Chapter 4. Objectives

- Learn the components common to every modern computer system.
- Be able to explain how each component contributes to program execution.
- Understand a simple architecture invented to illuminate these basic concepts, and how it relates to some real architectures.
- Know how the program assembly process works.

CPU

- Registers hold data that can be readily accessed by the CPU.
- They can be implemented using D flip-flops.
- A 32-bit register requires 32 D flip-flops.
- The arithmetic-logic unit (ALU) carries out logical and arithmetic operations as directed by the control unit.
- The control unit determines which actions to carry out according to the values in a program counter register and a status register.

BUS

- The CPU shares data with other system components by way of a data bus.
- A bus is a set of wires that simultaneously convey a single bit along each line.
- Two types of buses are commonly found in computer systems: point-to-point, and multipoint buses.

This is a point-to-point bus configuration:
Multipoint Bus

- Because a multipoint bus is a shared resource, access to it is controlled through protocols, which are built into the hardware.

BUS

- Buses consist of data lines, control lines, and address lines.
- While the data lines convey bits from one device to another, control lines determine the direction of data flow, and when each device can access the bus.
- Address lines determine the location of the source or destination of the data.

BUS Arbitration

- In a master-slave configuration, where more than one device can be the bus master, concurrent bus master requests must be arbitrated.
- Four categories of bus arbitration are:
  - Daisy chain: Permissions are passed from the highest-priority device to the lowest.
  - Centralized parallel: Each device is directly connected to an arbitration circuit.
  - Distributed using self-detection: Devices decide which gets the bus among themselves.
  - Distributed using collision-detection: Any device can try to use the bus. If its data collides with the data of another device, it tries again.

Traditional (ISA) (with cache)

High Performance Bus

Inside of Modern Computer
Clock

- Every computer contains at least one clock that synchronizes the activities of its components.
- A fixed number of clock cycles are required to carry out each data movement or computational operation.
- The clock frequency, measured in megahertz or gigahertz, determines the speed with which all operations are carried out.
- Clock cycle time is the reciprocal of clock frequency.
  - An 800 MHz clock has a cycle time of 1.25 ns.

Clock vs. Performance

- Clock speed should not be confused with CPU performance.
- The CPU time required to run a program is given by the general performance equation:
  \[
  \text{CPU Time} = \frac{\text{seconds}}{\text{cycles per instruction}} \times \frac{\text{cycles per second}}{} \times \frac{\text{instructions}}{\text{program}} \times \frac{\text{program}}{\text{seconds}}
  \]

  - We see that we can improve CPU throughput when we reduce the number of instructions in a program, reduce the number of cycles per instruction, or reduce the number of nanoseconds per clock cycle.

Input/Output Systems

- A computer communicates with the outside world through its input/output (I/O) subsystem.
- I/O devices connect to the CPU through various interfaces.
- I/O can be memory-mapped – where the I/O device behaves like main memory from the CPU's point of view.
- Or I/O can be instruction-based, where the CPU has a specialized I/O instruction set.

A Modern Memory Hierarchy

- By taking advantage of the principle of locality:
  - Present the user with as much memory as is available in the cheapest technology.
  - Provide access at the speed offered by the fastest technology.

Main Memory Organization

- Computer memory consists of a linear array of addressable storage cells that are similar to registers.
- Memory can be byte-addressable, or word-addressable, where a word typically consists of two or more bytes.
- Memory is constructed of RAM chips, often referred to in terms of width \( \times \) depth.
- If the memory word size of the machine is 16 bits, then a 4M \( \times \) 16 RAM chip gives us 4 megabytes of 16-bit memory locations.

Memory Organization

- How does the computer access a memory location corresponding to a particular address?
- We observe that 4M can be expressed as \( 2^2 \times 2^{10} = 2^{22} \) words.
- The memory locations for this memory are numbered 0 through 2\(^{22} - 1\).
- Thus, the memory bus of this system requires at least 22 address lines.
  - The address lines “count” from 0 to 2\(^{22} - 1\) in binary. Each line is either “on” or “off” indicating the location of the desired memory element.
Memory Organization

- Physical memory usually consists of more than one RAM chip.
- Access is more efficient when memory is organized into banks of chips with the addresses interleaved across the chips.
- With low-order interleaving, the low order bits of the address specify which memory bank contains the address of interest.
- Accordingly, in high-order interleaving, the high order address bits specify the memory bank.

Dynamic RAM

- Capacitor can hold charge
- Transistor acts as gate
- No charge is a 0
- Can close switch & add charge to store a 1
- Then open switch (disconnect)
- Can read by closing switch
- Destructive Read
  - When cell read, charge removed
  - Must be restored after a read
- Refresh
  - Also, there’s steady leakage
  - Charge must be restored periodically

Interrupt

- The normal execution of a program is altered when an event of higher-priority occurs. The CPU is alerted to such an event through an interrupt.
- Interrupts can be triggered by I/O requests, arithmetic errors (such as division by zero), or when an invalid instruction is encountered.
- Each interrupt is associated with a procedure that directs the actions of the CPU when an interrupt occurs.
  - Nonmaskable interrupts are high-priority interrupts that cannot be ignored.