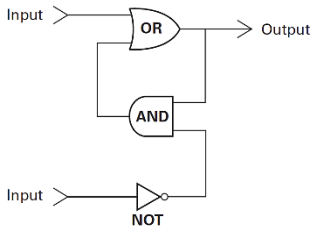
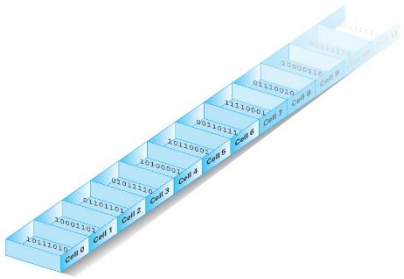
Notes on computer architecture

Logan Thrasher Collins

Main memory



- Some computers store data using flip-flop circuits. Each flip-flop circuit possesses a configuration of logic gates (including AND, OR, and NOT gates) that allows switching between “on” and “off” states corresponding to 1 and 0.
- More modern machines often use conceptually similar ways of storing data that involve using tiny electric charges to represent 1 and 0 states.
- Each memory cell contains eight flip-flop circuits (or similar storage devices) that correspond to eight bits of memory. Together, eight bits are equal to one byte.
- The memory cell’s eight bits are depicted as arranged in a line. The leftmost end is called the high-order end and the rightmost end is called the low-order end. The leftmost bit is called the most significant bit and the rightmost bit is called the least significant bit.
- In order for the computer to find specific memory cells within main memory, every cell is assigned a unique numeric address. This can be visualized as a series of memory cells lined up and numbered starting with zero. In this way, individual cells are not only identifiable, but they are also ordered relative to other cells.

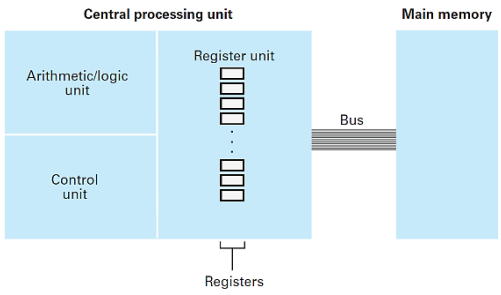


- Since the computer can independently access any cell that is needed for a computation (despite the cells possessing an ordered configuration), the main memory is called random access memory (RAM).
- For computers that use tiny charges (rather than flip-flop circuits) to store data, the main memory is called dynamic RAM or DRAM because the charges are volatile, dissipate quickly, and must be restored many times per second using a refresh circuit.

Central processing unit

- The central processing unit (CPU) includes an arithmetic unit that performs operations on data, a control unit that coordinates the machine’s activities, and a register unit that temporarily stores results from the arithmetic unit (and other data) in registers.

- The CPU is connected to the main memory (which is more permanent than the registers) via a collection of wires called a bus. To perform an operation on data from the main memory, the CPU uses an electronic address to find the desired data cell and send it to a set of registers. To write data into the proper location within main memory, the CPU uses a similar address system.



The stored program

- Instructions for the CPU's data manipulation can be stored in a computer's main memory because programs and data are not fundamentally distinct entities.
- The following steps summarize how stored programs operate.
    1. Retrieve a set of values from main memory and place each value within a register.
    2. Activate the circuitry that performs some operation upon the values (i.e. two values might be added together) and then store the result in another register.
    3. Transfer the result from its register to main memory for long-term storage. After this, stop the program.
- CPUs also store cache memory in order to increase their speed. The cache memory is a temporary copy of the portion of the main memory that is undergoing processing at a given time. Using cache memory, the CPU can rapidly retrieve relevant data without needing to go all the way to the main memory as often.

Machine language

- Data transfer group: instructions to "transfer" data from a memory cell to a register (or some similar process) are more accurately described as "copying" the data. Requests to copy data from are memory cell to a register are called LOAD instructions. Requests to copy data from a register and write it to a memory cell are called STORE instructions. Requests that control interaction of the CPU and main memory with external devices like printers and keyboards are referred to as I/O instructions.
- Arithmetic/logic group: the arithmetic/logic unit can carry out instructions that run data through basic arithmetic operations and Boolean logic gate operations (AND, NOT, OR, XOR, etc.) The arithmetic/logic unit also uses the SHIFT and ROTATE instructions. SHIFT moves bits to the left or right within a register. ROTATE is another version of SHIFT which moves bits to the slots at the other end of the register (rather than allowing them to "fall off" as would happen if SHIFT were used).
- Control group: contains instructions that direct program execution and termination. JUMP (also called BRANCH) commands cause a program to change the next action that it performs. JUMP commands can be unconditional or conditional (when conditional, they work like "if" statements). The STOP command also falls into this category.

Machine cycle

- The machine cycle involves two special purpose registers, the instruction register and the program counter.
- The instruction register contains the instruction that is undergoing execution.
- The program counter contains the address of the next instruction that will be executed and so keeps track of the machine's place within the program.
- Using three steps, the CPU performs the machine cycle.
    1. Fetch: the CPU retrieves an instruction from the main memory at the address specified by its program counter. The program counter then increments to specify the next instruction.
    2. Decode: the CPU breaks the instruction into appropriate components based on its operational code.
    3. Execute: the CPU activates the necessary circuitry to perform the command that was requested.
- The computer's clock is a circuit that generates oscillating pulses which control the machine cycle's rate. A faster clock speed results in a faster machine cycle. Clock speed is measured in Hertz. Typical laptop computers (as of 2018) run at clock speeds of several GHz.
- To increase a computer's performance, pipelining is often used. Pipelining involves allowing the steps of the machine cycle to overlap. Using pipelining, an instruction can be fetched while the previous operation is still underway, multiple instructions can be fetched simultaneously, and multiple operations can be executed simultaneously so long as they are independent of each other.

Multiprocessor machines

- Some computers possess multiple CPUs that are linked to the same main memory. This is called a multiple-instruction stream multiple-data stream (MIMD) architecture. The CPUs operate independently while coordinating their efforts by writing instructions to each other on their shared memory cells. In this way, a CPU can request another CPU to perform a specified part of a large processing task.
- Some computers use multiple CPUs that are linked together so as to perform the same sequence of instructions simultaneously upon distinct datasets. This is called a single-instruction stream multiple-data stream (SIMD) architecture. SIMD machines are useful when the application requires the same task to be performed upon a large amount of data.
- Parallel processing can also be carried out using large computers that are composed of multiple smaller computers, each with its own CPU and main memory. In these cases, the smaller computers coordinate the partitioning of resources to handle a given task.

Reference and image source: Brookshear, J. G., Smith, D. T., & Brylow, D. (2012). *Computer Science: An Overview*. Addison-Wesley.