COMPUTER ARCHITETURE

ASSOCIATE PROFESSOR VIORICA SUDACEVSCHI

Couse structure

lectures- 30 hours

seminars / laboratory works - 45 hours

Grading

Semester grade - 60 %

- Periodic evaluation1 – 15% (test)
- Periodic evaluation2 – 15% (test)
- Current evaluation1 – 15% (quizzes, laboratory works)
- Individual work – 15% (thematic reports)

Final evaluation - 40% (exam)

INTRODUCTION

A computer consists of a set of physical components:

- Hardware;
- Software: system programs (system software) that are responsible for data processing according to an algorithm, specified by the user through an application program (application software).

Computer systems have conventionally been defined through their interfaces at a number of abstraction levels, each providing functional support to its predecessor.

Included among the levels are:

• High-level programming language	Level 5
• Assembly language	Level 4
• Operating system	Level 3
• Machine instructions	Level 2
• Micro architecture	Level 1
• Digital logic	Level 0

Computer architecture is a set of rules and methods that describe the functionality, organization, and implementation of computer systems.

COMPUTER'S GENERATIONS

Modern electronic computers are typically grouped into four "generations." Each generation is marked by improvements in basic technology. Each advance has resulted in computers of lower cost, higher speed, greater memory capacity, smaller size and power consumption.

1. **First Generation (1945–1954)** based on **vacuum tube** invented in 1906 by an electrical engineer named Lee De Forest.

General-purpose computers:

ENIAC (Electronic Numerical Integrator and Computer)- 18,000 vacuum tubes, 30.5 meters, 10-digit registers for temporary calculations;

Colossus - 1,500 vacuum tubes,

UNIVAC - 5,000 vacuum tubes.

These early machines were typically controlled by plug board wiring.

ENIAC

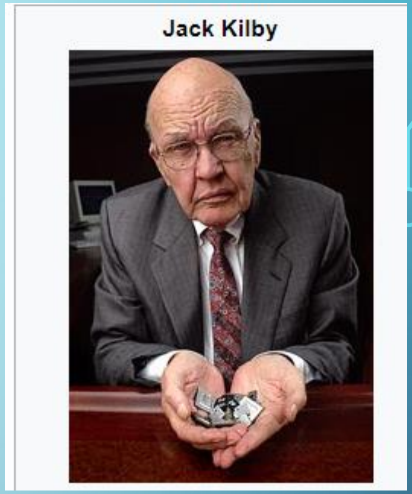
UNIVAC

- **Second Generation (1955–1964)** based on **transistors** invented in the mid-1940s by John Bardeen (1908–1991), William B. Shockley (1910–1989), and Walter H. Brattain (1902–1987).
- In this period appears and first supercomputers: UNIVAC LARC - Livermore Atomic Research Computer and IBM 7030 - named Strech Computer), used for weather prediction, nuclear research and artificial intelligence.
- These second generation machines were programmed in languages such as COBOL (Common Business Oriented Language) and FORTRAN (Formula Translator).
- Magnetic disks and tape were often used for data storage. Appears the concept of parallel processing.

UNIVAC LARC

IBM-7030

3. **Third Generation (1965–1978)** based on **integrated circuits** invented by Jack Kilby and Robert Noyce.



Computers:

IBM System/360 - was able to execute 500,000 additions per second. This computer was about 263 times as fast as the ENIAC.

During the third generation of computers, the central processor was constructed by using many integrated circuits.

It introduced single computer architecture over a range or family of devices. In other words, a program designed to run on one machine in the family could also run on all of the others. IBM spent approximately $5 billion to develop the System/360.

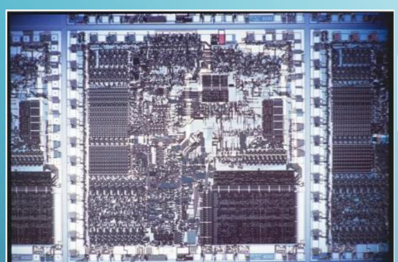
Appears first minicomputers.

The important characteristics of the computers of this generation: operating systems, multiprogramming, multiprocessing and virtual memory.

IBM SYSTEM/360

4. **Fourth Generation (1979–?)** based on the **microprocessors.** Microprocessors used Large Scale Integration (LSI) and Very Large Scale Integration (VLSI) techniques to pack thousands or millions of transistors on a single chip. Advantages: speed, high integration ratio, high reliability, small costs and dimensions.

The Datamaster was an all-in-one computer with text-mode CRT display, keyboard, processor, memory, and two 8-inch floppy disk drives all contained in one cabinet. (Photo: Oldcomputers.net)



Die shot of the 16-bit Intel 8086 microprocessor. The 8086 gave rise to the famed x86 architecture which eventually turned out as Intel's most successful line of processors.

Paul Allen and Bill Gates pose next to a few early desktop systems.

Mobile phone, laptop, printer, camera and tablet

The Summit is the first computer ever to reach mind-boggling exascale speeds, referred to as exaflops, or a billion billion calculations per second.

It is predicted that 175 zettabytes of data will be produced by 2025, which will make data centers to continue to play a vital role in the ingestion, computation, storage, and management of information.

1000^7	ZB	zettabyte	1024^7	ZiB	zebibyte

First microprocessor: Intel Company, I4004 – 4 bits organization (built in 1971) was the first processor to be built on a single silicon chip. It contained 2,300 transistors.

First successful microprocessor: Intel I8080 – 8 bits processor (1972).

First 16 bits processor: Intel I8086 (1978).

First 32 bit processor: Intel I80386 (1985).

Superscalar microprocessor architecture: Pentium Pro (1990)

64 bits processors, multi-core architectures:

ATOM, single, dual-core, quad-core, 8-, 12-, and 16-core processors for netbooks, embedded applications, and mobile internet devices (MIDs).

XEON Phi 57-, 60-, 61-, 64-, 68-, and 72-core processors

Other microprocessor families:

Motorola: 6800 (8 biti), 68000 (16 biti), 68020, 68030 (32 biti), 68040,

Zilog: Z80, Z8000

Texas Instruments: - digital signal processors: TMS320c10/20/30/50/80

Microchip: microcontrollers: PIC8/16/32

MIPS (Microprocessor without Interlocked Pipeline Stages) ,

ARM (*Advanced RISC Machine*), etc.

- **Multithreading** is the ability of a Central Processing Unit (CPU) (or a single core in a multicore processor) to provide multiple threads of execution concurrently, supported by the operating system.
- A **multi-core processor** is a computer processor on a singleintegrated circuit with two or more separate processing units, called cores, each of which reads and executes program instrauctions.
- A **superscalar processor** is a CPU that implements a form of parallelism called instruction level parallelism within a single processor. In contrast to a scalar processor that can execute at most one single instruction per clock cycle, a superscalar processor can execute more than one instruction during a clock cycle by simultaneously dispatching multiple instructions to different execution units.

TENDENCIES AND PERSPECTIVES

- Increase of integration ration - smaller switching elements (transistors): 45->35nm, increase of switching elements' number, processors - over 1 billion transistors, memory – over 64-512 billion;
- Power reduction - intelligent power distribution, dynamic power control: energy where and when it is needed, frequency limitation;
- Multi-core and multi-thread architectures (from 2 cores/chip to 128 cores and more, symmetric and asymmetric architectures
- Network-on-chip - network communication inside the chip instead of parallel buses;

TENDENCIES AND PERSPECTIVES

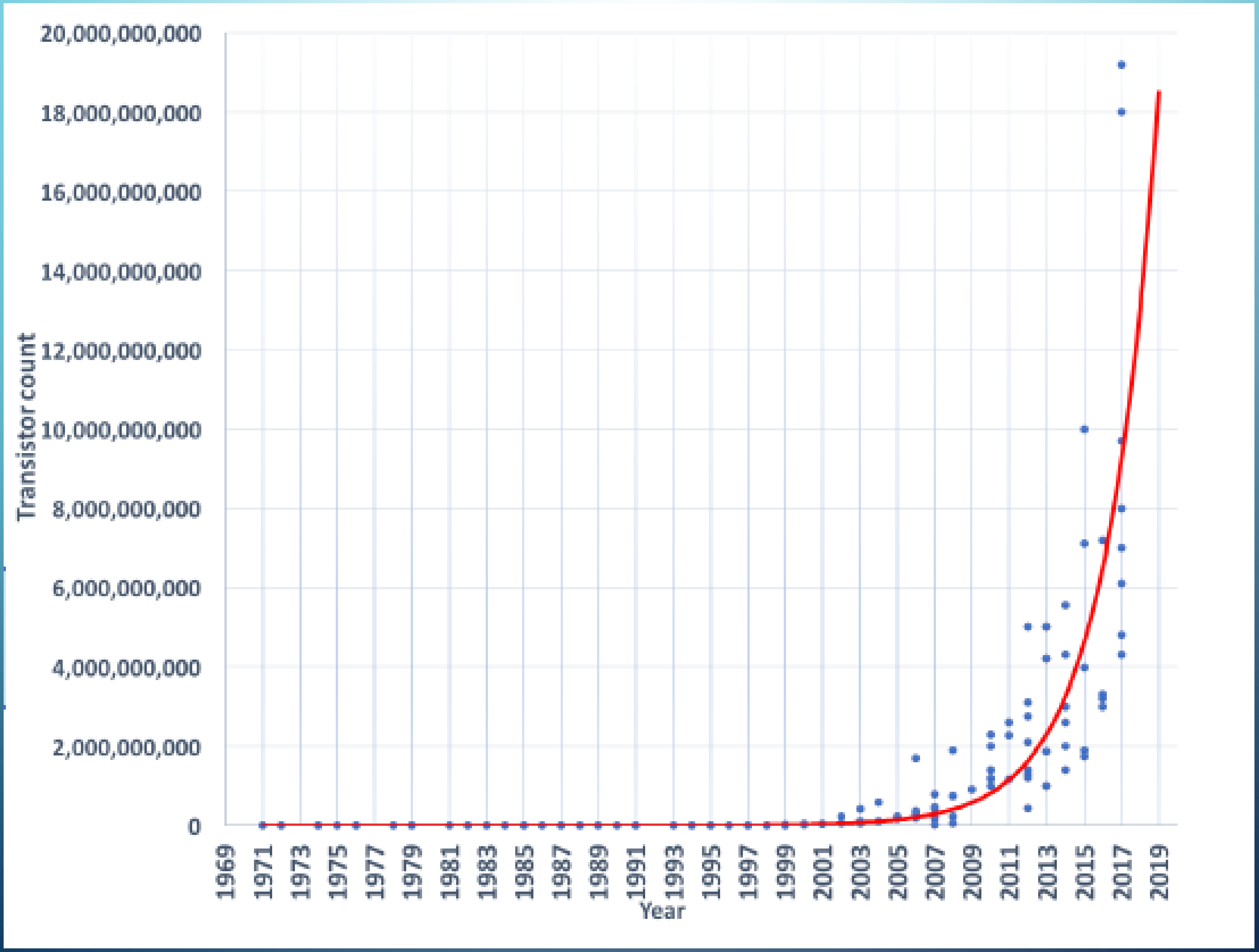
- Memory hierarchies - more cache memory levels (inside the processor), virtual memory, access request anticipation;
- External memories of silicon - no more hard and floppy disks of DVDs, flash instead;
- Multi-processor architectures - parallel architectures, distributed architectures;
- Computer networks - Internet – an indispensable computer resource, wireless networks;
- Mobile and portable computers: laptops, graphic tablets, PDA (personal digital assistant);
- GPS (Global Positioning System), intelligent phones.

TECHNOLOGICAL DEVELOPMENT

Numbers of Devices per Chip

Integration Technology	Typical number of devices	Typical functions
SSI Bipolar	10–20	Gates and flip-flops
MSI Bipolar & MOS	50–100	Adders & counters
LSI Bipolar & MOS	100–10,000	ROM & RAM
VLSI CMOS (mostly)	10,000–5,000,000	Processors
WSI CMOS	5,000,000	DSP & special purposes

A common law that governs the world of microprocessors is **Moore's Law**. Moore's Law states that the numbers of transistors on a single chip at the same price will double every 18 to 24 months. Current microprocessor chips contain millions of transistors and the number is growing rapidly.

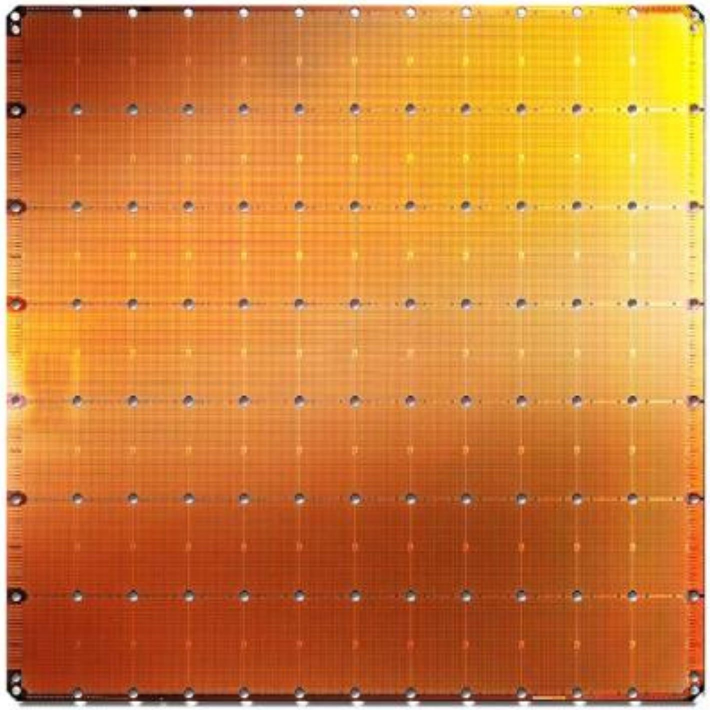


- New Computing Paradigms (Rethinking the Full Stack)
  - Processing in Memory, Processing Near Data
  - Neuromorphic Computing
  - Fundamentally Secure and Dependable Computers

- New Accelerators (Algorithm-Hardware Co-Designs)
  - Artificial Intelligence & Machine Learning
  - Graph Analytics
  - Genome Analysis

- New Memories and Storage Systems
  - Non-Volatile Main Memory
  - Intelligent Memory

Cerebras's Wafer Scale Engine (2019)



- The largest ML accelerator chip
- 400,000 cores

Cerebras WSE
1.2 Trillion transistors
46,225 mm^2

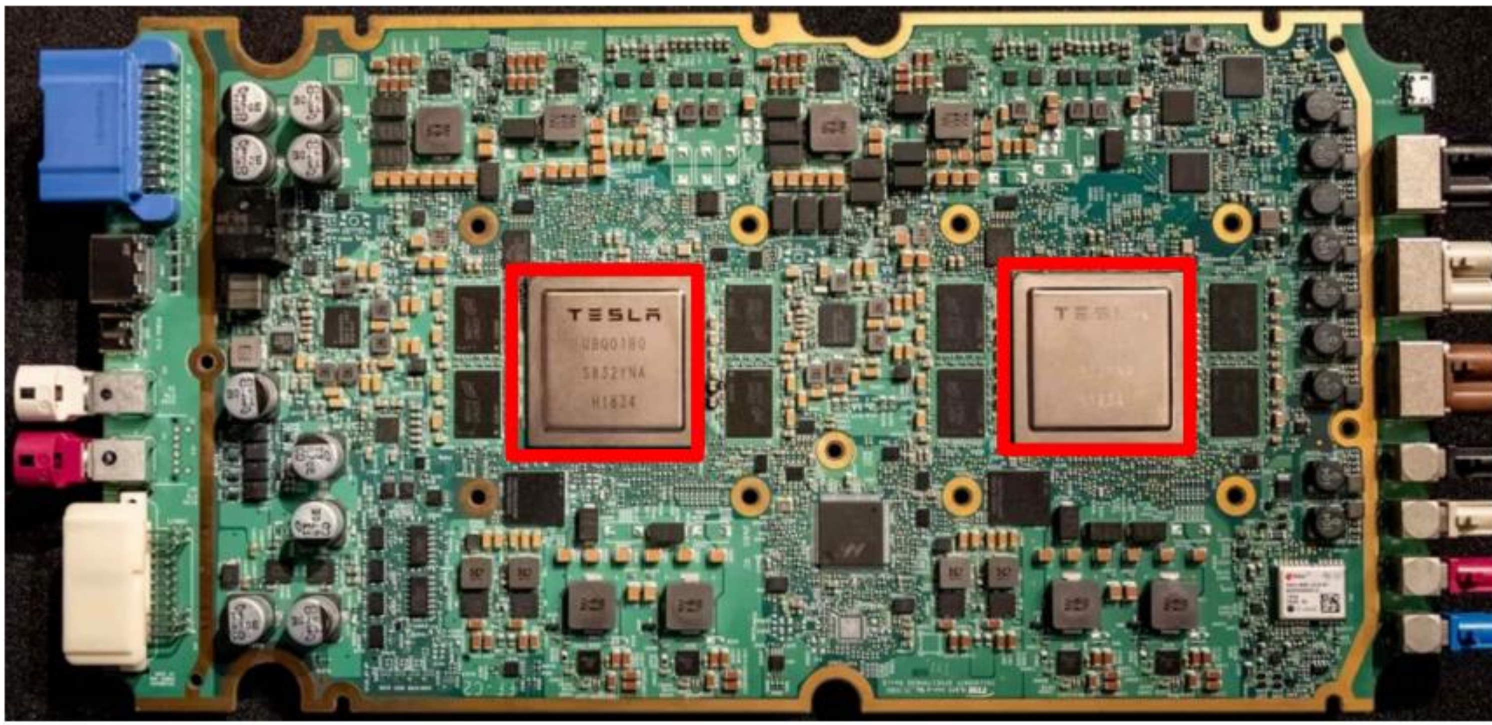
Largest GPU
21.1 Billion transistors
815 mm^2
NVIDIA TITAN V

https://www.anandtech.com/show/14758/hot-chips-31-live-blogs-cerebras-wafer-scale-deep-learning

https://www.cerebras.net/cerebras-wafer-scale-engine-why-we-need-big-chips-for-deep-learning/

TESLA Full Self-Driving Computer (2019)

- ML accelerator: 260 mm^2, 6 billion transistors, 600 GFLOPS GPU, 12 ARM 2.2 GHz CPUs.
- Two redundant chips for better safety.



In computing, **floating point operations per second** (**FLOPS**, **flops** or **flop/s**) is a measure of computer performance, useful in fields of scientific computations that require floating-point calculations.

Tensor Processing Units (TPUs) are Google's custom-developed application-specific integrated circuits (ASICs) used to accelerate machine learning workloads. These TPUs are designed from the ground up with the benefit of Google's deep experience and leadership in machine learning.

Google TPU Generation I (~2016)

Figure 3. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

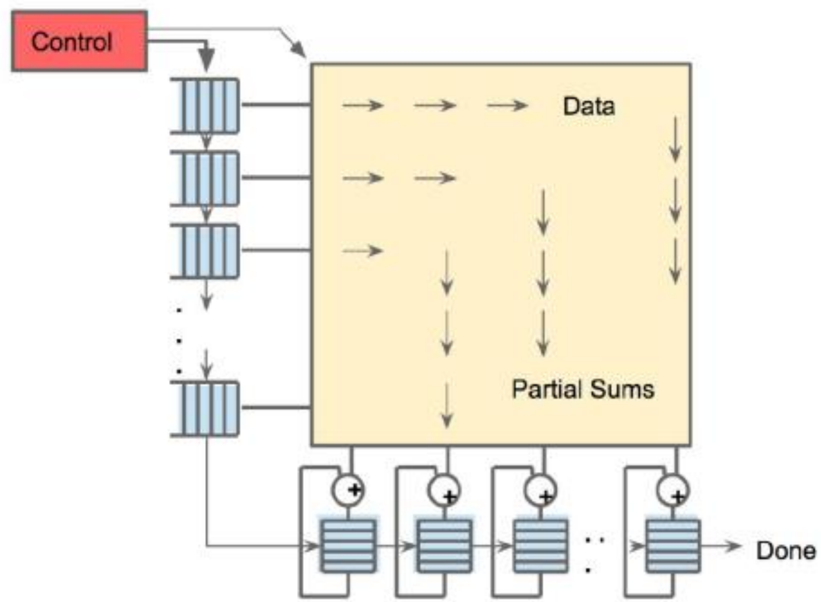


Figure 4. Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

Jouppi et al., “In-Datacenter Performance Analysis of a Tensor Processing Unit”, ISCA 2017.

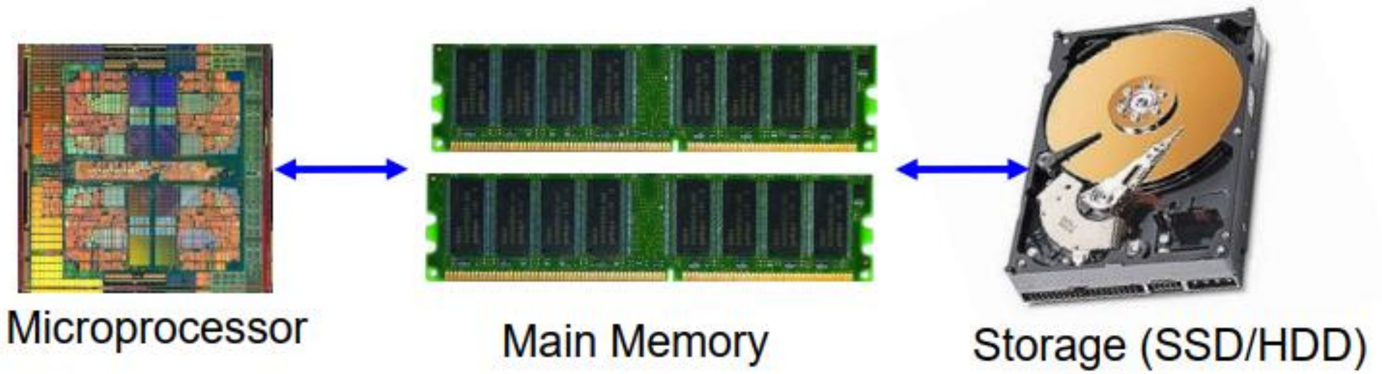
Many (Other) AI/ML Chips

- Alibaba
- Amazon
- Facebook
- Google
- Huawei
- Intel
- Microsoft
- NVIDIA
- Tesla
- Many Others and Many Startups…

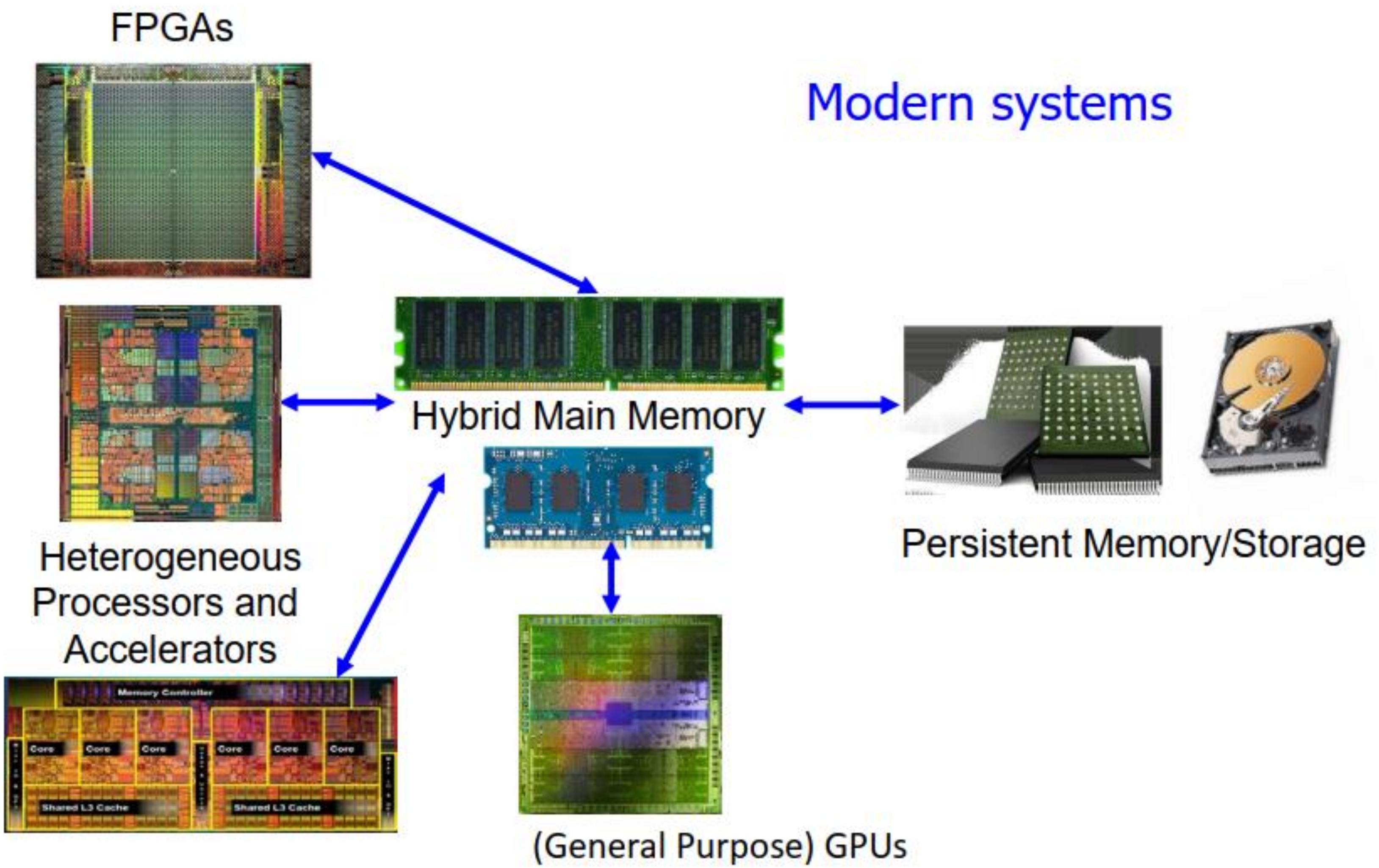
- In June 2019, Summit, an IBM-built supercomputer now running at the Department of Energy's (DOE) Oak Ridge National Laboratory (ORNL), captured the number one spot with a performance of 148.6 petaFLOPS on High Performance Linpack (HPL), the benchmark used to rank the TOP500 list. Summit has 4,356 nodes, each one equipped with two 22-core Power9 CPUs, and six NVIDIA Tesla V100 GPUs.
- In June 2020, Fugaku turned in a High Performance Linpack (HPL) result of **415.5 petaFLOPS**, besting the now second-place Summit system by a factor of 2.8x. Fugaku is powered by Fujitsu’s 48-core A64FX SoC, becoming the first number one system on the list to be powered by ARM processors. In single or further reduced precision, used in machine learning and AI applications, Fugaku’s peak performance is over 1,000 petaflops (1 exaflops). The new system is installed at RIKEN Center for Computational Science (R-CCS) in Kobe, Japan.

Increasingly Complex Systems

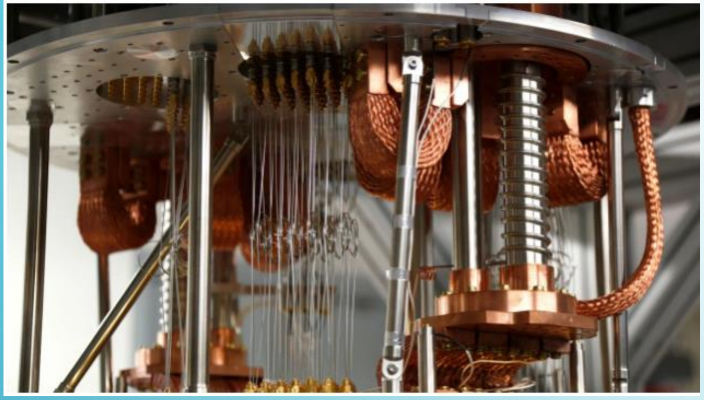
Past systems



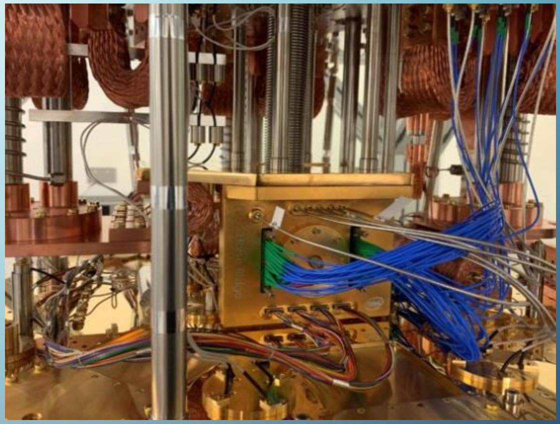
Increasingly Complex Systems



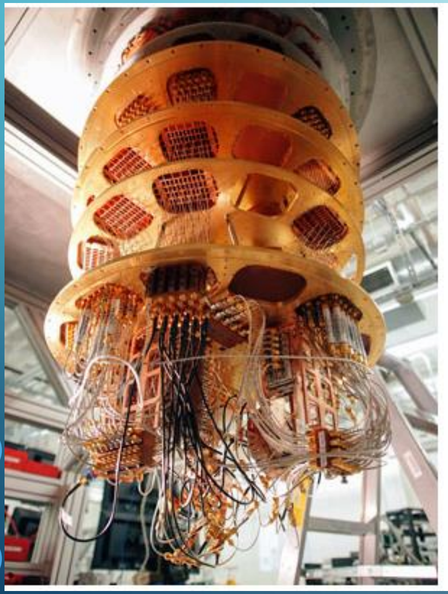
QUANTUM COMPUTERS



IBMs 53 quantum bits (qubits) quantum computer is available for researchers and companies to run applications via the cloud.



Intel Horse Ridge II -49 qubit quantum computer

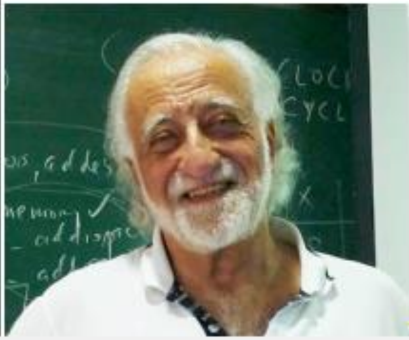


Googles 72 qubits quantum computer.

Role of the (Computer) Architect

Role of the Architect

- **_-- Look Backward (Examine old code)_**
- **_-- Look forward (Listen to the dreamers)_**
- **_-- Look Up (Nature of the problems)_**
- **_-- Look Down (Predict the future of technology)_**



from Yale Patt's lecture notes

COURSE MAIN TOPICS

1. Arithmetic basics
2. Logic Basics
3. Microprocessors
4. Memory
5. I/O devices
6. Assembly language basics