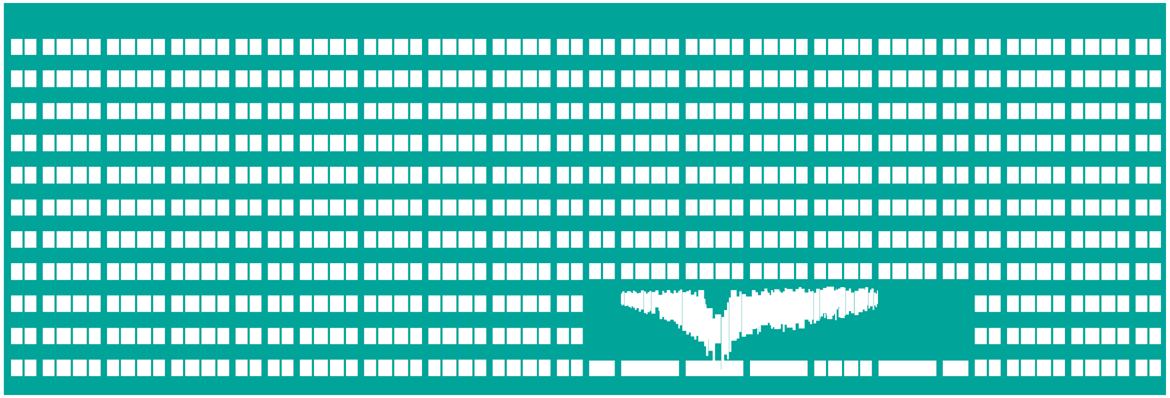
VŠB TECHNICKÁ UNIVERZITA OSTRAVA

VSB TECHNICAL UNIVERSITY OF OSTRAVA



www.vsb.cz

APPS
Architektury počítačů a paralelních systémů / Architecture of Computers and Parallel Systems

Part 01: Computer Architectures

Ing. Petr Olivka, Ph.D.
Department of Computer Science,
FEECS, VSB-TUO
petr.olivka@vsb.cz
http://poli.cs.vsb.cz

(Computer) Architecture

The word **architecture** was not invented by computers but it was taken over from other disciplines. The term does not provide exact definitions, specifications or principles. From a modern, construction point of view, it can be divided into four categories:

- **Structure, layout**: parts description and their interconnection,
- **Interaction, cooperation**: describe the dynamic communication of all working parts of a computer,
- **Realization, implementation**: describe internal structure of all working parts,
- **Functionality, activity**: final behavior of the whole computer.

To understand, how computer works, we have to deal with all of these aspects.

Brief History of Computing

- Second half of the 19^{tth} century: Charles Babbage constructed a mechanical computer “Difference and Analytical Engine”.
- 1936: Alan Turing provided a definition of a universal computer called Turing machine with algorithm execution capability.
- Before WWII: mechanical and electrical analog computers compute ballistics trajectories, momentum, inertia…
- 1946: ENIAC – first electronic general-purpose computer (18,000 vacuum tubes, 1,500 relays, 30 ton).
- 1947: John von Neumann introduced a design of a stored-program computer. This base design is known as “von Neumann architecture”. It started new epoch of computing and computers. All moderns computers nowadays use this design.

Von Neumann Architecture

John von Neumann introduced a universal computer. It must comply with some principles and criteria:

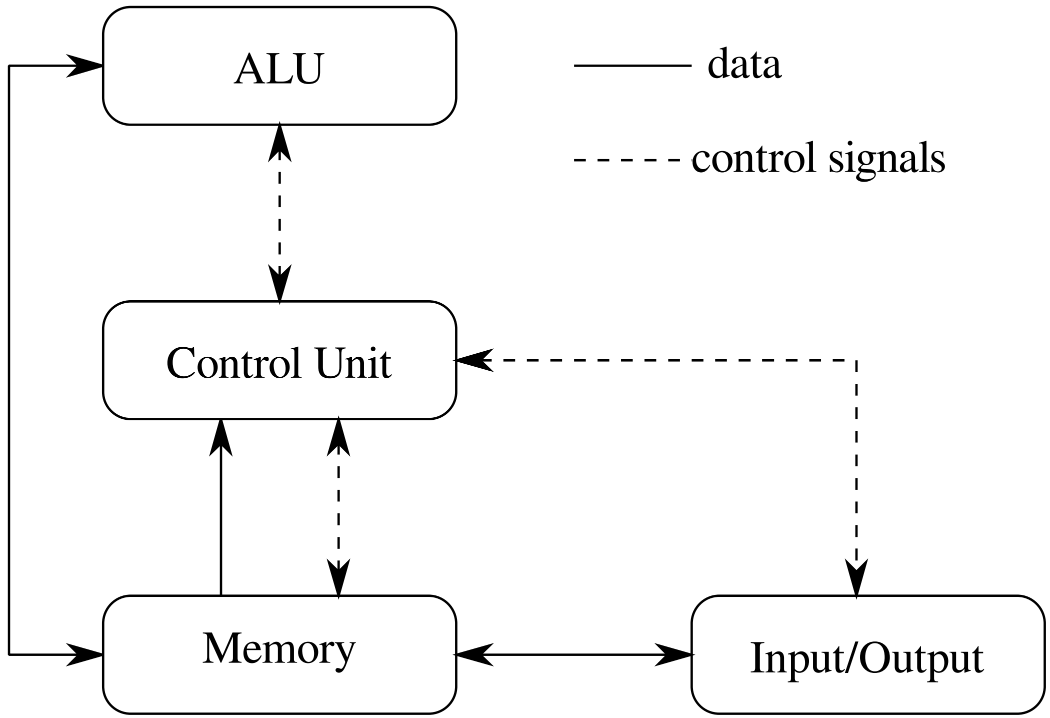
- Computer consists of memory, arithmetical-logical unit (ALU), control unit, input and output devices,
- All parts of a computer are connected together by Bus,
- Computer structure is independent on the computed problem, a computer is programmed with content of memory,
- Every computing step depends on the previous step,
- Machine instruction and data are in the same memory,
- Memory is split to small cells with the same size. Their ordinal numbers are called address numbers.

… Von Neumann Architecture

- Program consists of a sequence of instructions. Instructions are executed in order they are stored in memory.
- Sequence of instructions can be changed only by unconditional or conditional jump instructions.
- Instructions, characters, data and numbers are represented in binary form.

… Von Neumann Architecture

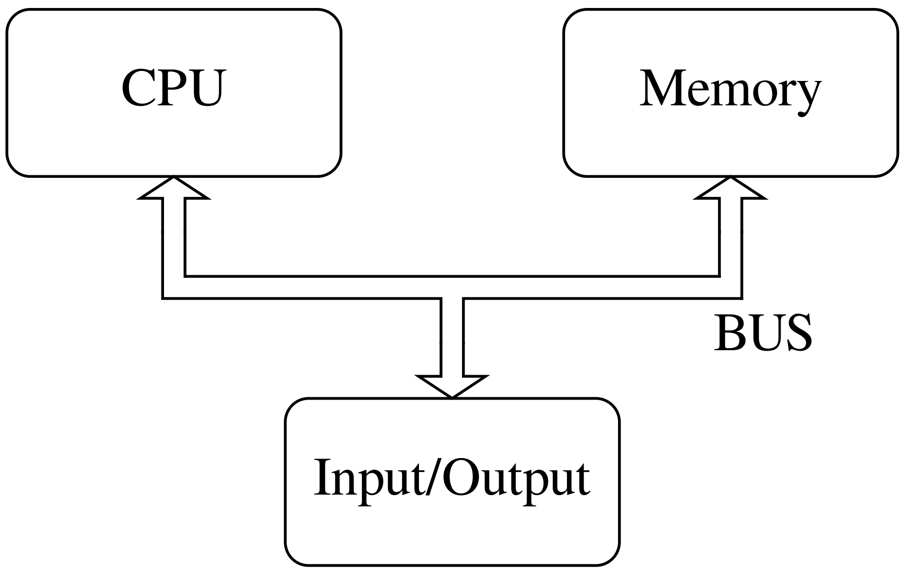
Design of the von Neumann computer is shown for clarity on next scheme:



- Data is passed between every blocks through Data Bus.
- Control Unit controls all blocks by signals passed through Control Bus.

Von Neumann Architecture Today

Thanks to the progress in computer technology, Control Unit and ALU are today integrated in one circuit called Processor or CPU (Central Processing Unit). Next scheme shows design of modern computer:



- Memory and Devices are controlled by CPU.
- Bus between blocks integrate Data Bus, Address Bus and Control Bus.
- Data can pass through bus in half duplex mode to or from CPU.

Advantages/Disadvantages of von Neumann Architecture

+ Control Unit gets data and instruction in the same way from one memory. It simplifies design and development of the Control Unit.

+ Data from memory and from devices are accessed in the same way.

+ Memory organization is in the hands of programmers.

- Serial instruction processing does not allow parallel execution of program. Parallel executions are simulated later by the Operating system.

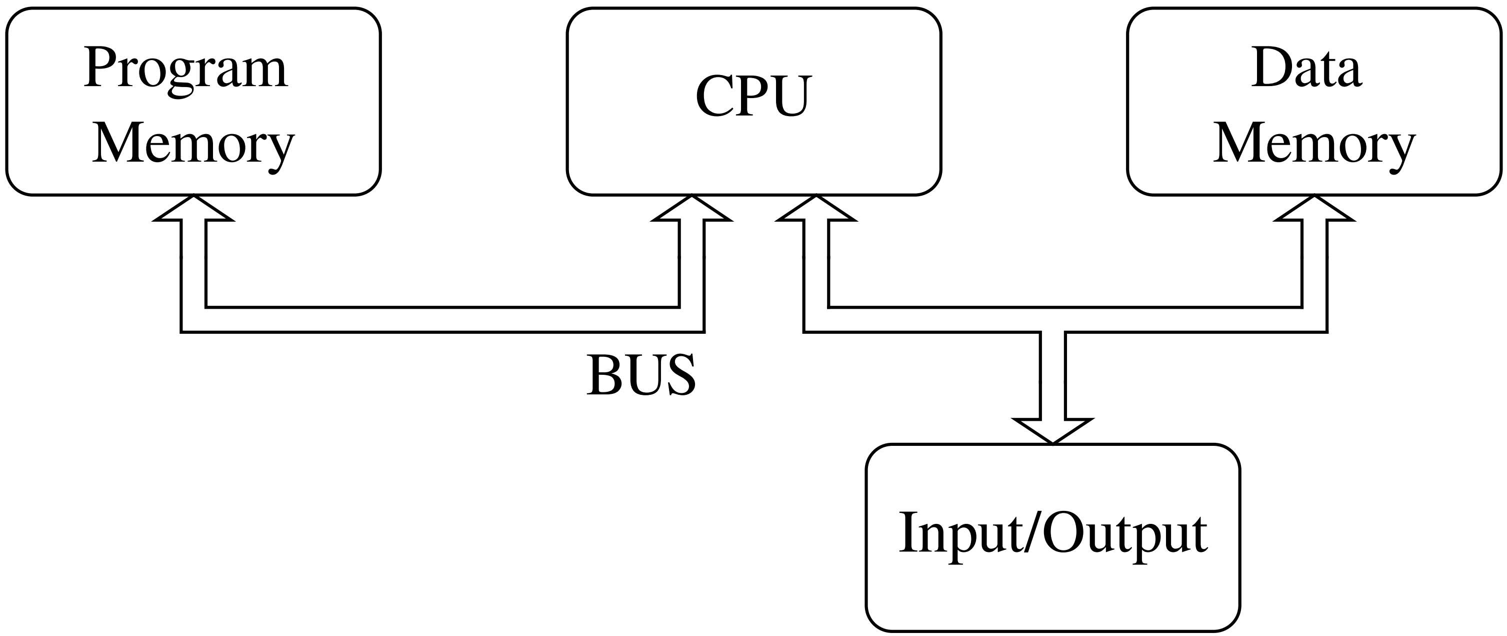
- One bus is a bottleneck. Only one information can be accessed at the same time.

- Instruction stored in the same memory as the data can be accidentally rewritten by an error in a program.

Harvard Architecture

MARK II computer was finished at Harvard University in 1947. It wasn't so modern as the computer from von Neumann team. But it introduced a slightly different architecture. Memory for data was separated from the memory for instruction. This concept is known as the **Harvard architecture**.

Modern view on Harvard architecture is on next scheme:



Harvard vs. von Neumann

Both architectures have advantages and disadvantages. It is impossible to decide, which is better. Both are still used in modern computers. We can now compare both in more detail:

- von Neumann

+ Programmers organize the content of the memory and they can use the whole capacity of the installed memory.

+ One bus is simpler for the Control Unit design.

+ Development of the Control Unit is cheaper and faster.

+ Computer with one bus is cheaper.

+ Data and instruction are accessed in the same way.

- One Bus (for data, instructions and devices) is a bottleneck.

- Error in a program can rewrite instructions and crash program execution.

... Harvard vs. von Neumann

- Harvard

  + Two memories with two Buses allow at the same time access to data and instructions. Execution can be 2x faster.

  + Both memories can be produced by different technologies (Flash/EEPROM, SRAM/DRAM).

  + Both memories can use different cell sizes.

  + Program can't rewrite itself.

  - Control unit for two Buses is more complicated and more expensive.

  - Production of a computer with two Buses is more expensive.

  - Development of a complicated Control Unit needs more time.

  - Free data memory can't be used for instruction and vice-versa.

… Harvard vs. von Neumann

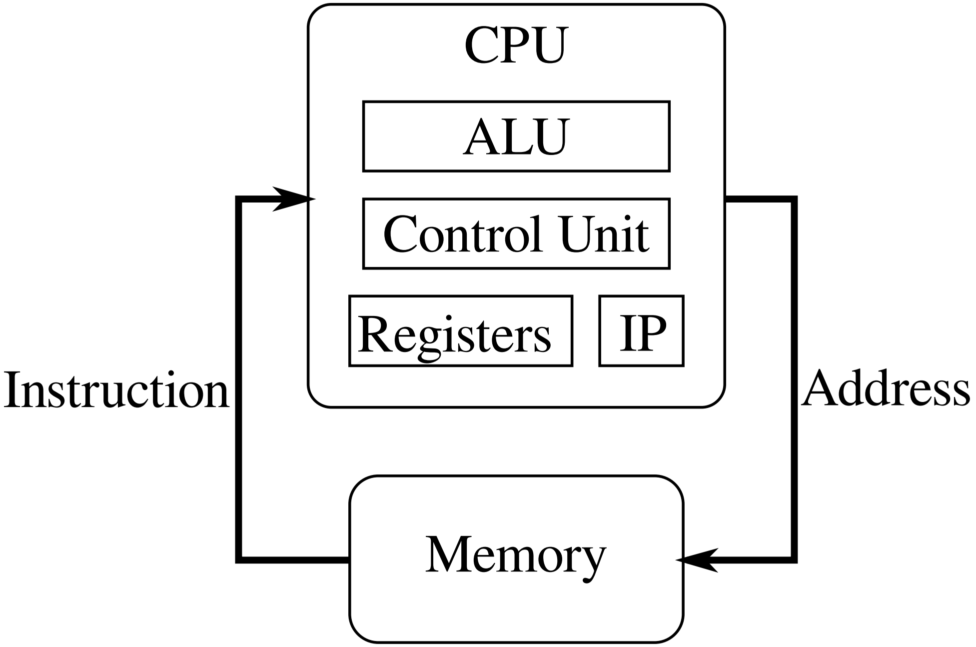
Until today both architectures are used in modern computers. Over the years few different architectures were introduced: transputers, biologically inspired computers, distributed computing, etc. But von Neumann and Harvard architecture are used massively in mainstream production.

Harvard architecture is used primary for small embedded computers and signal processing (DSP). Von Neumann is better for desktop computers, laptops, workstations and high performance computers.

Some computers may use advantages from both architectures. Typically they use two separated memories. The first one is used for programs and the second one to store dynamic data. A good example are smart watches and mobile phones.

How computer works

Only the schema with CPU and memory can be used to explain how computer works.



Von Neumann said: program is in memory, it is executed sequentially and every steps depends on previous steps. Thus CPU must be a sequential circuit.

… how computer works

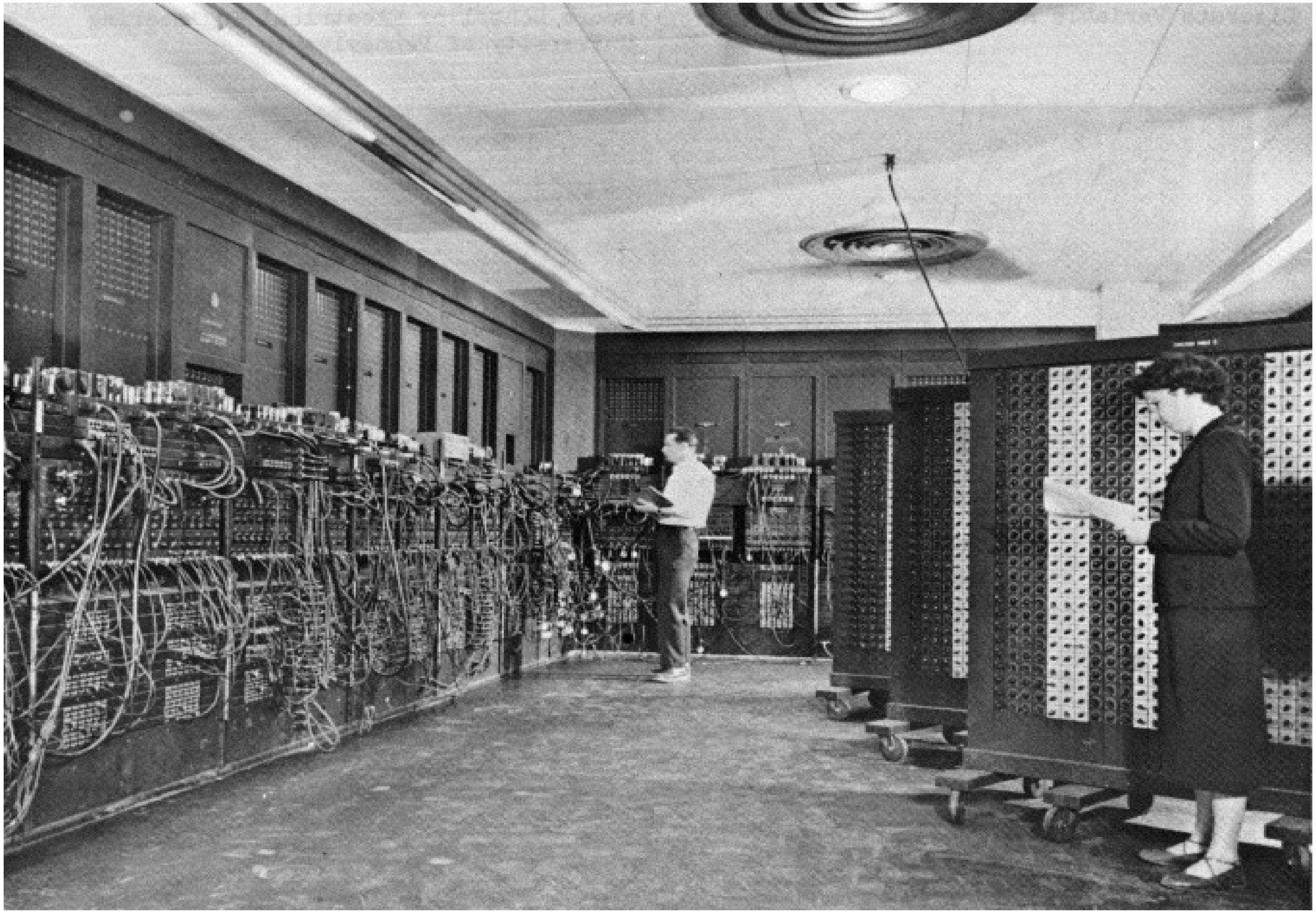
- CPU is sequential circuit where a small part of memory is implemented directly in CPU. This memory is called registers.
- One register is used to point to instructions in memory. It is mostly called IP (instruction pointer) or PC (program counter).
- IP register points to instruction which is currently executed, or will be executed. CPU uses this register to address instruction by bus in memory and this instruction is readed by bus into CPU.
- When the instruction is in CPU, it can be executed.
- If instruction requires data, CPU will use them from registers or from memory. Data from memory is readed in the same manner as instructions.
- When instruction is executed, the result is stored back into registers or memory.

Generations of Computers: 1st Generation 1945 - 56

The first generation of computers was constructed from vacuum tubes, relays, switches, crystal diodes and passive components. Drum memory was used for program storage. The typical example of the first generation computer was ENIAC (Electronic Numerical Integrator And Computer) computer mentioned earlier. Its technical parameters were:

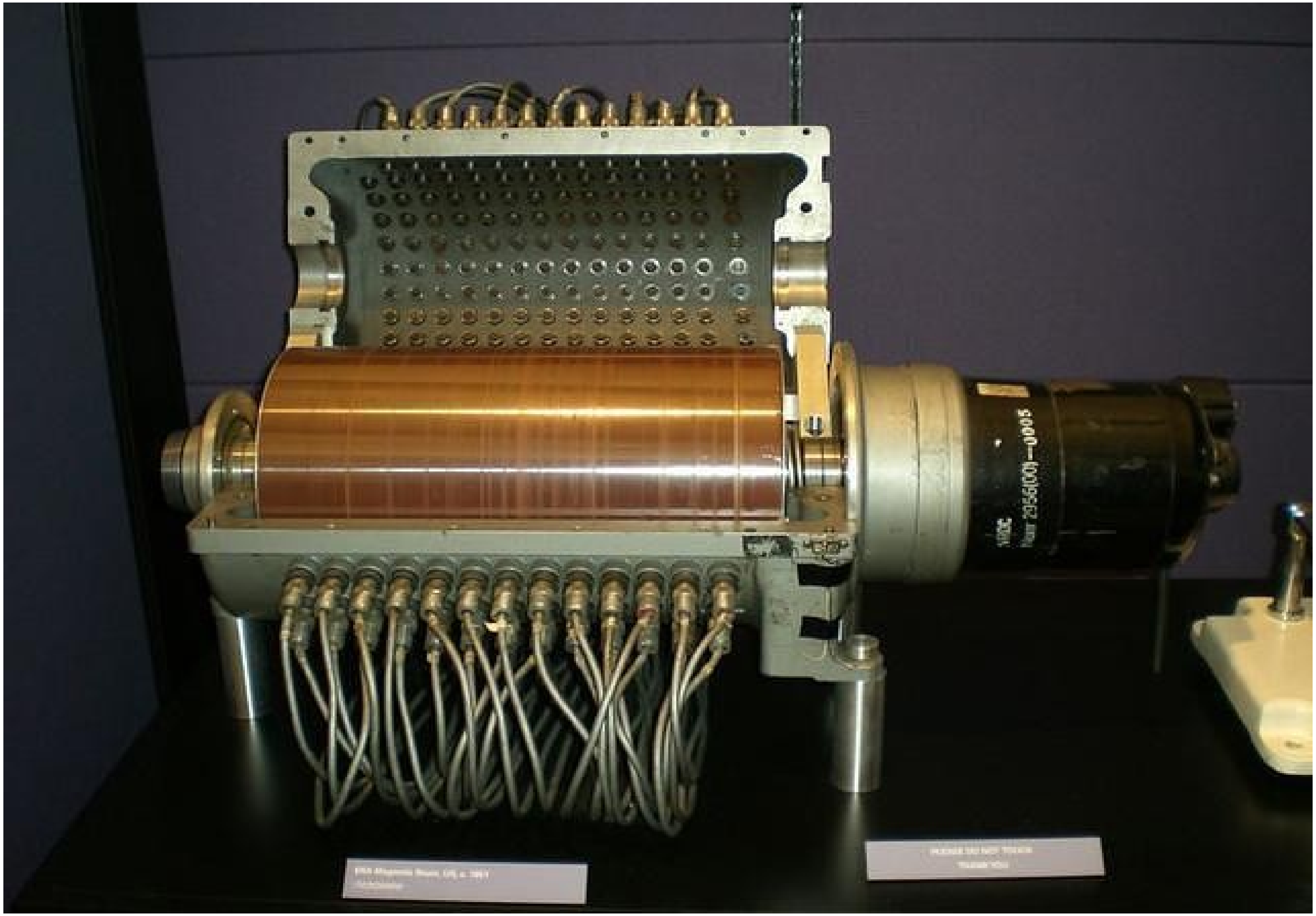
- Up to 18,000 vacuum tubes and 7,200 crystal diodes,
- 70,000 resistors, 10,000 capacitors and inductors,
- 1,500 relays, 6,000 switches and 5 million hand-soldered joints,
- Weight of 30 tonnes, power consumption of up to 200kW,
- Performance of up to 5,000 operations per second,
- Programmed in machine code, to change a program one week was required.

… 1st Generation - ENIAC



Which of 18,000 tubes is bad?

… 1st Generation – Magnetic Drum

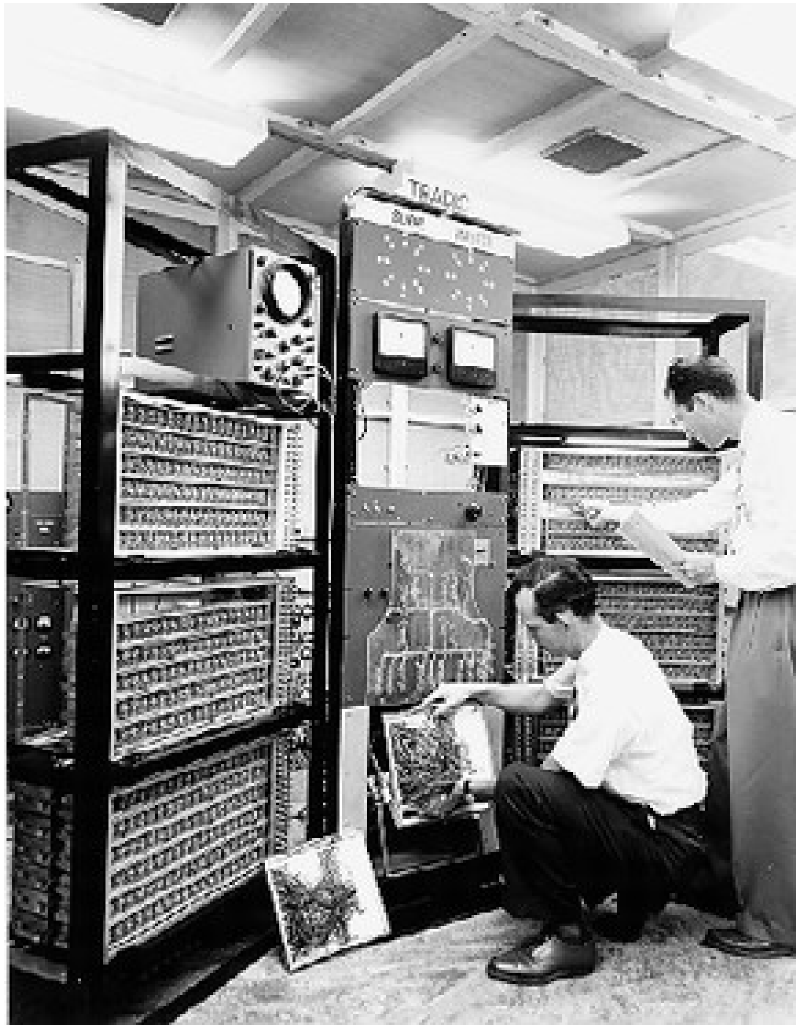


Generations of Computers: 2nd Generation 1956-65

The replacement of vacuum tubes by transistors is characteristic for the second generation of computers. The transistor is a semiconductor invented in Bell Labs in 1947. Transistors allow computers to become smaller, faster, cheaper, more energy-efficient and more reliable than their first-generation predecessors.

- Magnetic drum memory was replaced by magnetic ferrite core memories technology. Up to 16kB of core memory available.
- Program was read from punched card or tape, output printed.
- Magnetic tapes and discs were used as external memory.
- Programming moved from cryptic binary machine language to symbolic language or early version of FORTRAN and COBOL.
- First operating system was developed, batch processing used.

… 2nd Generation – Computer TRADIC



First fully transistorized computer from AT&T – using about 800 transistors.

… 2nd Generation – Mainframe ICT 1301

Typical fully transistorized mainframe from UK - ICT 1301.

Generations of Computers: 3rd Generation 1965-80

The usage of integrated circuit with small and medium scale integration is typical for the 3rd generation of computers. Miniaturized transistors on silicon drastically increased the speed and efficiency of computers.

- Fixed discs were used to save program from punched cards.
- Spooling – parallel I/O operation during process execution.
- First memory made from IC – higher capacity and speed.
- Multiprogramming to run more jobs at the same time.
- Timesharing to allow multi-users access to computer at the same time with keyboard and CRT monitors.
- Complex OS such as OS/360 (developed by IBM) were used.
- Used for various applications including scientific and business applications.
- First minicomputer was introduced.

… 3rd Generation – IBM 360

The computer legend – IBM 360.

… 3rd Generation – Terminal

Typical terminal used by users in timesharing systems.

Generations of Computers: 4th Generation 1980 …

The fourth generation of computer is characterized by VLSI IC usage. Computers were smaller, mainframes turned into workstations and desktops. Speed and efficiency is still growing up.

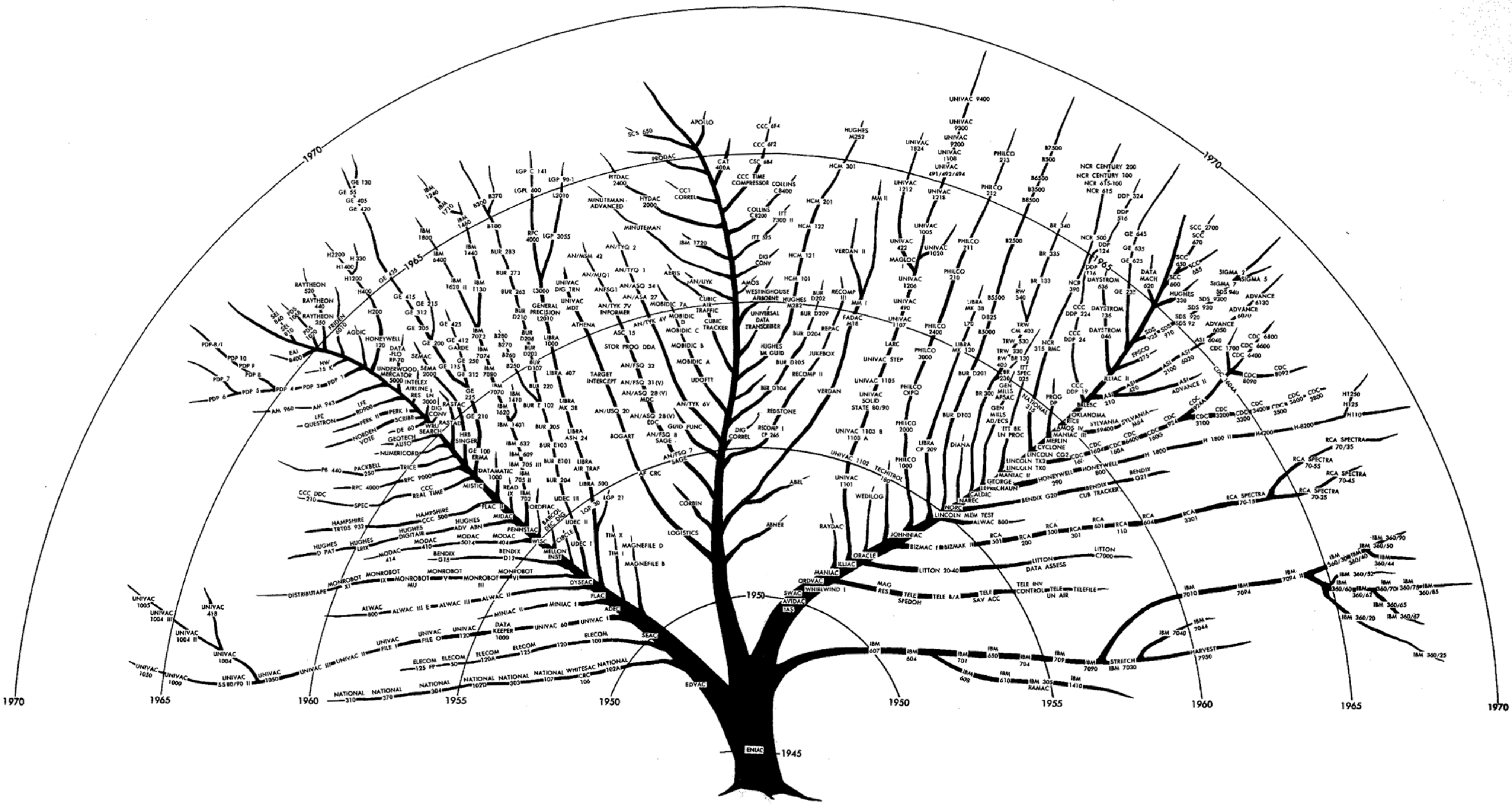
- Thousands of ICs were integrated to one silicon chip. First single-chip CPUs (ALU with Control Unit) were developed.
- Semiconductor memory DRAM and SRAM were used as the main computer memory.
- Secondary memory was composed of hard discs and floppies.
- Multiuser OS with GUI, network OS, distributed OS developed.
- Parallelism, pipelining cache memory and virtual memory were applied in a better way.
- Many new programming languages (C, Pascal, …).
- Color CRT monitors, keyboard and mouse.

Generations of computers: Tabular Summary

Generation	1st	2nd	3rd	4th
Active element	Vacuum tube	Transistor	SSI and MSI IC	(V)LSI IC
Memory	Drum	Ferrite	Ferrite and IC	(V)LSI IC
Memory capacity	In kB	In 10kB	In 1MB	In 10MB
Communication with devices	Synchr. CPU	Asynchr. CPU	Channels	Peripheral Contollers

Generations of Computers: Tree from ENIAC - 1970

Enlarge it.



Units for Memory Capacity (technical vs. business)

All information in computer is stored in binary form. In the same way we compute memory capacity – in multiples of powers of 2:

- 1 B (Byte) is 8 bits – usually smallest addressable cell
- 1 “kB” = 2^{10} B = 1,024 B
- 1 “MB” = 2^{20} B = 1,048,576 B
- 1 “GB” = 2^{30} B = 1,073,741,824 B
- 1 “TB” = 2^{40} B = 1,099,511,627,776 B

These units are used in operating systems and in programs to compute and manage the capacity of primary or secondary memory. However many promotional products and advertisement use **decimal** units:

$2^{10} \cong 10^3$, thus $1024 \cong 1000$.

The decimal units twist reality and for the 1”TB” capacity the discrepancy is as high as **10%**!

... Units for Memory Capacity - IEC 60027-2

The international technical standard specify binary prefixes:

- 1 B (Byte) is 8 bits – usually smallest addressable cell
- 1 kiB (kibi) = 2^{10}
- 1 MiB (mibi) = 2^{20} B
- 1 GiB (gibi) = 2^{30} B
- 1 TiB (tebi) = 2^{40} B

In 2008, the IEC prefixes were incorporated into the International System of Quantities alongside the decimal prefixes of the international standard system of units