Von Neumann Computers

CS 350: Computer Organization & Assembler Language Programming

A. Why?

- The computers we use follow the von Neumann model/architecture.

B. Outcomes

After this lecture, you should know

- The von Neumann architecture and how it differs from other architectures.
- The basic design of a von Neumann computer and CPU.
- The different parts of the instruction cycle and what happens during them.

C. Mechanical Computation Devices

- **Jacquard's Loom** (early 1800s): First machine to perfect use of punch cards to direct its activity (controlling pattern of weave). Not a computer but introduced notion of changeable control of a device.
- **Difference Engine (Charles Babbage, 1820s–1840s)**: Designed to do calculations of values of polynomials. Was technically possible but not actually completed. (Manufacture required stronger tolerances than was easily available, British government cut off funding, and he had become interested in a more general machine.
- **Analytical Engine (Babbage; 1830s on)**: Designed to be a general purpose computer. Instructions and data would've been entered using punch cards. Instructions included tests and loops; machine would've been equivalent to Turing machines.
- Programming language Ada is named after Ada Byron (Countess of Lovelace), who wrote programs for the Analytical Engine.

D. Electro-Mechanical Computation Devices

- **Konrad Zuse (1930s-1940s)**: Built Z machines that did computations using mechanical and electric relays. The Z3 was equivalent to Turing machines and could be programmed using punch tape (not rewiring).
- **Harvard Mark I (1940s)**: also used switches, relays, other mechanical devices. Could execute long instructions automatically. Controlled by punch tape (a loop could be written by connecting the two ends of the tape.

E. Early Electronic Computation Devices

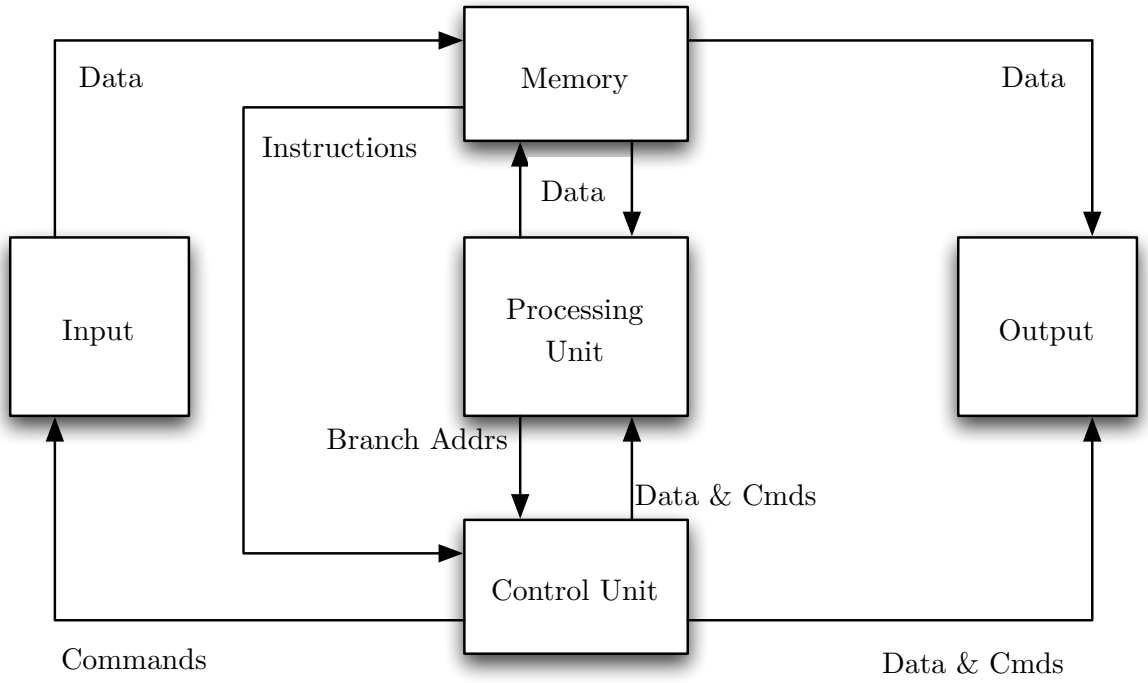
Fixed-Program Computer

- In a fixed-program computer, the program is not changeable at runtime.
- 1939: ABC (Atanasoff-Berry Computer) [Wikipedia photo]
  - Electronic machine. No CPU, but did mathematical computations & binary logic. Not programmable (built for one kind of problem — solving linear equations).
  - Adapted from Wikipedia: The ABC weighed > 700 lb. and was 800 feet^2. It contained ~1 mile of wire, 280 dual-triode vacuum tubes, 31 [vacuum tube diodes], and was about the size of a desk.
- 1943: ENIAC (Electronic Numerical Integrator And Computer) [Wikipedia photo]
  - Presper Eckert and John Mauchly -- first general purpose electronic computer. Decimal computer, not binary. Program hard-wired through dials & switches.
  - Adapted from Wikipedia: ENIAC contained 17,468 vacuum tubes, 7,200 crystal diodes, 1,500 relays, 70,000 resistors, 10,000 capacitors and around 5 million hand-soldered joints. It weighed 27 tons, was roughly 8.5'×3'×80', took up 1800 feet^2, and consumed 150 kW of power. On average a tube failed every 2 days and took 15 min to find.

Stored-Program Computer (von Neumann Computer)

- A stored-program computer has modifiable programs, stored in read-write memory along with data.

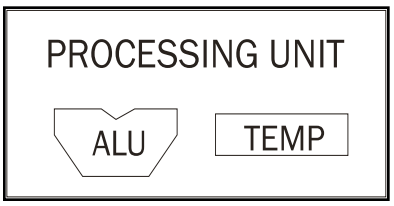
- 1944, EDVAC (Electronic Discrete Variable Automatic Computer) [Wikipedia photo], John von Neumann co-authors report on stored program concept.
- Adapted from Wikipedia: The EDVAC contained ~6,000 vacuum tubes, 12,000 diodes, 1,500 relays. It weighed 8.5 tons, took 490 feet^2, and consumed 56 kW of power. The EDVAC had ~ 5.5 kB of *ultrasonic serial memory*: Electrical pulses were transformed by a crystal speaker into pressure pulses that traveled through a column of mercury. Pulses received at the other end were converted back to electricity using crystal microphones, amplified, and sent back to repeat.



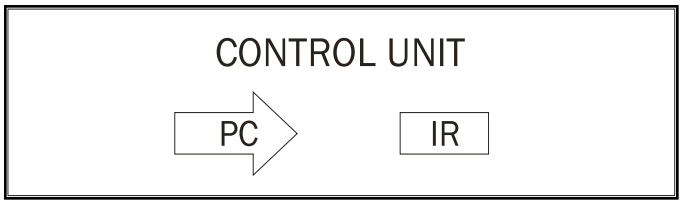
Typical von Neumann Computer

F. Basic von Neumann Architecture

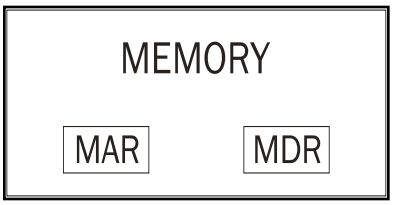
- Under the von Neumann architecture, a computer has 3 main parts: Read-Write **Memory**, containing instructions and data; Input and Output devices; and a **CPU (Central Processing Unit)**, which includes a **Control unit**, which interprets instructions and a **Processing unit**, which does arithmetic/logical operations.



- The traditional processing unit is an **ALU (Arithmetic Logical Unit)**. It contains
  - "General-purpose" **Registers**: Small fast temporary storage.
    - Limited number (expensive)
    - The machine's **Word Size** is the width of the registers.
  - Circuitry to perform calculations on data
  - ALU action is directed by the **Control Unit**
    - ALU can get data from control unit, may send addresses to the control unit for branches or jumps (i.e., goto's)
    - ALU also sends data to memory/receives data from memory.



- The **Control Unit (CU)** directs execution of the program. The **Instruction Register** (`IR`) contains the current instruction. The **Program Counter** (`PC`) contains the memory address of the next instruction to be executed. (**"Instruction Pointer"** would probably be a better name, but `PC` is traditional.)



- **Memory:** Each address is K bits long and stores M bits. In theory K and M are independent; in practice $K \leq M$.
  - **MAR: Memory Address Register** (K bits wide): Contains address to read/write
  - **MDR: Memory Data Register** (M bits wide): Contains value read/ written
- **Input & Output Devices**
  - Keyboard, mouse, disk, video, printer, etc. Receive commands from processing unit (or control unit; depends on design). Send/Receive data. Sends **interrupt signals** to control unit when done with operation.

- Actual modern computers are more sophisticated than the original von Neumann design
  - Many processing units (e.g. floating-point)
  - I/O processing off-loaded to special hardware (e.g. video cards)
  - I/O goes directly into/out of memory (avoids CPU bottleneck)
  - Control unit works on > 1 instruction at a time (execute current instruction while decoding next instruction while loading one after that into memory).
  - Can have multiple control unit/processing unit pairs (multi-core CPUs)

G. Instruction Format, Instruction Set Architecture

- An instruction has an identifying **Opcode** (Operation code) and some **fields** (bit strings) that specify some numbers (possibly 0) of operands and results. It might also contain extra flags (like to specify when to a go to), or to differentiate between related operations (read a character? A string?)
- **Instruction length**
  - **Fixed-length** instruction machine: Easier to design, limits instruction designs.
  - **Variable-length** instruction machine: More flexible instruction sets, more compact use of memory, hardware more complicated.
- **Instruction Set Architecture (ISA)**
  - The set of instructions for a computer, their formats, the way they specify opcodes and operands and results.
  - It's the fundamental interface between programmers and hardware designers.

H. Addressing Modes

- A large part of how instructions work has to do with accessing operands: they might be in memory or a data register or they might be contained within the instruction itself. An **addressing mode** is a technique for specifying the location of an operand or result.

- First off, an operand might be part of the actual instruction; these are **immediate** operands. (E.g., the 1 implicit in `++x`.)
- If the data is in a **register**, then you need some bits to specify which one.
- The most obvious way for an instruction to indicate a memory address is to just have the address as a bitstring within the instruction. This is called **absolute addressing** and actually isn't as useful as you might think. First, you can only use it if (instruction size $\geq$ the opcode width + address width); second, you can't change what address you're accessing unless you change the instruction itself. (Makes going through an array trickier.)
- In practice, a data register can contain an address, so there are a number of techniques that use the value of a data register (a **base** register) or the program counter. In addition, we might have a constant "offset" — that way we can handle a bunch of addresses that are near each other without using multiple registers.
- In addition, there are addressing modes that use pointers for indirection (which we will see) and auxiliary index registers (which we won't).

I. The Instruction Cycle

- A computer executes a program by executing a sequence of instructions: Get instruction from memory, execute it; get instruction from memory, execute it, etc. Whether a bitstring is an instruction or data depends on how we use it.
- The **instruction cycle** is how one instruction is executed. Different people break up the cycle into different numbers of phases; the textbook uses 6:

  (1) Fetch instruction
  (2) Decode instruction
  (3) Evaluate address(es) of operands
  (4) Fetch operand(s)
  (5) Execute instruction
  (6) Store results of execution.
- Details vary by instruction; not all instructions have phases 3, 4, 6 (depends on the numbers of operands and results).
- Look at generic instruction cycle first; study a specific version in a bit.

- **Program Counter (`PC`)**: The `PC` is a special register that (most of the time) holds the address of the next instruction to execute. It only holds the address of the current instruction at the very beginning of the Fetch Instruction phase of the instruction cycle.

1. Fetch Instruction

- Read instruction into Instruction Register: Treat the Program Counter value as a memory address; get the contents of that address; copy those contents into the Instruction Register. Microcode (shorthand): `IR` ← `Mem`[`PC`]
  - Increment Program Counter (so that it points to next instruction in memory). Microcode: `PC` ← `PC+1`

2. Decode Instruction

- Feed opcode bits to decoder to figure out which instruction we have. (E.g. `IR`[12:15], for a 4-bit opcode stored at the left end of a 16-bit instruction.) Depending on the instruction, retrieve bits from other parts of the `IR`.

3. Evaluate Addresses (= Get effective addresses of operands)

- If the instruction has operands, then for each operand, figure out which register or memory location the operand is stored at. E.g., `IR`[4:7] could be a 3-bit register number, or `PC+IR`[7:15] could be an address. Not all instructions have this phase.

4. Fetch Operands

- If the instruction has operands then retrieve their values. They might be in registers, in memory, or encoded within the instruction itself. Not all instructions have this phase.

5. Execute Instruction

- There are 3 general kinds of instructions.
  - **Data movement instruction**: Load value from/Store value to memory.
  - **Calculation instruction**: Perform some operation on data (add, etc).
  - **Control Instruction** (Branch/Jump): Go to a different part of the program (used in decision and iteration statements).

6. Store Results

- Take result(s) of load-from-memory instructions and calculation instructions and put them somewhere. (Might be register(s), memory locations.) Instructions that don't produce results (like store to memory, or go to location) don't have this phase.
-

Von Neumann Computers

CS 350: Computer Organization & Assembler Language Programming

A. Why?

- The von Neumann architecture is the one used by modern computers

B. Outcomes

After this activity, you should be able to

- Describe the basic design of a von Neumman computer and discuss how it differs from other architectures.
- Describe the parts of the instruction cycle and what happens during them.

C. Questions

1. (a) What are the three main parts of a von Neumann computer?
   (b) What makes the von Neumann architecture different from earlier computer architectures. (Hint: "Stored program" architecture.)
   (c) What are the parts of the CPU?
2. (a) What is the Memory Address Register?
   (b) What is the Memory Data Register?
   (c) What are the steps involved in reading memory?
   (d) What are the steps involved in writing memory?
3. What does the Program Counter point to? Why is it badly named?
4. When we look at how instructions execute, we find three basic kinds of instructions. What are they and how do they differ?
5. (a) What are the phases of the instruction cycle?
   (b) What are the steps of the Fetch Instruction phase?
   (c) During the Decode Instruction phase of the instruction cycle, where is the instruction being decoded?

(d) What happens during the Evaluate Addresses (a.k.a Get Effective Addresses) phase of the instruction cycle? Why is this phase sometimes skipped?

(e) What happens during the Fetch Operands phase of the instruction cycle? From where can operands be fetched? Why is this phase sometimes skipped?

(f) What happens during the Store Results phase of the instruction cycle? To where can results be stored? Why is the phase sometimes skipped?

Solution

1. (a) CPU, Memory, and I/O devices

   (b) Programs are stored as data in memory.

   (c) Control unit and (Arithmetical/Logical) Processing Unit

2. (a) The Memory Address Register (MAR) holds the location to read or write.

   (b) The Memory Data Register (MDR) either holds the value read from memory or the value to write to memory.

   (c) To read: Set MAR to the address to read; signal Read; Find the value in the MDR and copy it out to wherever.

   (g) To write: Set MAR to the address to read; Set MDR ← value to write; Signal Write.

3. The PC points to the next instruction to execute. It's badly named in that it doesn't actually count anything.

4. Calculation instructions use and create values; Data movement instructions move data to/from memory; Control instructions perform goto operations.

5. (a) Fetch Instruction, Decode Instruction, Evaluate Addresses, Fetch Operands, Execute Instruction, Store Results.

   (b) Read the instruction pointed to the program counter from memory into the instruction register; then increment the program counter.

   (c) During Decode Instruction, the instruction is in the instruction register.

   (d) During Evaluate Addresses we figure out where the operands are stored; this could be a register number or a memory address. This phase is skipped if there are no operands.

   (e) During Fetch Operands we actually retrieve the operand values from a register, from memory, or from part of the instruction register. This phase is skipped if there are no operands.

   (f) During Store Results, values calculated during Execute Instruction are moved to memory or CPU registers. This phase is skipped if there are no instructions

