in a readable form like graphs, tables, vector files, audio, video, documents, etc. This output can be stored and further processed in the next data processing cycle.
6. **Storage:** The last step of the data processing cycle is storage, where data and metadata is stored for further use. This allows for quick access and retrieval of information whenever needed, and also allows it to be used as input in the next data processing cycle directly.

**Types of Data Processing**

There are different types of data processing based on the source of data and the steps taken by the processing unit to generate an output.

1. **Batch Processing:** Data is collected and processed in batches. Used for large amounts of data. eg: payroll system
2. **Real-time Processing:** Data is processed within seconds when the input is given. Used for small amounts of data. eg: withdrawing money from ATM
3. **Online Processing:** Data is automatically fed into the CPU as soon as it becomes available. Used for continuous processing of data. eg: barcode scanning
4. **Multiprocessing:** Data is broken down into frames and processed using two or more CPUs within a single computer system. Also known as parallel processing. eg: weather forecasting
5. **Time-sharing:** Allocates computer resources and data in time slots to several users simultaneously.

**Data Processing Methods**

1. **Manual Data Processing:** In this data processing method, data is processed manually. The entire process of data collection, filtering, sorting, calculation and other logical operations are all done with human intervention without the use of any other electronic device or automation software. It is a low-cost method and requires little to no tools, but produces high errors, high labor costs and lots of time.
2. **Mechanical Data Processing:** Data is processed mechanically through the use of devices and machines. These can include simple devices such as calculators, typewriters, printing press, etc. Simple data processing operations can be achieved with this method. It has much lesser errors than manual data processing, but the increase of data has made this method more complex and difficult.
3. **Electronic Data Processing:** Data is processed with modern technologies using data processing software and programs. A set of instructions is given to the software to process the data and yield output. This method is the most expensive but provides the fastest processing speeds with the highest reliability and accuracy of output.

**Function**

1. **Data storage:** Even if the computer is processing data on the fly (i.e., data come in and get processed, and the results go out immediately), the computer must temporarily store at least those pieces of data that are being worked on at any given moment. Thus, there is at least a short-term data storage function. Equally important, the computer performs a long-term data storage function. Files of data are stored on the computer for subsequent retrieval and update.
2. **Data movement**: The computer’s operating environment consists of devices that serve as either sources or destinations of data. When data are received from or delivered to a device that is directly connected to the computer, the process is known as input–output (I/O), and the device is referred to as a peripheral. When data are moved over longer distances, to or from a remote device, the process is known as data communications.
3. **Control:** Within the computer, a control unit manages the computer’s resources and orchestrates the performance of its functional parts in response to instructions.

**Structure**

1. We now look in a general way at the internal structure of a computer.
2. There are four main structural components:
3. **Central processing unit (CPU):** Controls the operation of the computer and performs its data processing functions; often simply referred to as processor.
4. Main memory: Stores data.
5. **I/O:** Moves data between the computer and its external environment.
6. System interconnection: Some mechanism that provides for
7. Communication among CPU, main memory, and I/O. A common example of system interconnection is by means of a system bus, consisting of a number of conducting wires to which all the other components attach.

**Central processing unit (CPU):**

Its major structural components are as follows:

1. **Control unit:** Controls the operation of the CPU and hence the computer.
2. **Arithmetic and logic unit (ALU):** Performs the computer’s data processing functions.
3. **Registers:** Provides storage internal to the CPU.
4. **CPU interconnection:** Some mechanism that provides for communication among the control unit, ALU, and registers.

* First calculating ‘machines’ were fingers, stones, beads, etc
* Examples of counting aids from ancient civilizations includes
  + **Greece:** The Abacus (287BC)
  + **Africa:** decorated necklace and wrist bands
  + **Peru:** rows of knotted strings, etc
* 1st mechanical calculating machine was the Antikythera used by the ancient Greeks for astronomical calculations
* **Mechanical computation did not get going until the 17th century**
* Blaise Pascal, aged 19, invented a gear-based machine for addition and subtraction in 1642
* In 1672, von Leibnitz added multiplication and division
* Then came the Charles Babbage’s Difference Engine
  + Only addition and subtraction but with hard-copy output
  + Laid the foundation for 1st programmable mechanical computer
* Charles Babbage’s Analytical Engine
  + The Store (Memory)
  + The Mill (CPU)
  + The Input section (punch card reader)
  + The Output (Punched cards)
* Read instructions from punch cards and execute them
* 1st fully programmable yet fully mechanical
* Programming was via a simple assembly language developed by Ada Lovelace
* Never worked – the gears couldn’t be made accurate enough

**First Electronic Computers**

* Babbage’s Analytical Engine laid the foundation
* No significant advances until late 1930
  + Conrad Zuse used electronic relays to build an electronic computing machine
  + John Atanasoff invented a machine which used binary arithmetic and had a capacitor memory (now known as DRAM). Never worked
  + George Stibbitz demonstrated a working machine in 1940
  + Howard Aiken built Babbage’s analytical engine using electronic relays in 1944, and made it work.

**The Vacuum Tube Machines**

* Compelled by the necessity to crack the ENIGMA code, Alan Turing was commissioned by the British MoD to design the COLOSSUS
  + The world’s 1st electronic digital computer
  + Worked, but highly classified
  + Maulchy and Erkert build the ENIAC in 1943-46
  + 18000 vacuum tubes, 1500 relays, 6000 programming switches, 30 tons
  + Influenced a host of similar machines

– EDSAC I, ILIAC, MANIAC, etc.

**The von Newmann Machines**

* John von Newmann designed the 1st programmable electronic computer at IAS
  + The basic architecture for almost all modern machine follows the IAS design
* **The Invention of Transistors**
  + Shockley, Bardeen and Brittain invented the transistor in 1948, and won the 1956 Nobel Prize for Physics
  + Made vacuum tube machines obsolete from early 50s
  + Was very expensive: the DEC PDP-1 had 4K 18-bit word memory, a visual display, ran at 0.2MHz and cost $120,000.00
  + The PDP-8 later introduced the idea of a bus, to allow components to communicate with each other, rather than via the memory
* The IBM 7090 dominated scientific computing in the 60s, while 1401 dominated the business market

**Seymour Cray**

* The 6600, designed by Seymour Cray was introduced by CDC in 1964
* It was superscalar with multiple functional units that worked in parallel
* Contained smaller slave units that handled admin tasks, leaving the CPU to crunch numbers
* Decades ahead of anything else that existed
* Cray later formed his own company manufacturing super computers

**The IC Revolution**

* Invention of IC by Jack Kilby and others drove the next revolution
* They figured out how to pack many transistors on a single microchip
* Led to cheaper computers and ignited the microcomputer revolution in the 70s
* Notable machines included the IBM 360, suitable for both scientific and commercial applications

**VLSI**

* IC-based micros where still very large
* VLSI technology allow tens hundreds of millions of transistors to be fabricated on a single chip
* Ignited the home computer revolution
* Early home computers cost a few hundred pounds and were plugged into TV
* They were mainly used for games, word processing and spreadsheet
* Commodore 64, Intel 8080, Zilog Z80, BBc Micro, etc

**The PC**

* Intel 8088 CPU, MS-DOS operating system
  + Windows came much later (after Apple had introduced the GUI on the Lisa, and then the Mac)
* IBM made the design public so that others could build add-ons
* But others sold IBM-compatible machines at much lower cost

Microsoft and Intel eventually gained control of the market created by IBM, and the result is the modern PC industry

**The Modern VLSI Microprocessors**

* The discrete-transistor machines of the 60s had a few thousand transistors which filled a large room
* Early VLSI chips had a few tens of thousands transistors
* The modern Athlon 64 X2 has 221 million transistors on one microchip that is just 118mm2  in area
* The transistors are < 0.1microns in width

**Instruction vs. Data**

* Modern computers accomplish computations by interpreting and manipulating binary strings
* Binary strings are either 4 or 8 bytes long in most machines (i.e. 32 or 64 bits)
* A single 32- (or 64-) bits string is referred to as a **word,** which are of two types
  + **Data**: Numbers/ASCII letter codes that correspond to the variables/constants in our programs
  + **Instructions**: interpreted by the computer and used to control what happens to the data
* Both are stored in a single read-write memory, although in diff locations
* Both words look the same but are processed differently
* Both can be manipulated using digital logic

A **computer** needs to be able to:

* Manipulate data
* Interpret instructions and decide what to do
* Store words (both types)
* Keep track of what is happening next
* Get data/instructions in and out
* Move data/instructions around

The core of modern computer where software is executed is the **CPU**, which must be able to

1. Read the program and its data from some external device (memory)
2. Process instructions of the program in the correct order
3. Manipulate and compute with data in accordance with the instructions
4. Take decisions according to the results of the computations
5. Send the results of the computations to some external device (screen, printer) or store them for further use

Typical **CPU** must therefore have units which can:

1. Load/store instruction and data in memory
2. Interpret the instructions to allow the machine to do the necessary computations
3. Send the results to external devices
4. Take appropriate decisions and control what is happening
5. Keep track of things

Instr. Register

Program Counter

Load/Store Unit

Control Unit

Registers

ALU

**Storage**

Memory can, roughly, be divided into

**Main memory**

* Used to store both instructions and data
* Can hold whole programs and all their data
* Generally physically separate from the main CPU and therefore slow compared to the Registers

**Registers**

* Used to store active instructions and data which are currently being used
* Each register holds one word
* Physically within the CPU, therefore very fast compared to MM
* Some registers are special-purpose

» Instruction registers

» Program counters

**I/O**

* Computers are useless if there is no means of putting stuff in and getting stuff out
* Most devices have some I/O controller which handles external devices
* These controllers work in several ways
  + Assign the external devices some memory addresses
  + Use interrupts

**Bus**

* The transport network within a computer is known as the bus Consists of:
  + A set of wires down which information can be sent (instruction, data, addresses, control signal, etc)
  + A protocol which allows components to
    - Write info on to the bus for transmission to other components of the machine
    - Read info from the bus
    - Control where the info is being sent and what is reading it
* A bus controller which
  + Grants/refuses access to the bus
  + Balances the use of the bus between components

**The von Neumann Design**

MAIN MEMORY

CPU

LONG-TERM MEMORY

I/O & PERIPHERALS

THE BUS

S