

CS202 : COMPUTER ORGANIZATION

Lecture 1 Course Introduction History of Computers

2023 Spring

Today's Agenda

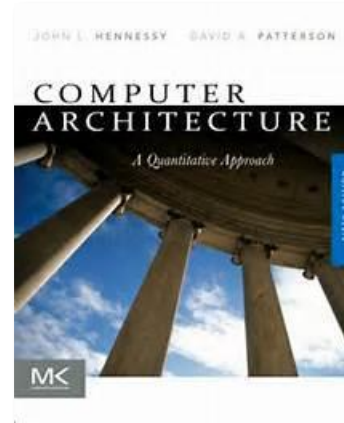
- Introduction to course
- Context
 - Computer: History, Abstractions and Technology
- Reading: Textbook, Sections 1.1 - 1.5.

Course Information

- Course website: [CS202-30022126-2023SP](#)
- Instructor:
 - Dr. Yuhui BAI (baiyh@sustech.edu.cn)
 - Office: 411 College of Engineering South
 - Office hour: Wednesday 14:00-16:00
- Session 01
 - Lecture:
 - Monday 14:00-15:50, Room 301, Lecture Hall #3
 - Lab:
 - Monday 16:20-18:10, Room 511, Lecture Hall #3
- Session 04
 - Lecture:
 - Tuesday 10:20-12:10, Room 101, School of Business
 - Lab:
 - Wednesday 10:20-12:10, Room 501, Lecture Hall #3
 - Wednesday 10:20-12:10, Room 503, Lecture Hall #3 (by Wei WANG)

Textbook

- Textbooks
 - Computer Organization & Design, the Hardware/Software Interface, D. A. Patterson and J. L. Hennessy, 5th edition
- Reference book:
 - Computer Architecture - a quantitative approach, Hennessy and Patterson, 5th edition



Motivations

- Why to learn Computer Organization?
 - Must know how to reason about the program performance and energy consumption
 - **CPU Performance**: if you understand how CPU processes data, you can enhance the computing efficiency
 - **Memory Management**: if you understand how and where data is placed, you can ensure that relevant data is easily accessible
 - **Thread Management**: if you understand how threads interact, you can write better multi-threaded programs
 - **I/O Management**
- Why the course is important?
 - Computer organization principles are everywhere:
 - Complex system design:
 - How to partition a problem?
 - Functional Specifications -> Control & Datapath -> Physical implementation
 - Modern CAD (computer-aided design) tools
 - Both EEs and CSEs need this information in almost all jobs!

Prerequisites

- Binary numbers
- Read and write basic C/Java programs
- Understand the steps in compiling and executing a program
- Digital Circuit, Logic design:
 - Logical equations, schematic diagrams
 - Combinational vs. sequential logic
 - Finite state machines (FSMs)

Course contents

- Course Goals
 - Learn the components of a computer and their relations
 - Learn the interface between software and hardware
 - Design a simple CPU
- Course Contents
 - Introduction (Chapter 1)
 - Basic terms
 - Moore's Law, power wall
 - Core ideas in computer architecture
 - Processors(Chapter 2-4)
 - Instruction set architecture(Chapter 2)
 - Computer arithmetic (Chapter 3)
 - Single Cycle CPU (Chapter 4)
 - Pipelining (Chapter 4)
 - Memory (Chapter 5)
 - Parallel Processors (Chapter 6)

Tentative Schedule

WEEK	LECTURE	DATE	TOPIC
1	Lecture #1	Feb. 13(14), 2023	Introductions
2	Lecture #2	Feb. 20(21), 2023	MIPS ISAs: basics
3	Lecture #3	Feb. 27(28), 2023	MIPS ISAs: Procedure Call
4	Lecture #4	Mar. 6(7), 2023	MIPS ISAs: Addressing
5	Lecture #5	Mar. 13(14), 2023	Performance; Overview of RISC-V
6	Lecture #6	Mar. 20(21), 2023	Arithmetic
7	Lecture #7	Mar. 27(28), 2023	Floating Point Arithmetic
8	Lecture #8	Apr. 3(4), 2023	The Processor
9	Lecture #9 Mid-term Exam	Apr. 10(11), 2023	The Pipeline Mid-term Exam Lecture #1—#8
10	Lecture #10	Apr. 17(18), 2023	Instruction-Level Parallelism
11	Lecture #11	Apr. 24(25), 2023	Exploiting Memory Hierarchy
12	N.A.	May 1(2), 2023	Holiday
13	Lecture #12	May. 8(9), 2023	Exploiting Memory Hierarchy(cont.)
14	Lecture #13	May. 15(16), 2023	Exploiting Memory Hierarchy(cont.)
15	Lecture #14	May 22(23), 2023	Parallel Processors
16	Lecture #15	May 29(30), 2023	Revision

Grading criteria

- ~10%-20% Assignments & Quizzes
 - No late submission is allowed without formal approval from the lecturer before deadline.
 - No re-submission is allowed for student reasons, e.g., corrupted file uploaded.
- ~25%-35% Mid-term examination
- ~25%-35% Final examination
- ~20%-30% Lab project
- ~5% Attendance

Honor policy

- All course work should be completed entirely on your own. You are encouraged to discuss general concepts and ideas in homework or lab assignments.
- Students who commit an act of academic dishonesty may receive a zero on the assignment (first conduct) or in the course (multiple conducts).
- Unless otherwise noted, exams and individual assignments will be pledged that you have neither given nor received unauthorized help.
- If you have questions on what is allowable, ask!

Computer Abstractions and Technology

- Classes of Computing Applications
- Modern Computing (or Post-PC Era)
- Eight Great Ideas in Computer Architecture
- Computer Components
- Processor and Storage Technologies
- Programs and Hardware

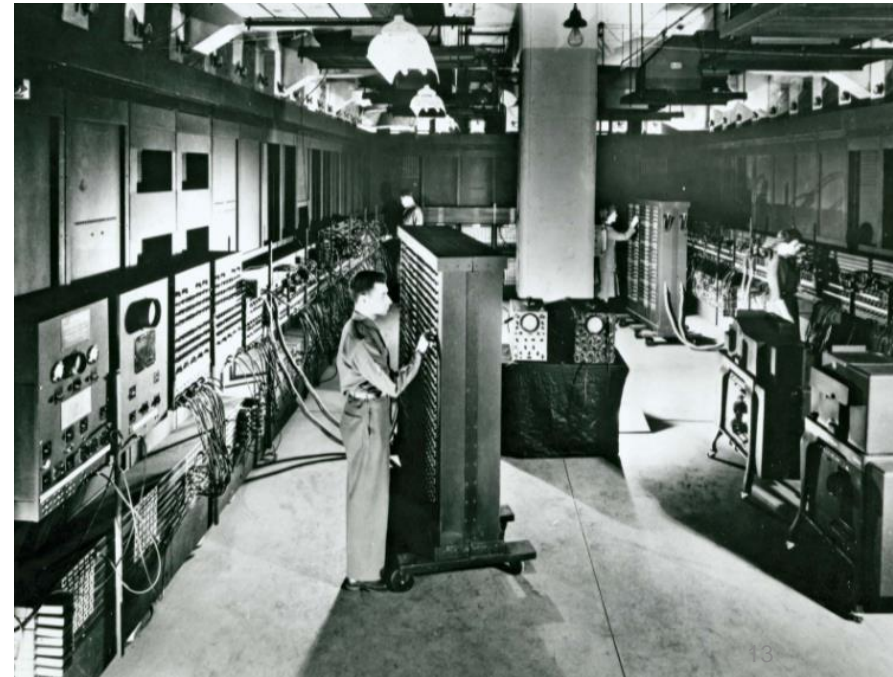
History of Computers

- 5 Generations of Computers
- First Generation (1940s - 1950s)
 - **Vacuum Tubes**
- Second Generation (1950s - 1960s)
 - **Transistors**
- Third Generation (1960s - 1970s)
 - **Integrated Circuits**
- Fourth Generation (1970s - Present)
 - **Microprocessors**
- Fifth Generation (Present and Beyond)
 - **Artificial Intelligence**

First Generation Computers

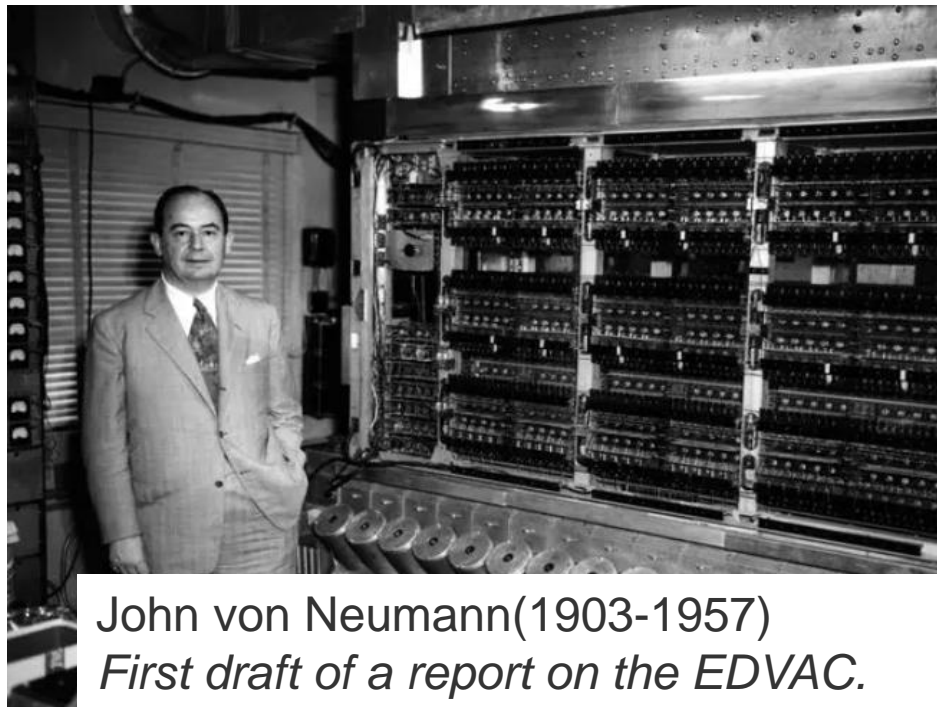
- ENIAC (Electronic Numerical Integrator and Calculator)
 - built in World War II was the first general purpose computer.
 - contained 18,000 vacuum tubes, very heavy, large
 - Program by setting switches or wiring cables
 - 5,000 additions/second
 - decimal system

Main electronic component	Vacuum tube.
Programming language	Machine language.
Main memory	Magnetic tapes and magnetic drums.
Input/output devices	Paper tape and punched cards.
Speed and size	Very slow and very large in size (often taking up entire room).



First Generation Computers

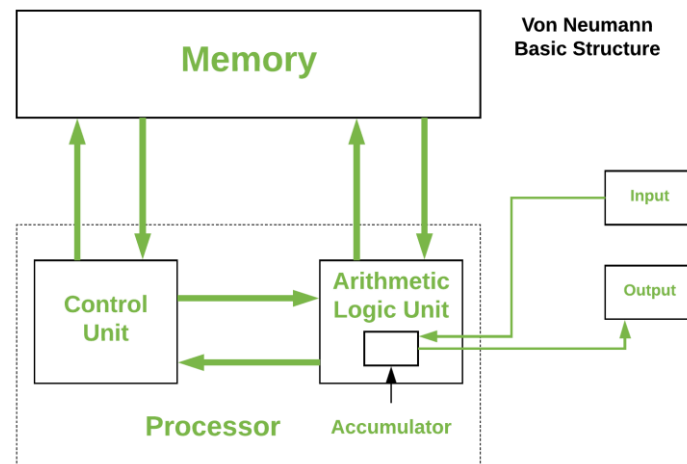
- EDVAC (Electronic Discrete Variable Automatic Computer)
 - Invented by Von Neumann
 - First Stored Program Computer. Uses Memory
 - Using binary system, could store data also as instruction and thus the speed was enhanced.



John von Neumann(1903-1957)
First draft of a report on the EDVAC.

Von Neumann architecture

- Also known as IAS (Institute for Advanced Studies) computer
- General structure of the Von Neumann computer consists of
 - The Central Processing Unit (CPU)
 - Control unit (CU) interprets instructions in memory to be executed.
 - Arithmetic and logic unit (ALU) capable of operating on binary data.
 - Main memory stores data and instructions.
 - Input and output (I/O) equipment operated by the control unit.
- Importance: all of today's computers have this same general structure and function and are thus referred to as von Neumann machines.



Von Neumann Bottleneck

- Von Neumann architecture uses the same memory for instructions (program) and data.
- The time spent in memory accesses can limit the performance. This phenomenon is referred to as *von Neumann bottleneck*.
- To avoid the bottleneck, later architectures restrict most operands to registers (temporary storage in processor).

Ref.: D. E. Comer, *Essentials of Computer Architecture*, Upper Saddle River, NJ: Pearson Prentice-Hall, 2005, p. 87.

Second Generation Computers

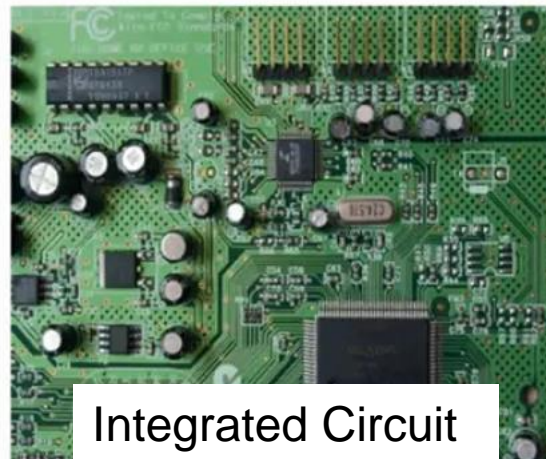
- Transistor(invented in 1947 at Bell Labs) replaced vacuum tubes
- smaller in size, low power consumption, and generated less heat
- batch processing
- Machine language and assembly language.
- Floating-point arithmetic



IBM 7094

Third Generation Computers

- Integrated circuit (IC) technology
- Semiconductor memories
- Memory hierarchy, virtual memories and caches
- Time-sharing
- Parallel processing and pipelining
- Microprogramming



C Programming Language and UNIX Operating System



Dennis Ritchie
(1941-2011)



Ken Thompson
1943-

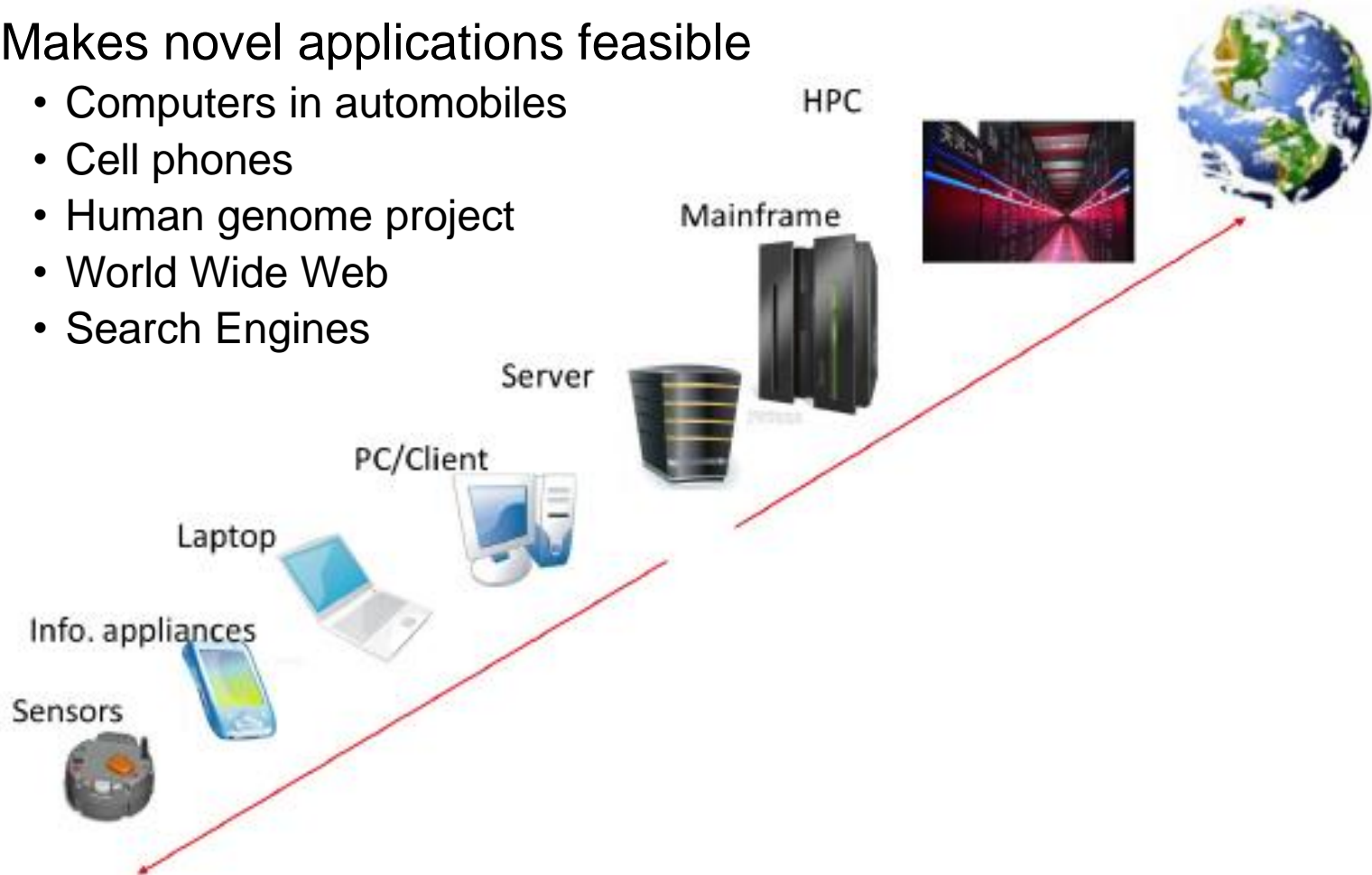
The Now Generation

- Very large-scale integration (VLSI) and the microprocessor
- semiconductor memory (such as RAM, ROM, etc.)
 - Personal computers
 - Laptops and Palmtops
 - Networking and wireless
 - SOC and MEMS technology
- And the future!
 - Biological computing
 - Molecular computing
 - Nanotechnology
 - Optical computing
 - Quantum computing



Computers now

- Progress in computer technology
 - Underpinned by Moore's Law
- Makes novel applications feasible
 - Computers in automobiles
 - Cell phones
 - Human genome project
 - World Wide Web
 - Search Engines



Classes of Computers

- Broadly speaking, computers are used in three different classes of applications:
 - Personal computers (PCs)
 - Servers
 - Embedded computers.
- PostPC Era
 - personal mobile device (PMD).
 - Warehouse Scale Computers (WSCs).

Personal Computers

- Personal computers
 - Computers designed for use by an individual
 - General purpose, variety of software
 - Subject to cost/performance tradeoff

Servers

- Server computers
 - Computers used for running larger programs for multiple users, often simultaneously
 - Network based
 - High capacity, performance, reliability
 - Range from small servers to building sized
- Supercomputers
 - High-end scientific and engineering calculations
 - Highest capability but represent a small fraction of the overall computer market

Top10 Supercomputers

- <https://www.top500.org/lists/top500/2022/11/>

Rank	Site	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	DOE/SC/Oak Ridge National Laboratory United States	Frontier - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11 HPE	8,730,112	1,102.00	1,685.65	21,100
2	RIKEN Center for Computational Science Japan	Supercomputer Fugaku - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D Fujitsu	7,630,848	442.01	537.21	29,899
7	National Supercomputing Center in Wuxi China	Sunway TaihuLight - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway NRCPC	10,649,600	93.01	125.44	15,371
10	National Super Computer Center in Guangzhou China	Tianhe-2A - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 NUDT	4,981,760	61.44	100.68	18,482

- Common size terms: KMGTPPEZY

Tianhe-2 supercomputer

- Rank 10, LINKPACK performance evaluation November, 2022
 - 61 Petaflop super computer located in Guangzhou, china
 - Used for simulation analysis in government security applications
 - Intel Xeon CPUs
 - 4,981,760 CPU cores



Sunway TaihuLight supercomputer

- Rank 7, LINKPACK performance evaluation November, 2022
 - 93 Petaflop super computer located in Wuxi, china
 - 40960 Chinese design SW26010 processors
 - 10,649,600 CPU cores



IBM Blue Gene

(2 processors/chip) • (2 chips/compute card) • (16 compute cards/node board) • (32 node boards/tower) • (64 tower) = 128k = 131072 (0.7 GHz PowerPC 440) processors (64k nodes)

System Location: Lawrence Livermore National Laboratory

Networks:
3D Torus point-to-point network
Global tree 3D point-to-point network
(both proprietary)

Cabinet
(32 Node boards, 8x8x16)

System
(64 cabinets, 64x32x32)

Node Board
(32 chips, 4x4x2)
16 Compute Cards

Compute Card
(2 chips, 2x1x1)

Chip
(2 processors)

2.8 Gflops peak
per processor core

2.8/5.6 GF/s
4 MB

5.6/11.2 GF/s
0.5 GB DDR

90/180 GF/s
8 GB DDR

2.9/5.7 TF/s
256 GB DDR

180/360 TF/s
16 TB DDR

Design Goals:

- High computational power efficiency
- High computational density per volume

LINPACK Performance:

280,600 GFLOPS = 280.6 TeraFLOPS = 0.2806 Peta FLOP

Top Peak FP Performance:

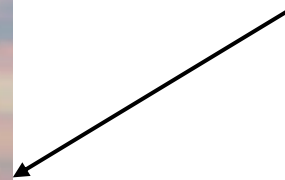
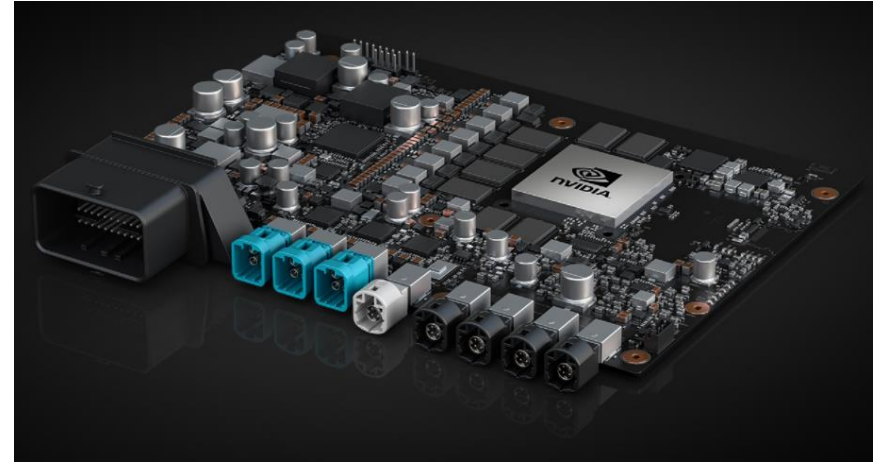
Now about 367,000 GFLOPS = 367 TeraFLOPS = 0.367 Peta FLOP

Embedded Computers

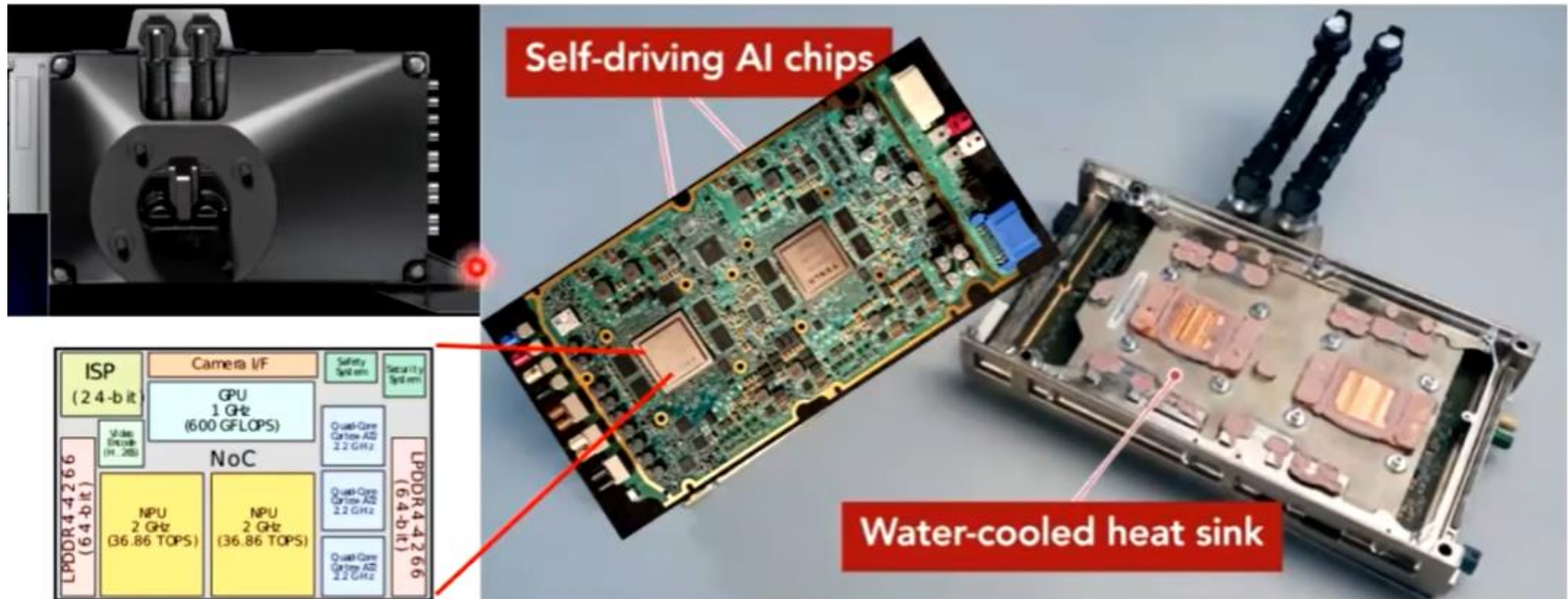
- Embedded computers
 - Hidden as components of systems
 - For running one predetermined application or collection of software.
 - Stringent power/performance/cost constraints
- Despite the large number of embedded computers, most users never really see that they are using a computer!

NVIDIA DRIVE AGX Xavier

- World's first AI computer for Autonomous Machines
 - 512-core Volta GPU
 - Deep Learning Accelerator (DLA)
 - 8-core ARM CPU
 - Vision Accelerator (DLA)
 - 30 TOPS
- Embedded system in Xpeng P7



Tesla FSD Self-Driving Chip



Full Self-Driving Chip: 3 quad-core Cortex-A72, 2.2GHz, Mali G71 MP12 GPU operating 1 GHz, 2 Neural Processing Units operating at 2 GHz, 14 nm, 2300 frames/s, NPU:36.86 trillion operations per second

Each chip makes its own assessment of what the car should do next.

The computer compares the two assessments, and if the chips agree, the car takes the action. If the chips disagree, the car just throws away that frame of video data and tries again

Personal Mobile Device

- Personal Mobile Device (PMD)
 - Battery operated
 - Connects to the Internet
 - Hundreds of dollars
 - Smart phones, tablets, electronic glasses



Cloud Computing

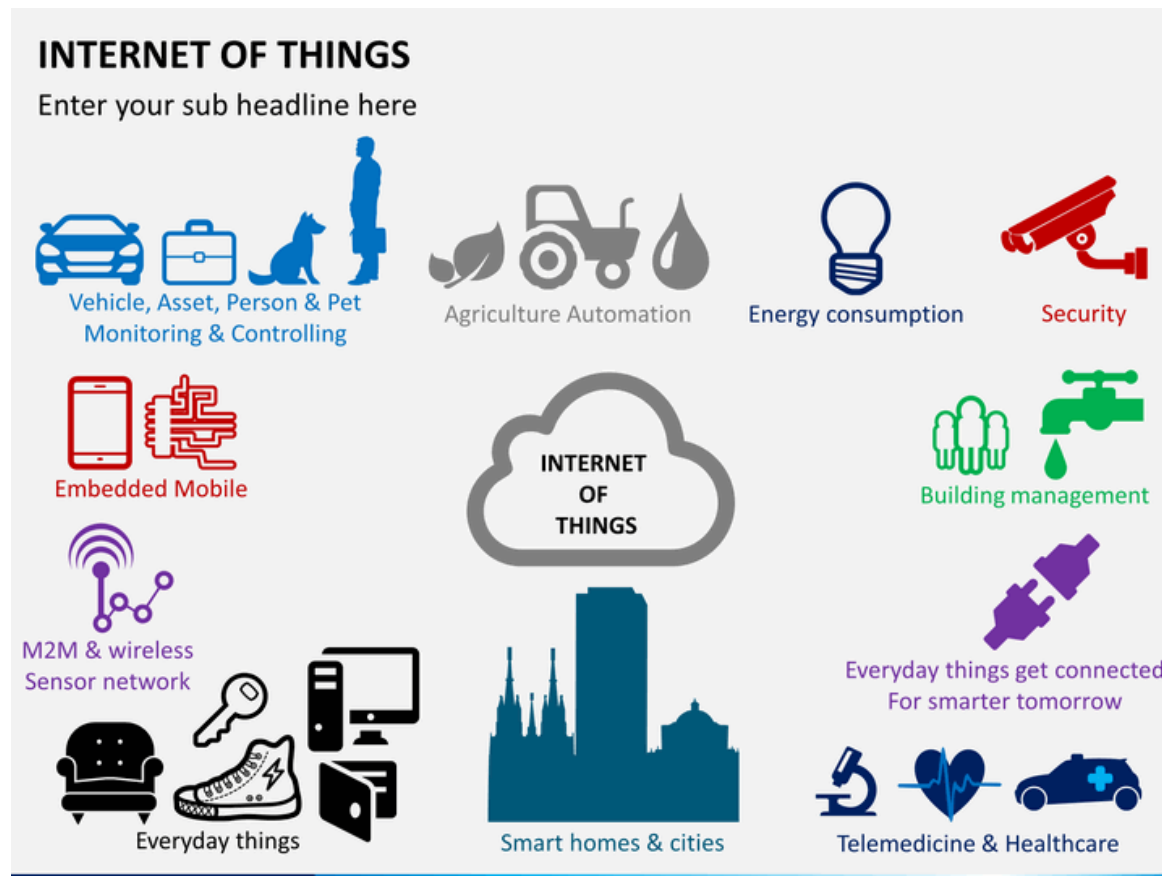
- Cloud computing
 - Warehouse Scale Computers (WSC)
 - Amazon and Google Data centers
 - Millions of computers connected by off-the-shelf networking devices
 - Software as a Service (SaaS)
 - Delivers software and data as a service over the Internet
 - Portion of software run on a Personal Mobile Device and a portion run in the Cloud



Google Data centers

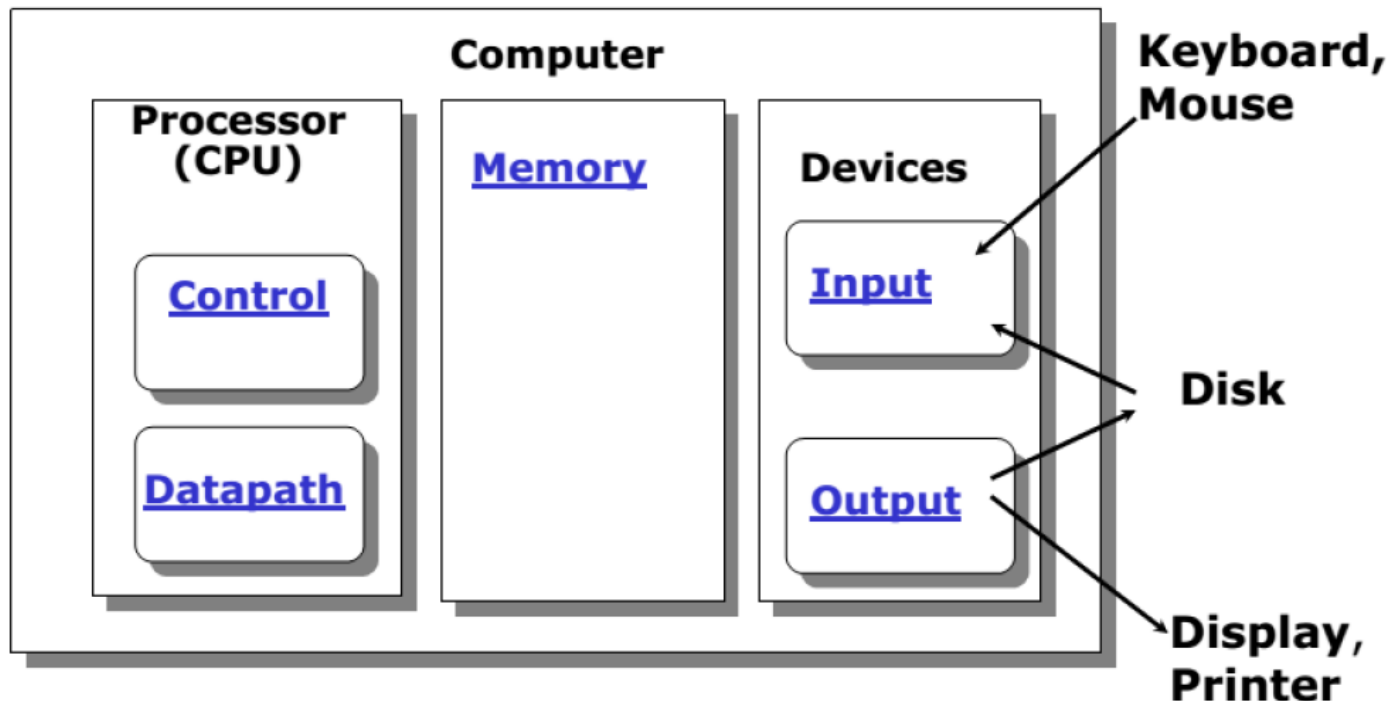
Internet of Things (IoT)

- A network of physical objects, that use sensors and Application Program Interfaces (APIs) to connect and exchange data over the Internet



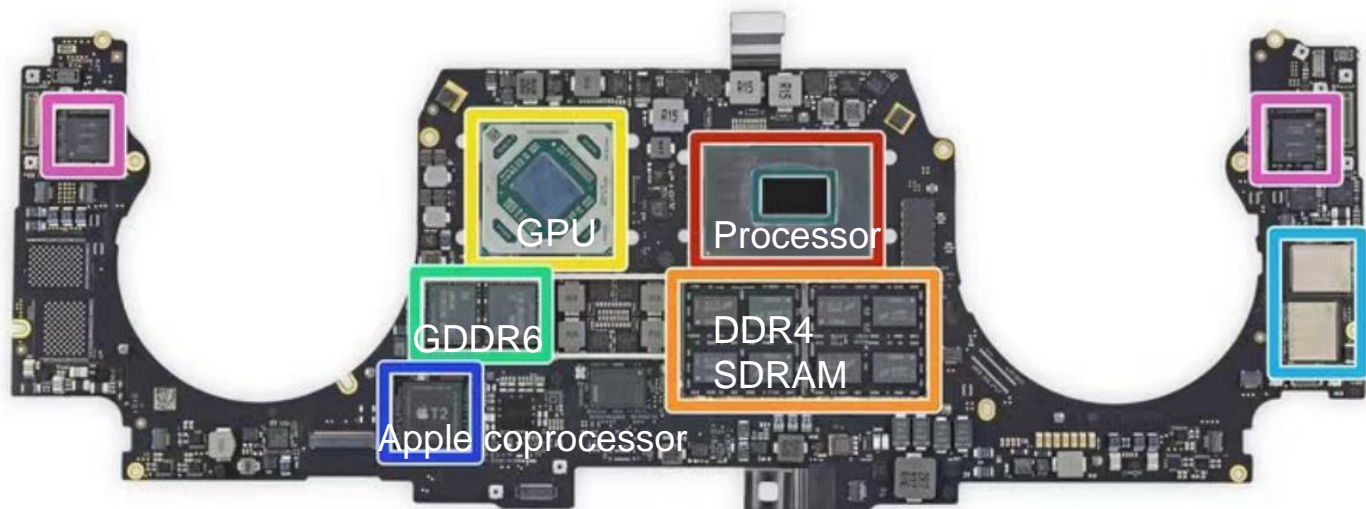
The Hardware of a Computer

- 5 classical computer components
- The processor gets instructions and data from memory. Input writes data to memory, and output reads data from memory.
Control sends the signals that determine the operations of the **datapath**, **memory**, **input**, and **output**.



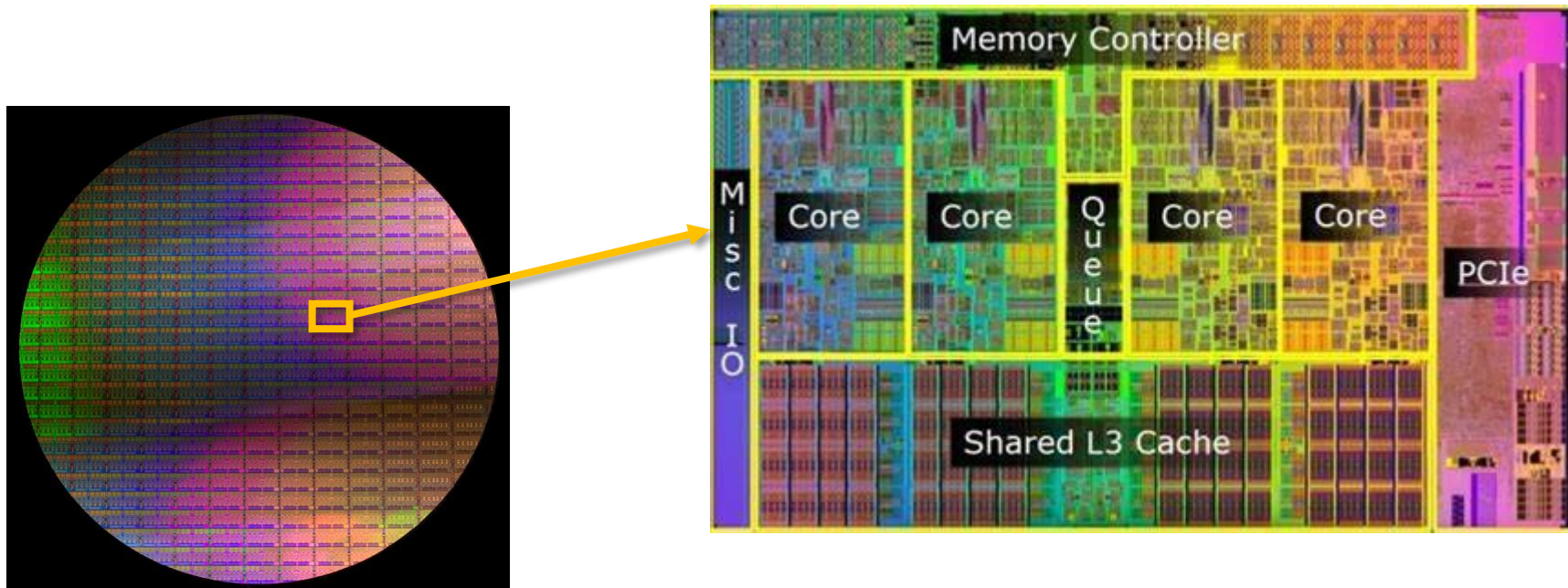
Teardown of MacBook

- 16" LED-backlit IPS Retina display
- Keyboard and Touch Bar
- 2.6 GHz 6-core Intel Core i7
- 16 GB of 2666 MHz DDR4 SDRAM
- 512 GB SSD
- 100 Watt-hour battery
- Speaker and microphone



Inside the Processors

- Datapath:
 - performs operations on data
- Control:
 - sequences datapath, memory, I/O
- Cache memory:
 - small fast Static RAM memory for immediate access to data

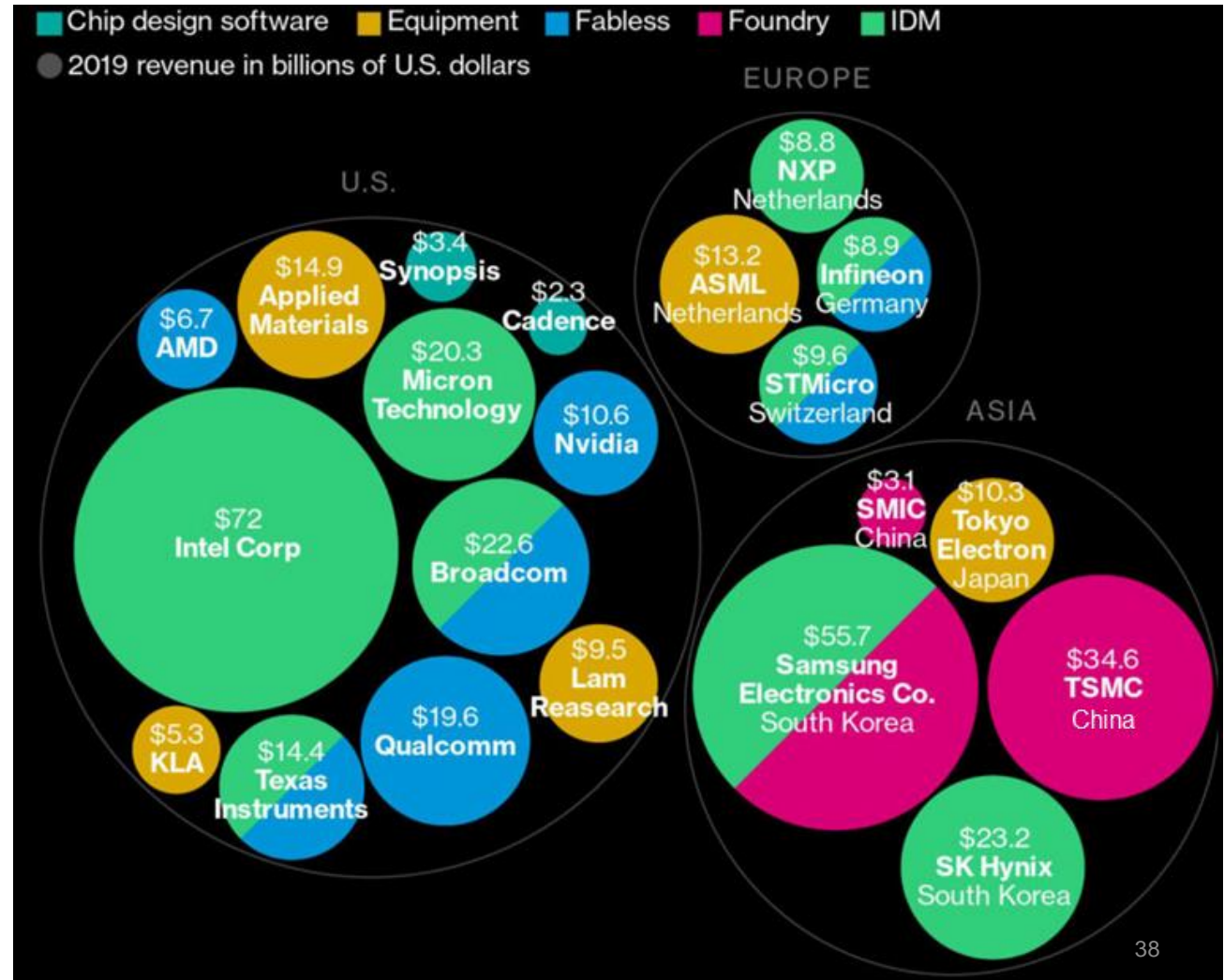


Intel's Core i7 wafer and Die map

Chip Industry Choke Points

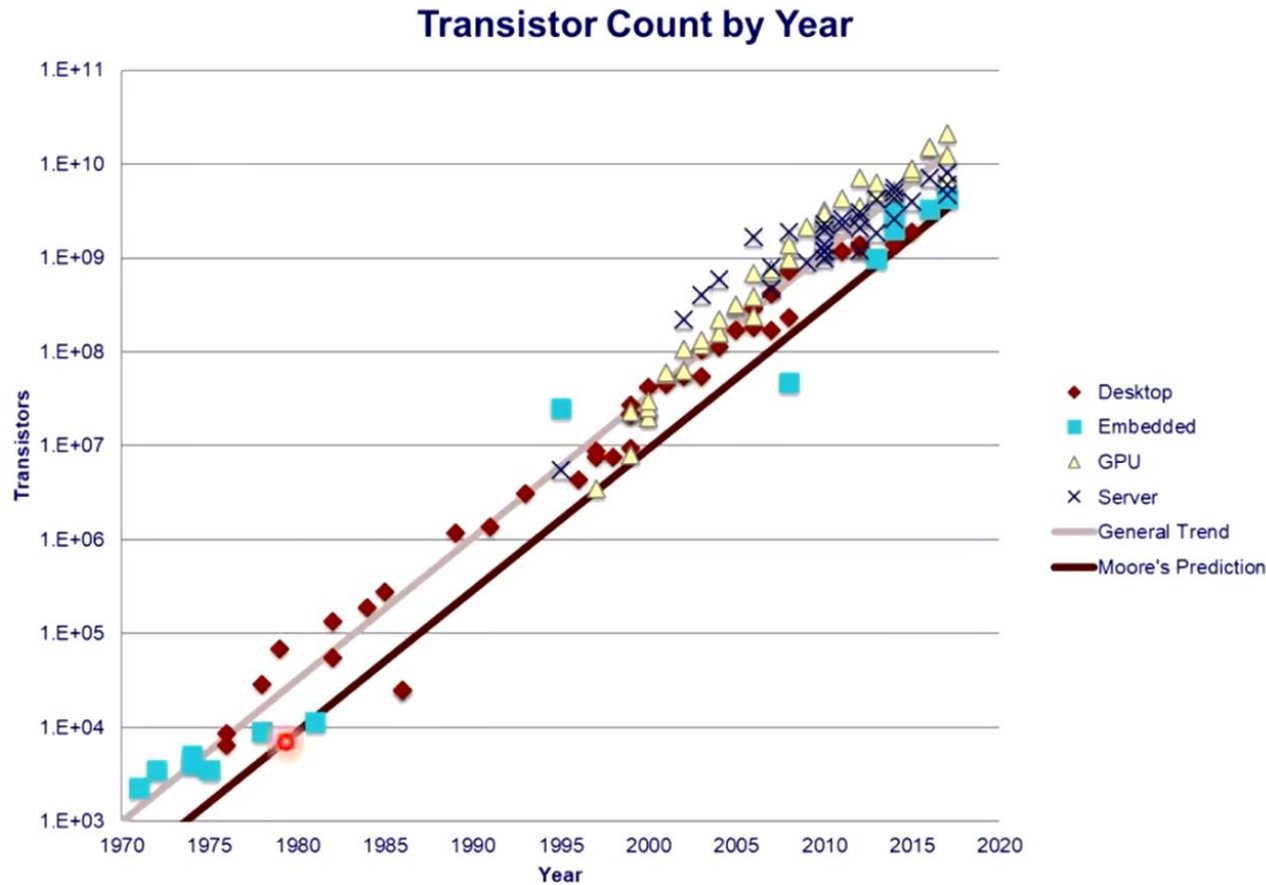
- Key players in chip industry

- Intel
- AMD
- Qualcomm
- Samsung
- TSMC
- Broadcom
- Nvidia
- ASML
- ...



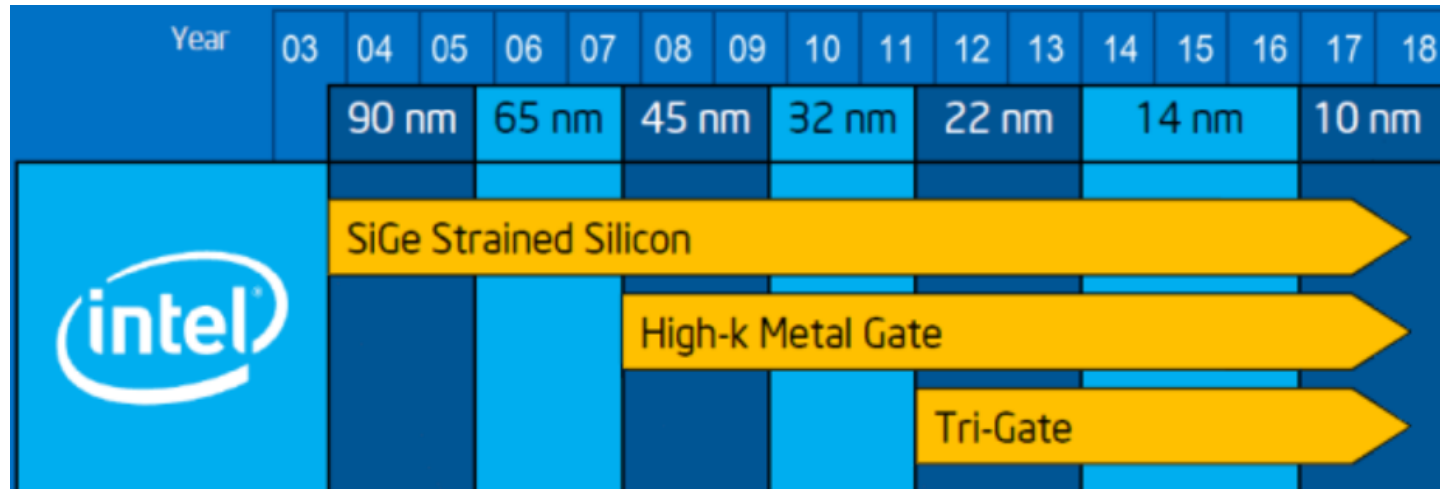
Moore's law

- **Prediction** made by Gordon Moore(1929-) in 1965 that the number of transistors per silicon chip doubles every 18 to 24 months.
- Is moore's law still valid?



Processor Technology Trends

- Shrinking of transistor sizes: 90nm(2004) -> 45nm(2008) -> 22nm(2012) -> 10nm(2017)



- Transistor density increases by 35% per year and die size increases by 10-20% per year... functionality improvements!
- Transistor speed improves linearly with size (complex equation involving voltages, resistances, capacitances)
- Wire delays do not scale down at the same rate as transistor delays

Eight Great Ideas

- Two Design rules
 - Design for **Moore's Law**
 - Use **abstraction** to simplify design
- Four ways to improve performance
 - Make the **common case fast**
 - Performance via **parallelism**
 - Performance via **pipelining**
 - Performance via **prediction**
- **Hierarchy** of memories
- **Dependability** via redundancy

Two Design rules

- Design for Moore's Law
 - Computer architects must anticipate where the technology will be when the design finishes rather than design for where it starts
- Abstraction
 - Increases productivity for hardware and software
 - Lower-level details are hidden to offer a simpler model at higher levels



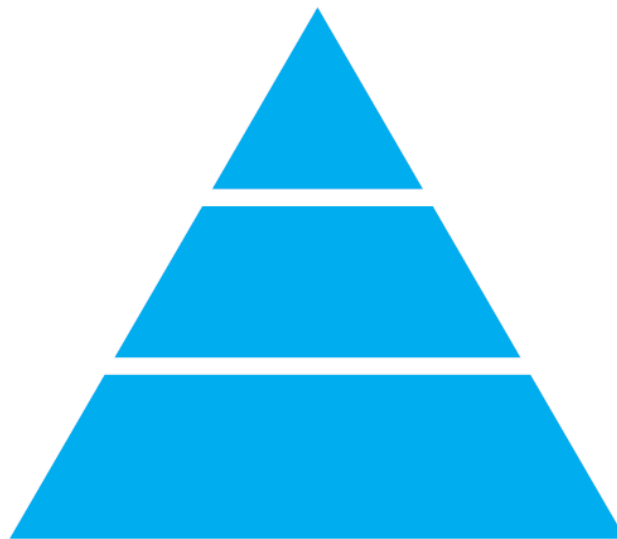
Four ways to improve performance

- Common Case Fast:
 - Enhances performance better than optimizing the rare case
 - the common case is often simpler than rare case and, hence, it is easier to enhance.
- Performance via Parallelism:
 - Running multiple operations in parallel enhances the computer performance
- Performance via Pipelining:
 - A particular pattern of parallelism
 - A set of data processing elements connected in series, so that the output of one element is the input of the next one
- Performance via Prediction
 - It can be faster on average to guess and start working rather than wait



Hierarchy of memories

- The fastest, smallest, and most expensive memory per bit at the top of the hierarchy;
- The slowest, largest, and cheapest per bit at the bottom.
 - Cache(SRAM)
 - Main memory(DRAM)
 - Secondary Storage: e.g. Hard disk



H I E R A R C H Y

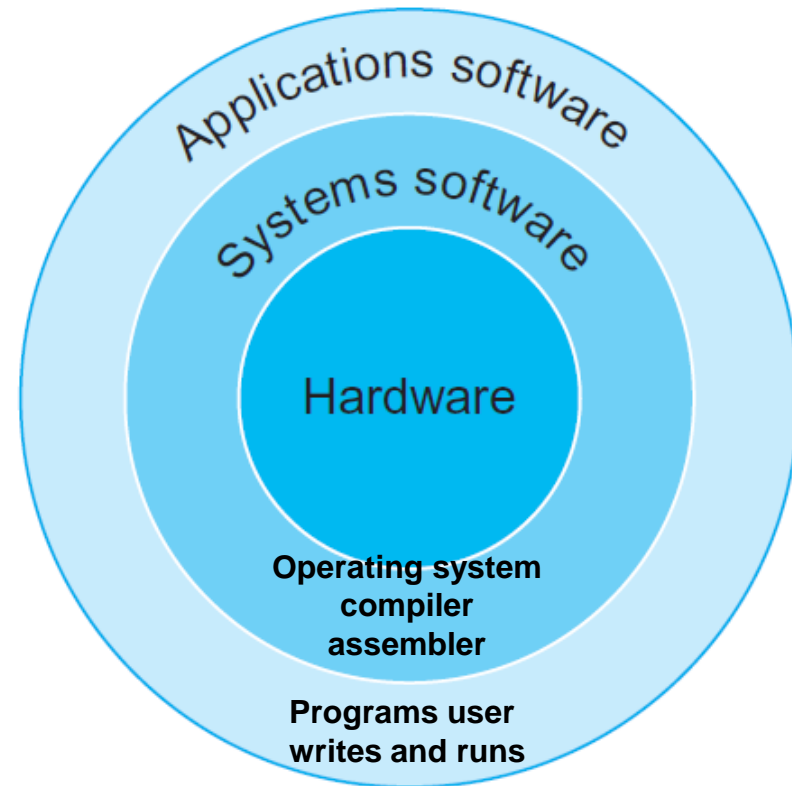
Dependability via Redundancy

- Any physical device, including a computer, can fail:
- we make systems dependable by including the redundant components that can take over when a failure occurs and help detect failures.



The Concept of a Computer

- Application software
 - Written in high-level language
- System software
 - Compiler: translates HLL code to machine code
- Operating System:
 - Handling input/output
 - Managing memory and storage
 - Scheduling tasks & sharing resources
- Hardware
 - Processor, memory, I/O controllers



Levels of Program Code

- C program compiled into assembly language and then assembled into binary machine language.
- High-level language
 - Level of abstraction closer to problem domain
 - Provides for productivity and portability
- Assembly language
 - Textual representation of instructions
- Machine language
 - Hardware representation
 - Binary digits (bits)
 - Encoded instructions and data

High-level
language
program
(in C)

```
swap(int v[], int k)
{int temp;
  temp = v[k];
  v[k] = v[k+1];
  v[k+1] = temp;
}
```

Compiler

Assembly
language
program
(for MIPS)

```
swap:
  multi $2, $5, 4
  add   $2, $4, $2
  lw    $15, 0($2)
  lw    $16, 4($2)
  sw    $16, 0($2)
  sw    $15, 4($2)
  jr    $31
```

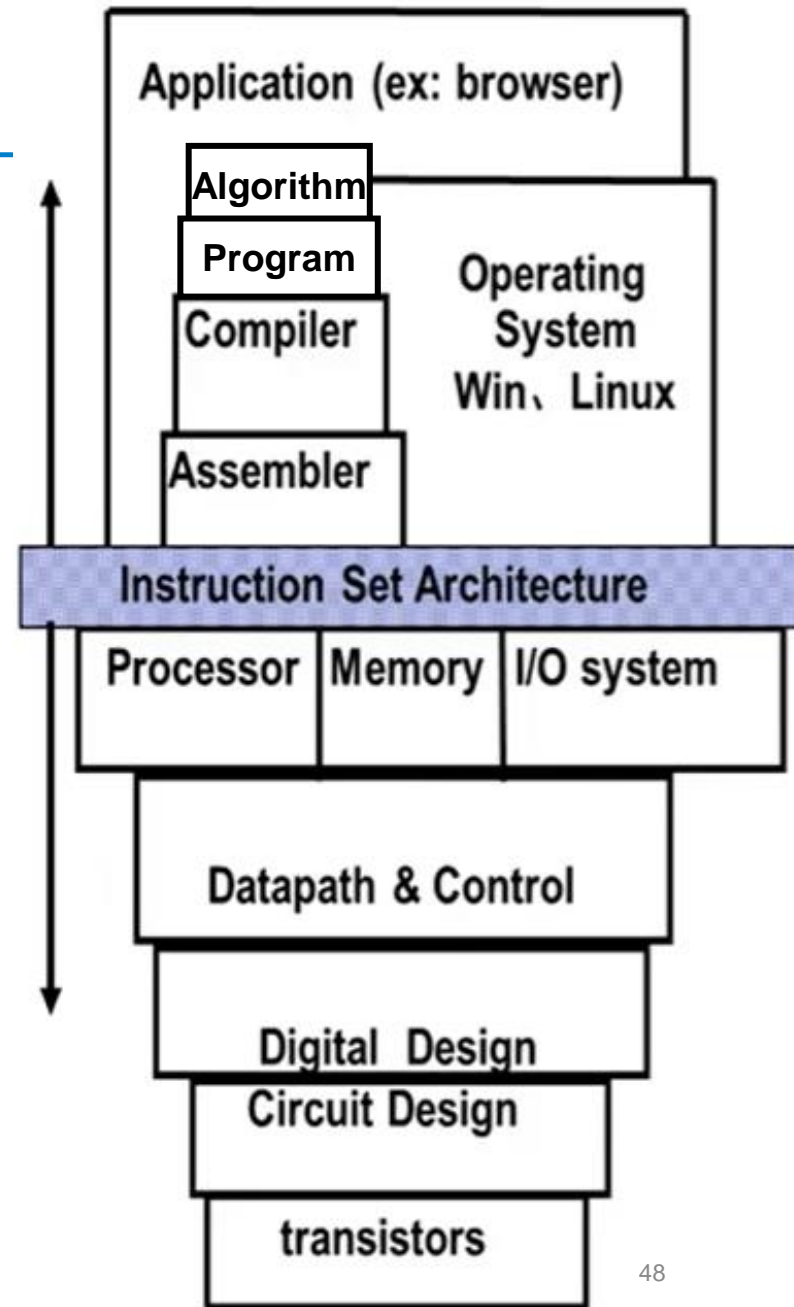
Assembler

Binary machine
language
program
(for MIPS)

```
000000001010001000000000100011000
000000001000001000010000000100001
100011011110001000000000000000000
1000111000010010000000000000000100
101011100001001000000000000000000
101011011110001000000000000000000
000000111110000000000000000001000
```

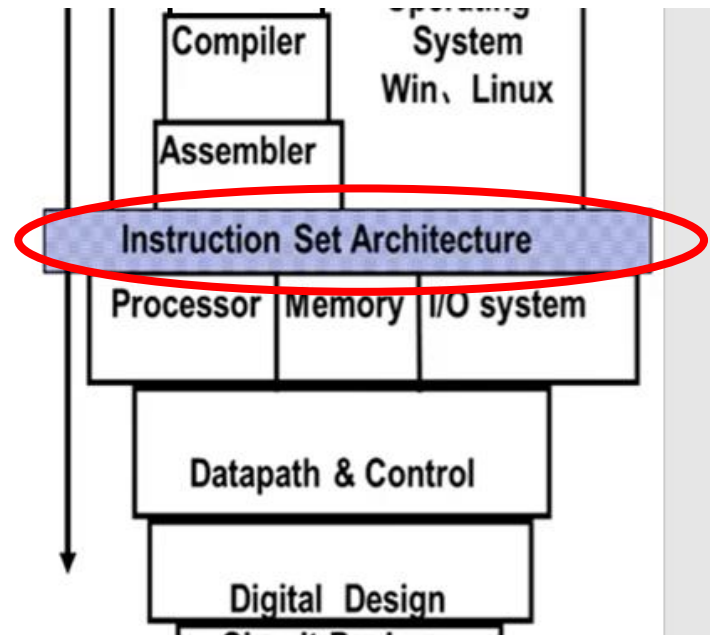
Abstractions

- Abstraction helps us deal with complexity
 - Hides lower-level details
- Instruction Set Architecture (ISA) or Computer Architecture
 - The hardware/software interface
 - Includes instructions, registers, memory access, I/O, and so on
- Operating system hides details of doing I/O, allocating memory from programmers



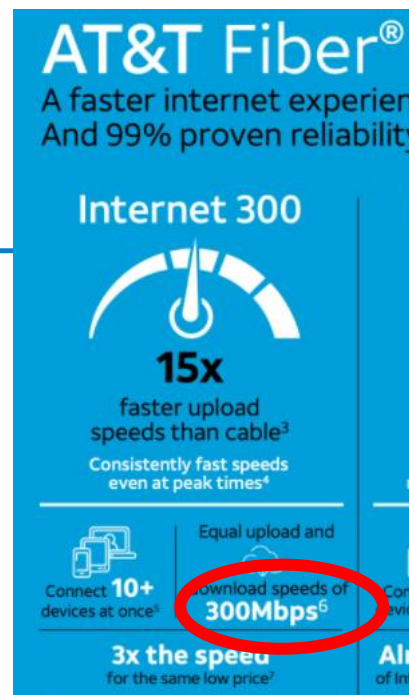
Instruction Set Architecture (ISA)

- A set of assembly language instructions (ISA) provides a link between software and hardware.
- Given an instruction set, software programmers and hardware engineers work more or less independently.
- Common types of ISA: RISC, CISC
- Examples:
 - IBM370/X86 (CISC)
 - MIPS (RISC)
 - ARM (RISC)



Value notions

- 1B(Byte) = 8b(bit)
- Byte: storage, 256GB
- bit: transmission rate, 300Mbps



AT&T Fiber®
A faster internet experience
And 99% proven reliability

Internet 300

15x
faster upload speeds than cable³
Consistently fast speeds even at peak times⁴

Equal upload and download speeds of 300Mbps⁵

Connect 10+ devices at once⁵

3x the speed for the same low price⁷



Decimal term	Abbreviation	Value	Binary term	Abbreviation	Value
kilobyte	KB	10^3	kibibyte	KiB	2^{10}
megabyte	MB	10^6	mebibyte	MiB	2^{20}
gigabyte	GB	10^9	gibibyte	GiB	2^{30}
terabyte	TB	10^{12}	tebibyte	TiB	2^{40}
petabyte	PB	10^{15}	pebibyte	PiB	2^{50}
exabyte	EB	10^{18}	exbibyte	EiB	2^{60}
zettabyte	ZB	10^{21}	zebibyte	ZiB	2^{70}
yottabyte	YB	10^{24}	yobibyte	YiB	2^{80}