Computer Architecture. Week 1

Alexander Tormasov

Innopolis University

a.tormasov@innopolis.ru

August 26, 2021

Topic of the lecture

- Introduction to computer architecture

Content of the class

- Computer Technology
- Classes of Computers
- What is Computer Architecture?
- Overall Structure
- Arithmetic Logic Unit (ALU)
- Control Unit (CU)
- The Registers
- The Bus
- Main Memory
- Disk Storage
- Anatomy: 5 Components of any computer
- I/O Devices
- Evolution in Computer Architecture Design
- Moore's Law

Computer Technology

I think it's fair to say that personal computers have become the most empowering tool we have ever created. They are tools of communication, they are tools of creativity, and they can be shaped by their user.

– Bill Gates, February 24, 2004

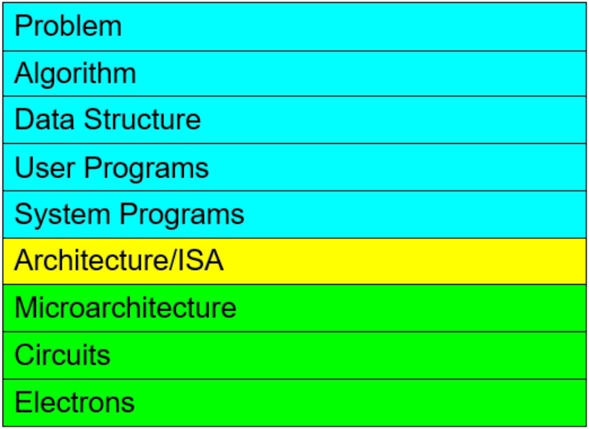
Some Classes of Computers

- Microcomputers
  - Personal mobile devices, PMD (power efficiency and real-time)
    - laptops, tablets, smartphones, etc.
  - Desktop computers (price-performance focus)
  - Video game consoles (price-performance, graphics processing)
- Servers (mail, web, file, etc.)
  - Provide services to remote users connected via network
  - Focus on large processing throughput, availability, and stability
- Clusters/Warehouse scale computers (incl. Supercomputers)
  - Focus on high performance and stability
  - Used for "Software as a Service (SaaS)", cloud systems
  - Supercomputers included (for intense numerical calculations)
- Embedded computers
  - "Embedded" into a device (e.g. part of a car), to constantly perform a single or few limited functions
  - Focus on availability and price

What is Computer Architecture?

The science and art of **designing**, selecting, and interconnecting **hardware components** and designing the hardware/software **interface** to create a computing system that meets functional, performance, energy consumption, cost, and other specific goals.

Problem Solution Stack in Modern World



Problem

Algorithm

Data Structure

User Programs

System Programs

Architecture/ISA

Microarchitecture

Circuits

Electrons

ISA: Instruction Set Architecture

Architecture vs Microarchitecture

- Architecture
  - Programmer's view of computer
    - Defined by instructions and operand locations
- Microarchitecture
  - How to implement an architecture in hardware

The Von Neumann Architecture

- Also called stored program computer (instructions in memory). It has two key properties:
- Stored program
  - Instructions stored in a linear memory array (directly addressable!)
  - Memory is unified between instructions and data
    - The interpretation of a stored value depends on the control signals
- Sequential instruction processing
  - One instruction processed (fetched, executed, and completed) at a time
  - Program counter (instruction pointer) identifies the current instruction.
  - Program counter is advanced sequentially except for control transfer instructions
- Other architectures are available, including Harvard (will be discussed later)

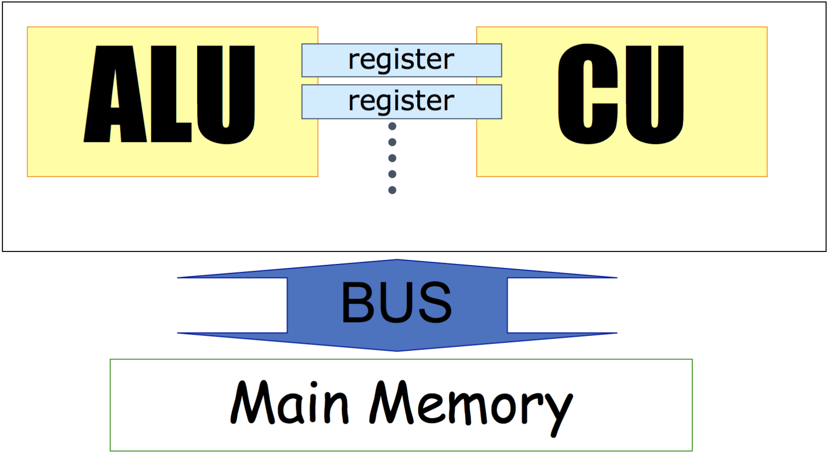
Microarchitecture

- How the underlying implementation actually executes instructions
- Microarchitecture can execute instructions in any order as long as it obeys the semantics specified by the ISA when making the instruction results visible to software
- Programmer should see the order specified by the ISA

Examples

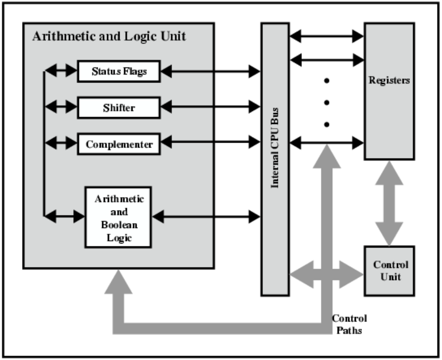
- All major instruction set architectures today use Von-Neumann model (with minor variations, such as separate instruction and data caches, and multi-core clusters).
  - For example: x86, ARM, MIPS, SPARC, Alpha, POWER, ...
- Underneath (at the microarchitecture level), the execution model of almost all implementations (or, microarchitectures) is very different
  - Pipelined instruction execution: Intel 80486
  - Multiple instructions at a time: Intel Pentium
  - Out-of-order execution: Intel Pentium Pro
  - Separate instruction and data caches
- But, what happens underneath that is not consistent with the von Neumann model is not exposed to software

An Oversimplified Computer Organization



The Control Unit

- It is the core of the sequencing of operations
- Picks the new operation to be executed
- Decodes it
- Coordinates its execution



The Arithmetic and Logic Unit (ALU)

- Is the core of the “computation”
- Performs arithmetic, logic and shift operations
- All other operations are combinations of these basic operations
- Works on numbers in base 2, usually (more detail in next lecture)

The Registers

- Are the places where we put the data we need for the actual execution
- Limited in size and very fast to access

The Communication Bus

- A communication system to transfer data between different components inside or outside of a computer, like
  - CPU <-> main memory (RAM)
  - main memory <-> peripheral devices (display, printer, etc.)
- Bus system components:
  - (a shared) communication medium (e.g. wire)
  - bus controller (to control bus usage, e.g. to avoid data collision)
  - communication protocol (rules of when, between which components, and how much data can be transferred)
- Multiple bus types are available, different in communication speed supported, purpose, etc.:
  - sequential and parallel
  - internal and external
  - on-chip and off-chip

Main Memory

- Addressed directly – sometimes said “randomly”, hence RAM
- Fast access, no as fast as register, but still fast
- Volatile structure

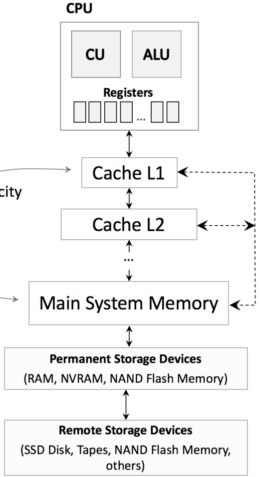
Disk Storage (incl. SSD and HDD)

- Slower to access than main memory (RAM), up to 100s times
    - SSD speed: 50-200 MB/sec
    - RAM speed: 2-20 GB/sec
- Sequential in accessing nature (while randomly accessible)
- Larger capacity
- Permanent storage

Anatomy of Computer

A faster memory unit,
But a smaller storage capacity

A slower memory unit,
But a larger storage capacity



CPU

CU

ALU

Registers

Cache L1

Cache L2

Main System Memory

Permanent Storage Devices
(RAM, NVRAM, NAND Flash Memory)

Remote Storage Devices
(SSD Disk, Tapes, NAND Flash Memory, others)

Note: Some hardware architectures allow a direct communication to main memory, bypassing some caches

Peripheral Storage Devices
(e.g. SSD disk, tape, NAND flash memory)
Details and complete classification will be studied later in the course

Input Devices

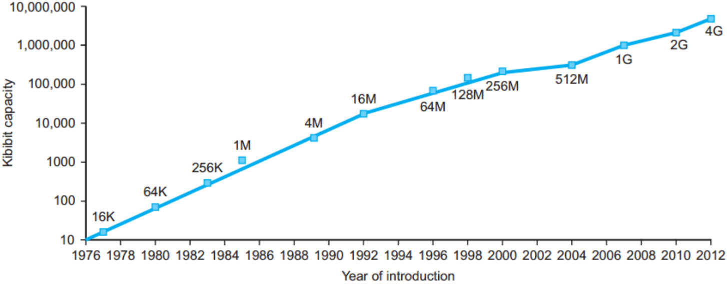
Peripheral devices are connected to the computer by using dedicated communication buses and controllers, either wired or wireless

Output Devices

Peripheral devices are connected to the computer by using dedicated communication buses and controllers, either wired or wireless

Memory Evolution

Memory Capacity Single-Chip DRAM



- The y-axis is measured in kibibits (2^{10} bits).
- The DRAM industry quadrupled capacity almost every three years
- In recent years, the rate has slowed down and is somewhat closer to doubling every two years to three years

Average State-of-the-Art PC Specs

- Processor clock speed: 2-3.5 GHz (up to 10.0 GHz, see note)
- 100X performance in last decade.
- Memory capacity: 10,000 MB (10.0 GB)
- Disk capacity: 20,000 GB (20.0 TB)
- New units! Mega =>Giga, Giga =>Tera =>Peta =>Exa =>Zeta =>Yotta

NOTE: 10GHz clock speed is not practical due to heat dissipation problems. On modern computers the clock rate achieved over 3.5Ghz and later was scaled back due to aforementioned problems. The current direction of improving performance is different from increasing clock rate: adding more parallelism to microarchitecture, adding hardware support for multithreading, using multi-core clusters, and specialized hardware, optimized for certain tasks (like GPU for 3D graphics)

Front-line State-of-the-Art Computer Specs

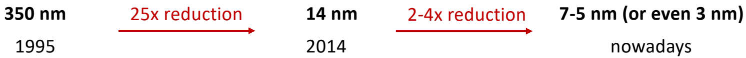
- Smartphone (expected specs for iPhone 13 Pro):
  - CPU: Apple A15 Bionic (5 nm); 6 cores*, 2.85GHz (?) speed
  - RAM: 8 GB RAM
  - Storage capacity: 128 GB - 1 TB
- Notebook (Dell XPS'17 (2020)):
  - Processor: 2.4-5.3 GHz Intel Core i9-10885H (10th Generation)
  - Memory capacity: Up to 64GB DDR4-2933MHz
  - Disk capacity: Up to 2TB SSD
- Server (Intel Xeon Server)
  - Processor: 2.2 GHz Xeon Processor E7 (60 MB Cache, 24 Cores*)
  - Memory capacity: Up to 3TB
  - Disk capacity: 12 TB

*A processor having more cores is capable to execute more instructions simultaneously, yielding a better performance.

Current Performance Trends

- Processor
  - 2X in speed every 1.5 years (from 1985 up to around 2010)
  - 100X performance in last decade.
- Memory
  - DRAM capacity: 2x / 2 years (since 96)
  - 64x size in last decade.
- Disk
  - Capacity: 2X / 1 year (since 97)
  - 250X size in last decade.

Processor Technology Trends



350 nm
1995

25x reduction →

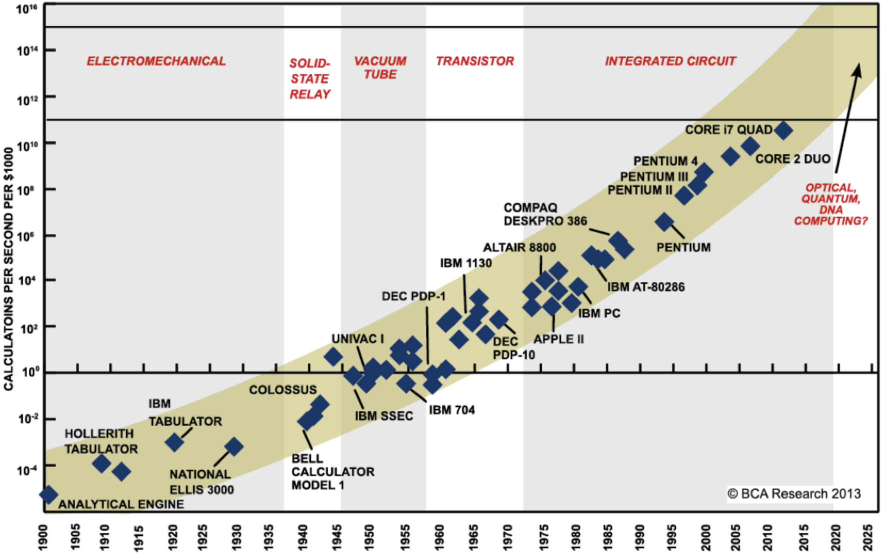
14 nm
2014

2-4x reduction →

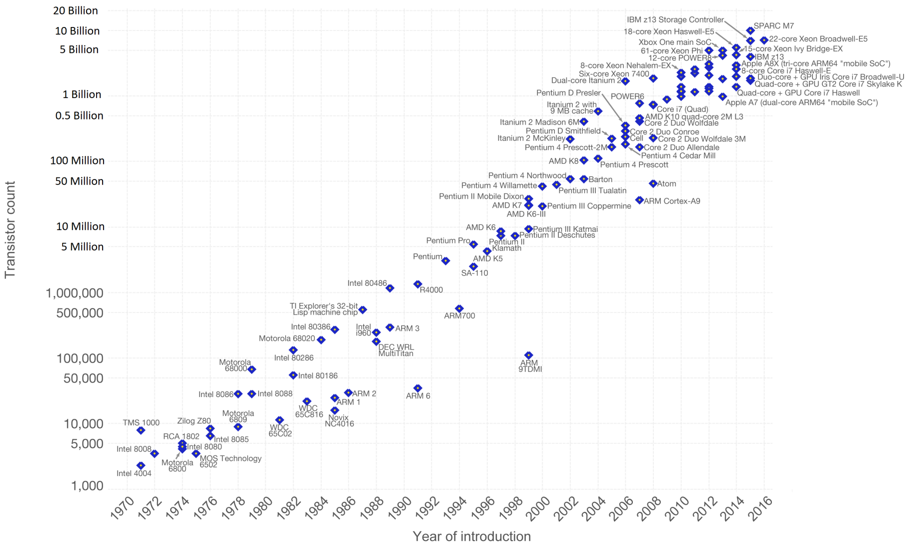
7-5 nm (or even 3 nm)
nowadays

- Real performance rising $<$ 25x
  - Multicore system use 25-50% of peak performance
  - Why 8-cores smartphone is better than 4-cores one?
- **Window of opportunity** for further performance increase:
  - SoC (System-on-a-Chip) customization for particular applications, to achieve the best performance at minimal cost and power consumption
  - customization and new generation of EDA (Electronic Design Automation)
  - New system level design technologies and EDA tools allow to achieve the best time to market in the industry

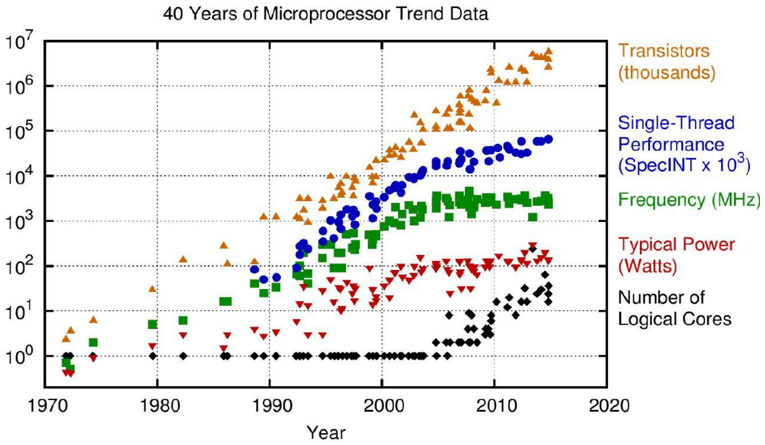
Microprocessor Complexity (Moore’s Law)



No. of transistors on IC chips (Moore's Law)



Processor Performance



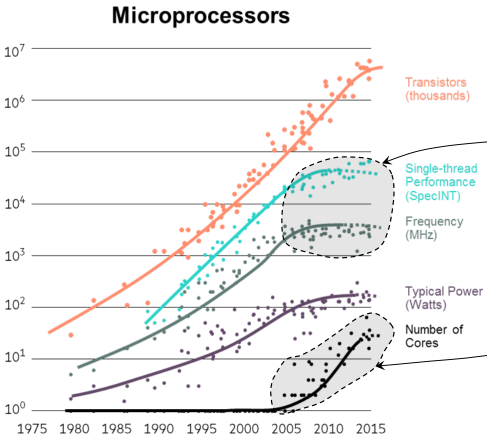
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

Processor Performance

In the next lecture we will learn:

- Performance evaluations is a complex issue
- We cannot judge performance by clock frequency or average number of instructions per cycle alone
- Processors with the same frequency can have different performances.
- For fair comparison a multitude of parameters should be taken into account (including clock rate, throughput, memory performance, etc.)
- Ideally, one would compare performance running his own application on different computers. However it is not practical, so the industry uses instead synthetic benchmarks that mimic the behaviour of typical user programs

Processor Cores vs. Core Speed Trend



CPU speed and single-thread performance seem not to increase since ~2008

The number of cores increases instead

Nowadays, the increase of performance is achieved through the number of CPUs (or CPU cores), rather than a further increase of CPU speed

Image is taken from https://www.quora.com/Why-is-Moores-law-no-longer-valid

Summary

- Computer architecture: Programmer's view of computer
- Microarchitecture: How to implement an architecture in hardware
- The Von Neuman model: Describes a design architecture for an electronic digital computer
- Computer structure
- Evolution in technology

Acknowledgements

- This lecture was created and maintained by Muhammad Fahim, Giancarlo Succi, Alexander Tormasov, and Artem Burmyakov