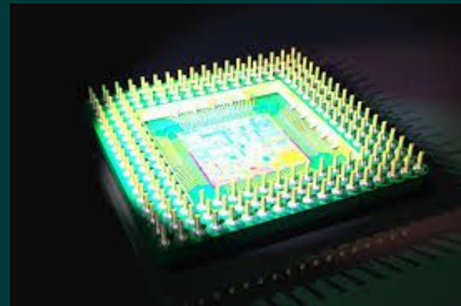



Computer Organization and Design

Introduction and Chapter 1

(Sections 1.1 – 1.5)

Computer Abstractions and Technology

Definitions & Distinctions

- Computer architecture and computer organization are related but distinct concepts in the field of computer science.
- Computer architecture refers to the design of the internal workings of a computer system, including the CPU, memory, and other hardware components.
  - It involves decisions about the organization of the hardware, such as the instruction set architecture, the data path design, and the control unit design.
  - Computer architecture is concerned with optimizing the performance of a computer system and ensuring that it can execute instructions quickly and efficiently.

Definitions & Distinctions

- Computer organization refers to the way in which the hardware components of a computer system are arranged and interconnected.
  - It defines the operational units and their interconnections that implement the architecture specification.
  - It deals with how the components of a computer system are arranged and how they interact to perform the required operations.
  - Computer organization is concerned with the physical implementation of the architecture design and includes decisions about the interconnection and communication between components
    - such as the bus structure, memory hierarchy, and input/output systems.

Categories of Computers

- Analog
  - Uses the continuous variation aspect of physical quantities such as electrical, mechanical, or hydraulic (analog signals) to model the problem being solved.
  - Norden bombsight, radar systems, aircraft flight computers
  - Pre-1950s but current interest in analog/digital hybrid systems
- Digital ← Focus of this class
  - Processes information that is represented as discrete, finite values
  - Majority of contemporary systems (1950 to current)
- Quantum
  - Exploits quantum mechanical phenomena, representing data as qubits using specialized hardware
  - Convergence of quantum mechanics and computer science (1980 to current)

Classes of Computers By Application

- Personal computers
  - General purpose, variety of software, individual use
  - Emphasis on delivery of good performance at reasonable cost
- Server computers
  - Network based, running larger programs accessed by many
  - High capacity, performance, reliability
  - Range from small servers to building sized systems
- Supercomputers (subset of server class)
  - High-end scientific and engineering calculations
  - Highest capability but represent a small faction of the overall computer market
- Embedded computers
  - Components of systems programmed for specific function
    - Smart devices often WiFi connected
  - Stringent power/performance/cost constraints

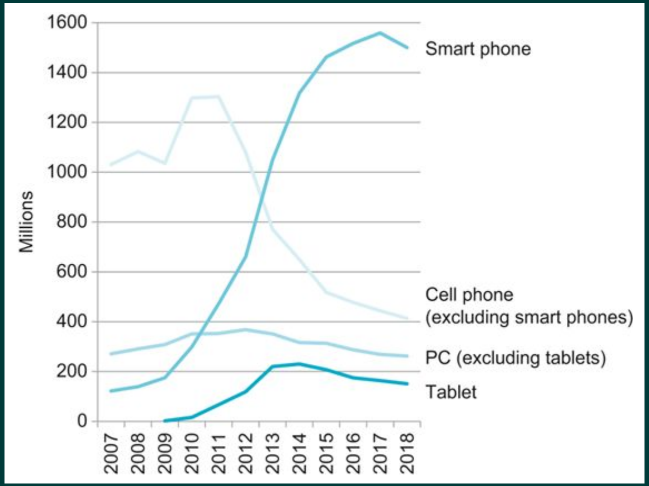
Computer Systems - Hardware

- Computing systems differ in their
  - Size
  - Cost
  - Power
  - Performance

- All computing systems
  - Process digital information
  - Execute binary programs
  - Have microprocessors and memory

Post-PC Era

- The PC is alive and doing well – but –
  - Mobile devices outnumber sales of PCs



- Personal Mobile Device (PMD)
  - Battery operated
  - Wireless connectivity to the Internet
  - Relatively low cost depending on features
  - Run “apps” but no external keyboard or mouse
  - Includes smart phones, smart watches, tablets, electronic eyewear

- Cloud computing
  - Warehouse Scale Computers (WSC)
  - Software as a Service (SAAS)
    - Portion of software run on PMD and a portion run in the cloud
    - Amazon, Google, Microsoft, social media

Understanding Program Performance

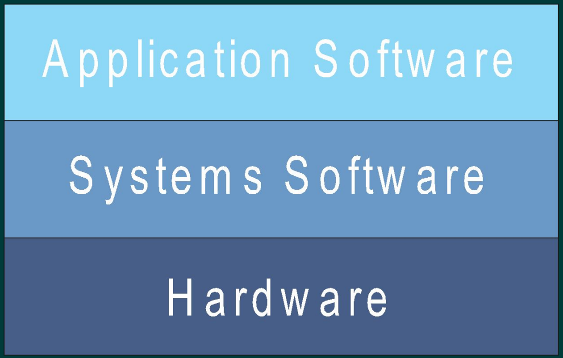
- The performance of a program depends on
  - A combination of the effectiveness of the algorithms used in the program
  - The software systems used to create and translate the program into machine instructions
  - The effectiveness of the computer hardware executing those instructions

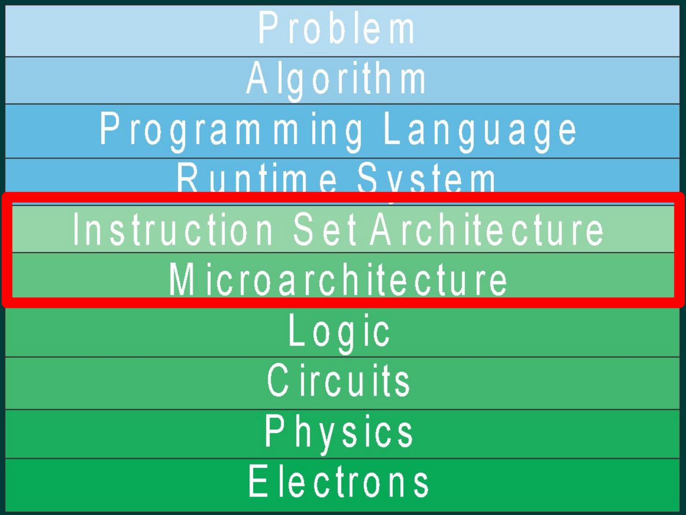
Hardware or Software Component	How this component affects performance
Algorithm	Determines both the number of source-level statements & number of I/O operations executed
Programming language, compiler, & architecture	Determines the number of computer instructions for each source-level statement
Processor & memory system	Determines how fast instructions can execute
I/O system (hardware & operating system)	Determines how fast I/O operations may be executed

- Computational aspects of computer performance will be examined in more detail in a future lecture.

Abstraction – Hardware/Software Stack

- Abstraction helps us deal with complexity
  - Hide lower-level details
- Instruction set architecture (ISA)
  - The hardware/software interface
  - The interface between the hardware and low-level software
  - This interface is binary
- Application binary interface
  - The ISA plus system software interface
- Microarchitecture
  - How the ISA is implemented
  - Organizational level
  - Implementation varies





Instruction Set Architecture (ISA)

- The interface between the hardware and the lowest level software – a.k.a. “architecture”
  - Defines the “rules” by which software must follow in order to be recognized in the hardware
  - One of the most important elements in computer system design
  - This is what distinguishes the performance and design of computer systems
- Two historical architectural categories:
  - CISC – Complex Instruction Set Computer
  - RISC – Reduced Instruction Set Computer
- The instruction set architecture encompasses everything necessary to write a machine language program that will run correctly (includes instructions, registers, memory access, and I/O devices and interfaces)

Technology

- The organization of a computer system is independent of any hardware technology
  - All computers have the same basic components
- The technology used to implement a system can differ among different classes of systems
  - Technology changes and evolves over time
  - How the technology is assembled into a computer system is governed by time-proven design concepts.
- There will always be differences in the implementation of computing systems but the organizational concepts will be consistent

Technology (continued)

- Process integration is the concept of making devices smaller while at the same time increasing the number of components per device
  - Moore's Law – doubling of number of transistors on a single chip every 2 years
- Integrated circuit technology is historically categorized by the degree of integration, also called the growth in chip density
  - Feature size or process node size
  - Measured today in nanometers
  - Measurement of different manufacturers not necessarily equivalent
- The higher the chip density, the more circuits are contained in the chip, thus more function and generally, more complexity

Technology Trends

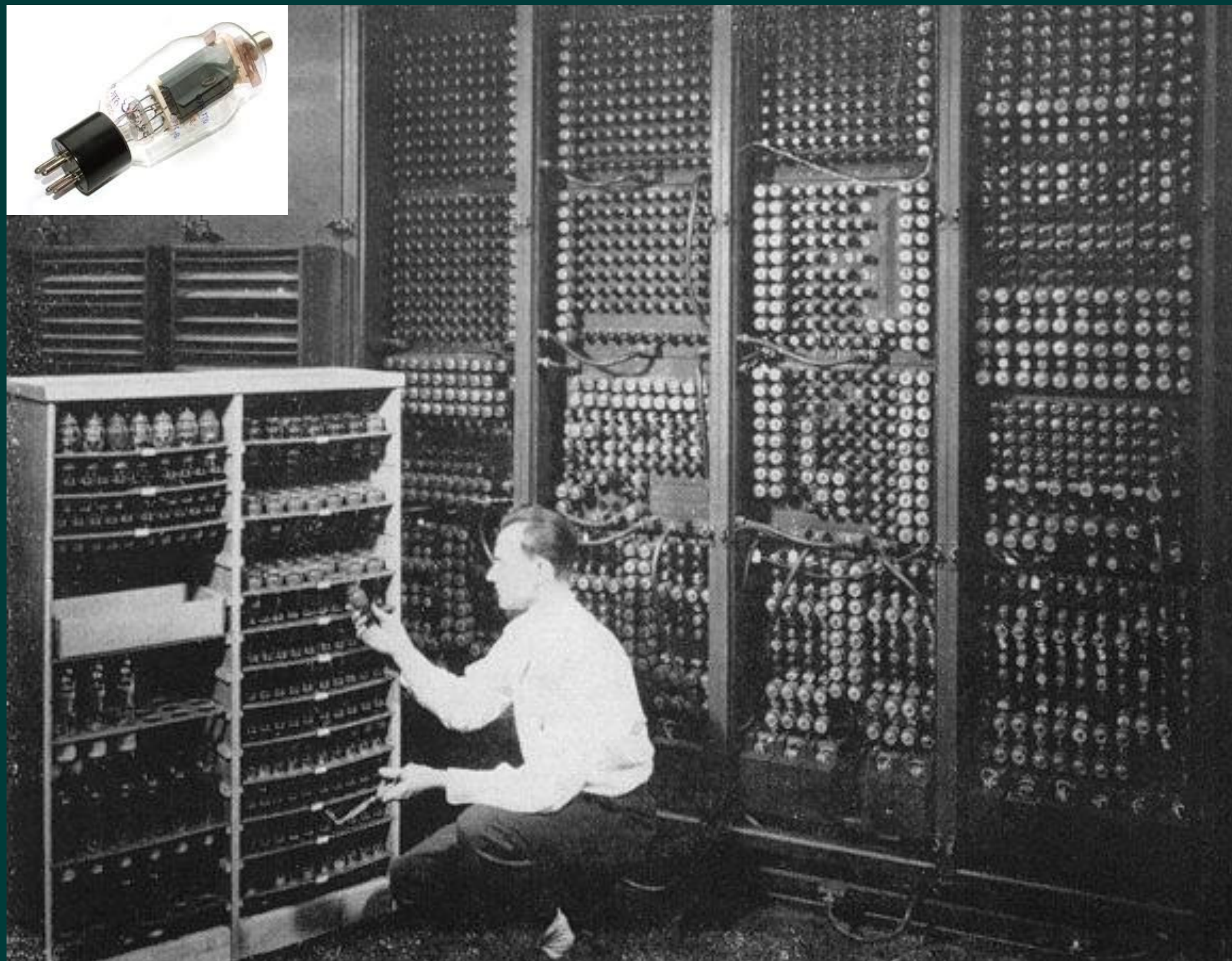
- Electronics technology continues to evolve
  - Increased capacity and performance
  - Reduced cost

Scales of Integrated Circuit Integration

Year	Technology	Relative performance/cost
1942 - 1959	Vacuum tube	1
1959 - 1965	Transistor	35
1965 - 1975	Integrated circuit (IC)	900
1975 -1988	Very large scale IC (VLSI)	2,400,000
1988 - 2008	Ultra large scale IC (ULSI)	250,000,000,000
2008 - Today	Giga scale IC (GSI)	14,880,000,000,000
2019 -	Tera scale IC (TSI)	? (too recent to know)

Relative performance per unit cost of technologies used in computers over time

ENIAC



Replacing a bad tube meant checking among ENIAC's 19,000 possibilities.

IBM Mainframes

- IBM System/360
  - Circa 1965
  - Transistor technology (SLT modules)
  - Solid logic technology

- IBM System/370
  - Circa 1972
  - Integrated circuit technology (memory)

Significant Small Computers For Home

Micral (1973)

Apple II (1977)

Atari 400 (1979)

IBM PC (1981)

Commodore 64 (1982)

IBM Watson

- Data analytics processor that uses natural language processing
- Jeopardy Champion
  - February, 2011

- After Jeopardy, Watson was devoted to research in healthcare, finance, ecosystems, and customer service as a development platform in the Cloud.
- Today, called Watsonx, the platform is being used to advance AI and machine learning technologies.

World's Most Powerful Supercomputer

- Frontier (Top500 – June 2024)
  - DOE/SC/Oak Ridge National Laboratory, Tennessee
  - 8,699,904 cores (AMD Optimized 3rd Generation EPYC 64C)
  - 1,714.81 peak petaflops (22,786 kWh)
  - First exascale computer

https://spectrum.ieee.org/frontier-exascale-supercomputer

World's 2^{nd} Most Powerful Supercomputer

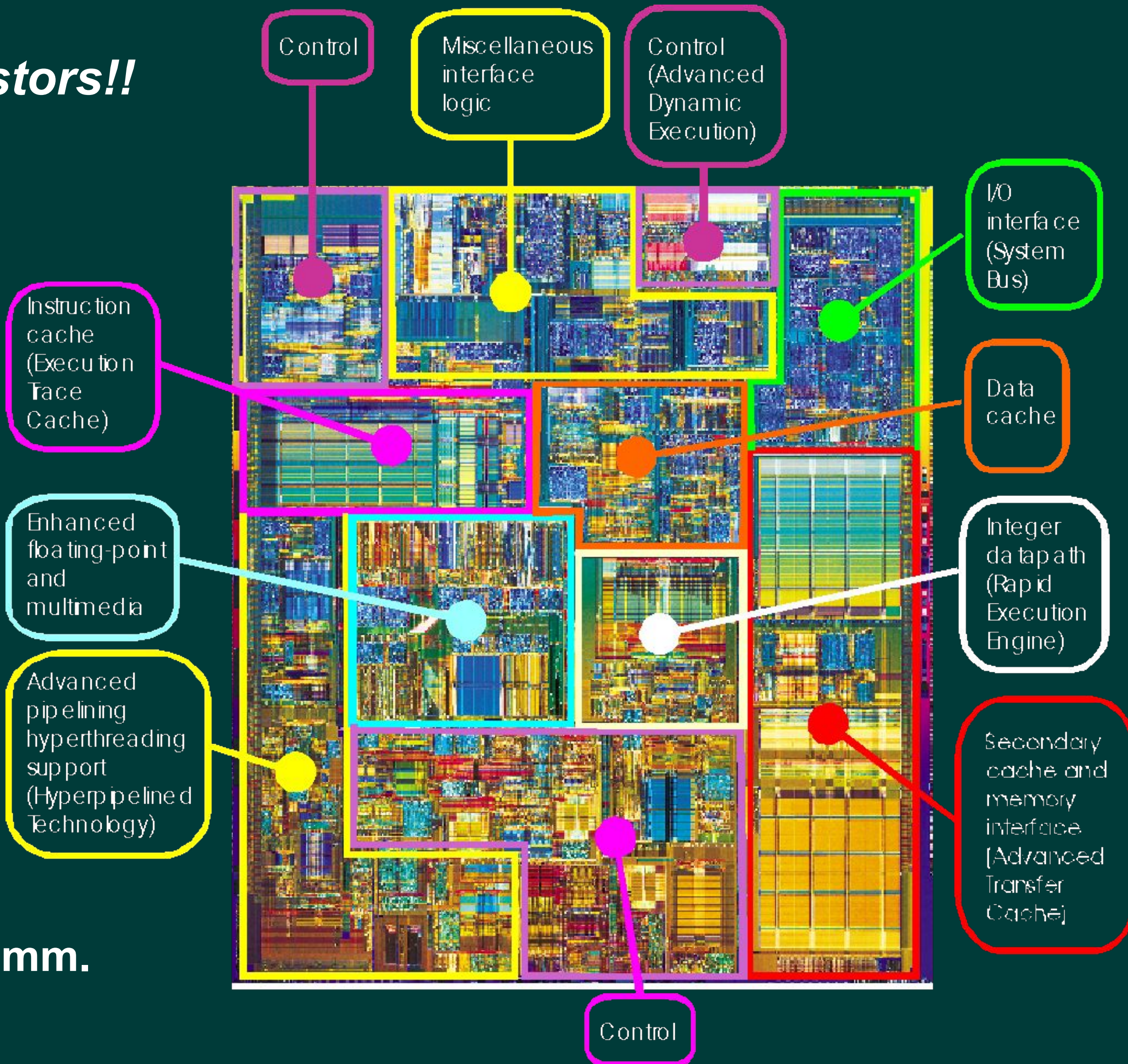
- Aurora (Top500 – June 2024)
  - DOE/SC/ Argonne National Laboratory, Illinois
  - 9,264,128 cores (Xeon CPU Max)
  - 1,980.01 peak petaflops (38,698 kWh)
  - Still being optimized for higher performance

Microprocessor Technology Development

Pentium 4 Processor Die 0.18 Micron Process (2000)

42 million transistors!!

Single core



Die size = 217 sq. mm.

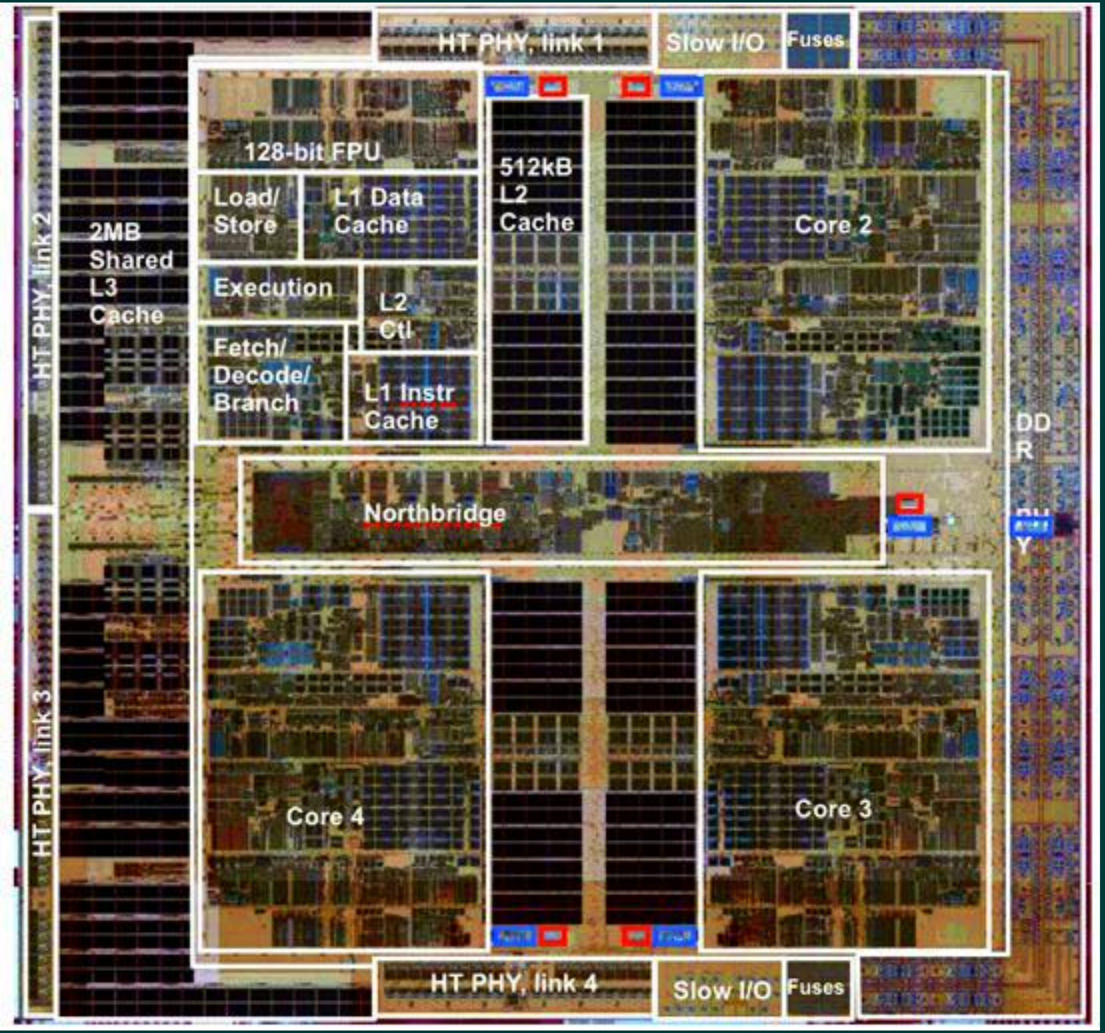
Microprocessor Technology Development

AMD Barcelona Die 65nm Process (2007)

463 million Transistors !!!!

4 CPU cores

Die size = 285 sq. mm

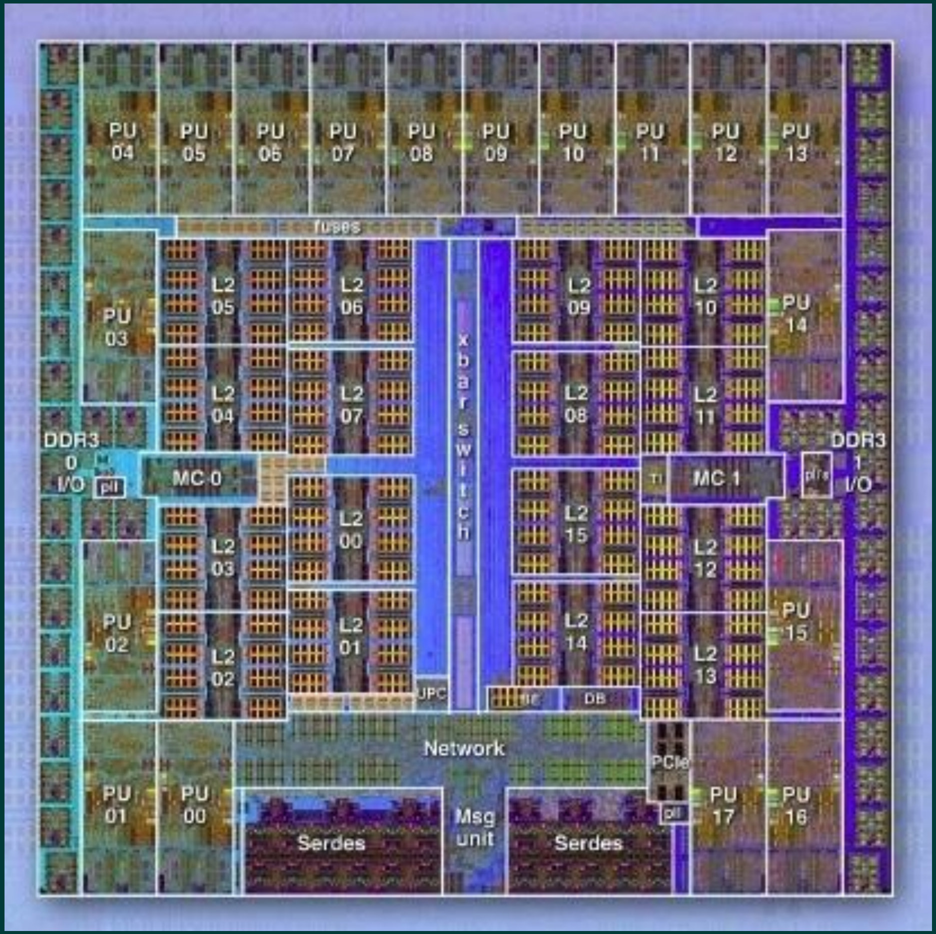


Microprocessor Technology Development

IBM Blue Gene/Q 45nm Process (2012)

1.47 billion Transistors !!!!!!!!

18 CPU cores



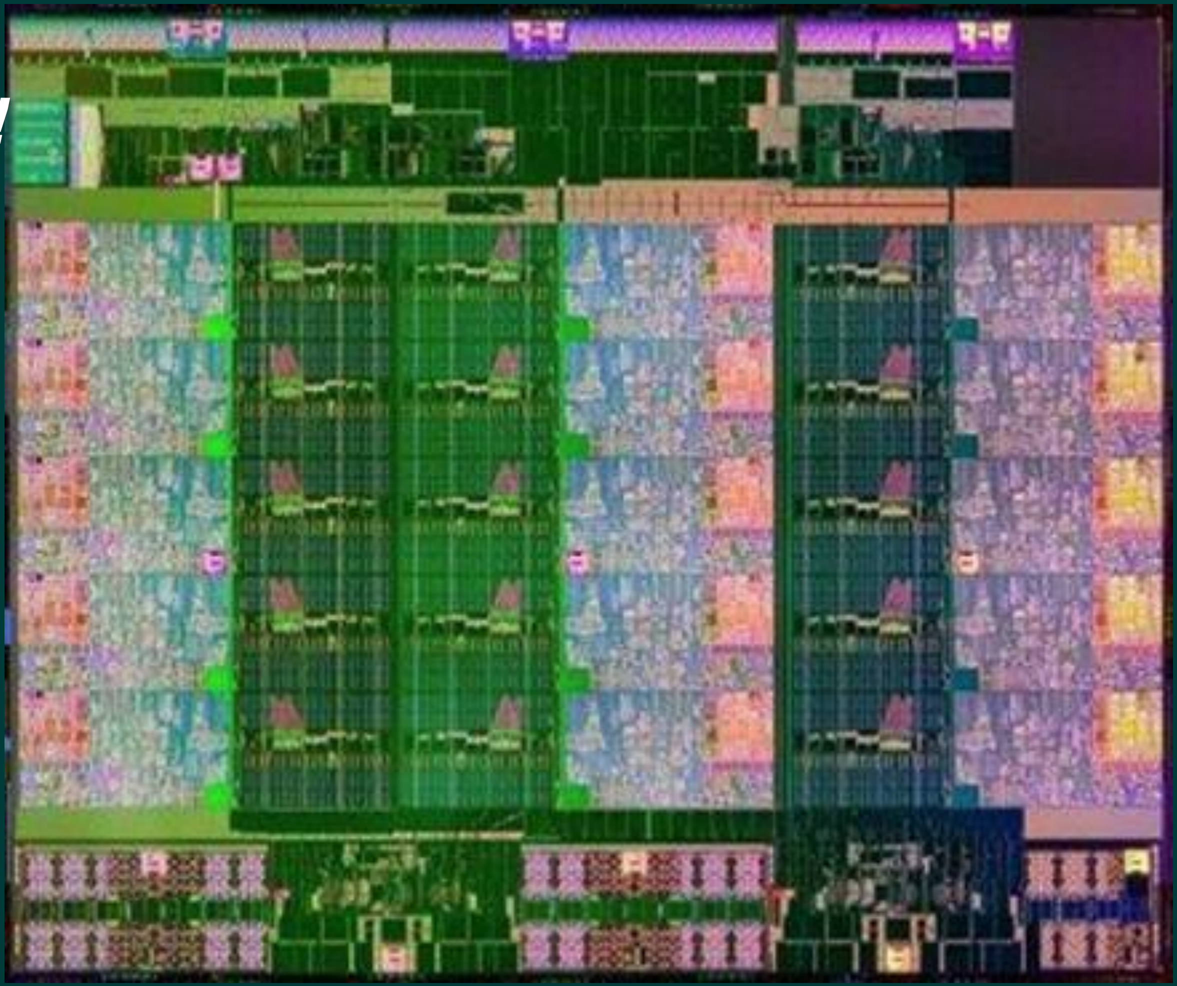
Die size = 359 sq. mm

Microprocessor Technology Development

Intel Ivy Bridge 22nm Process (2014)

4.31 billion Transistors !!!!!!!!

15 CPU cores



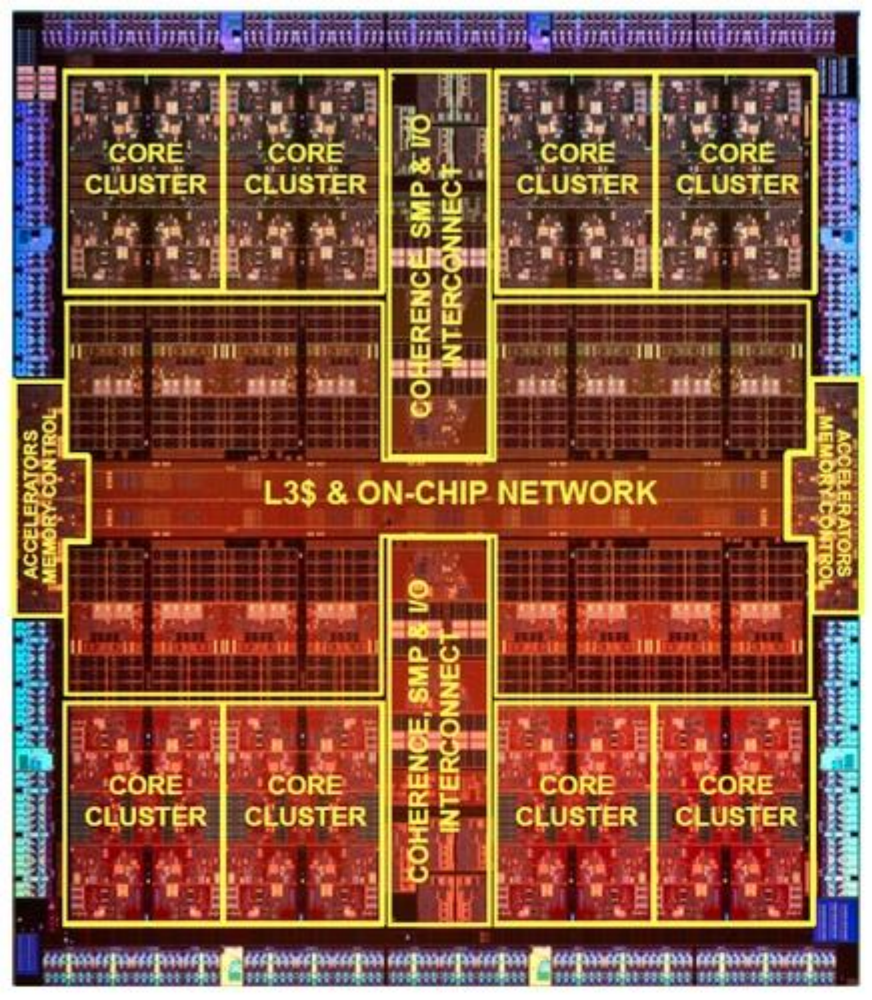
Die size = 541 sq. mm

Microprocessor Technology Development

SPARC M7 20nm Process (2015)

>10 billion Transistors !!!!!!!!

32 CPU cores



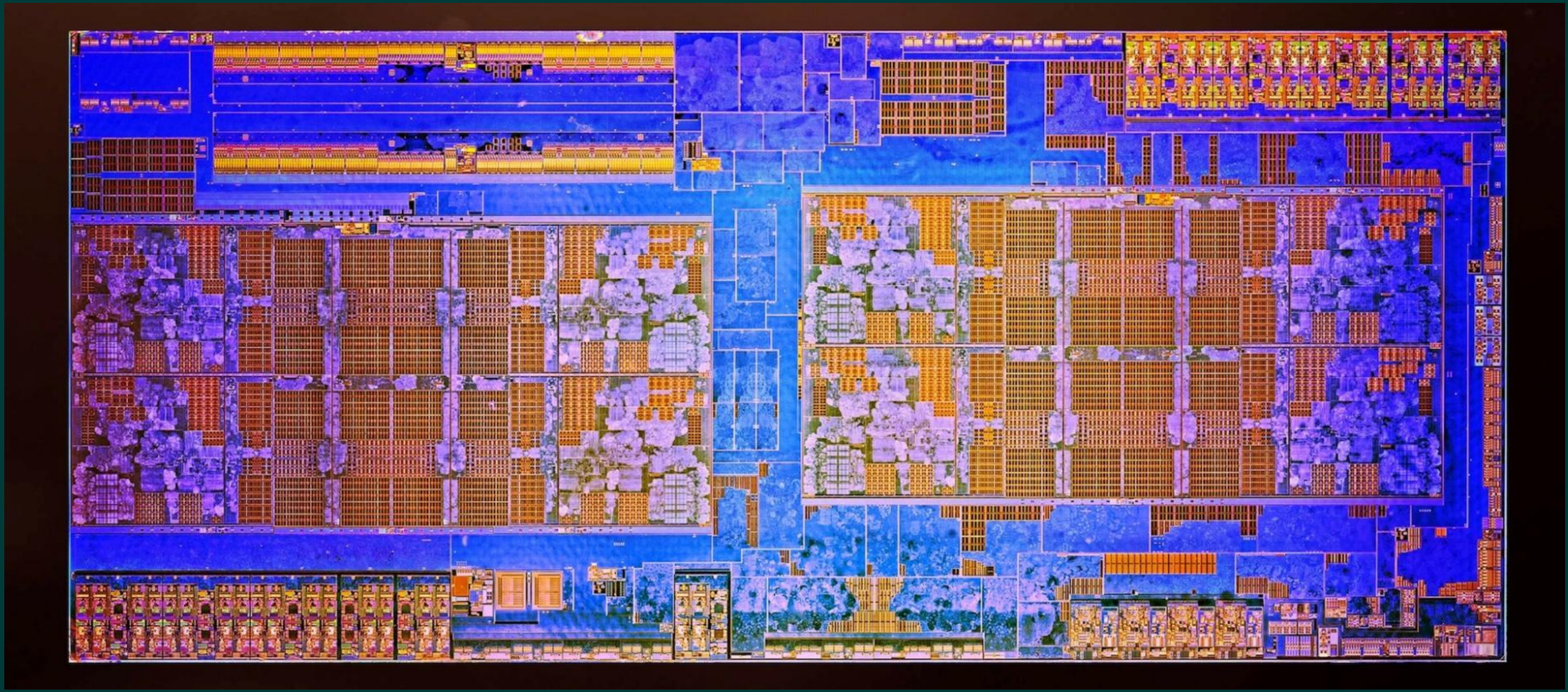
Die size = ~600 sq. mm

Microprocessor Technology Development

AMD EPYC 14nm Process (2017)

19,200,000,000
Transistors !!!!!!!!

8 CPU cores/chip
4 chips/module



Microprocessor Technology Development

Intel Alder Lake 10nm Process (2021)

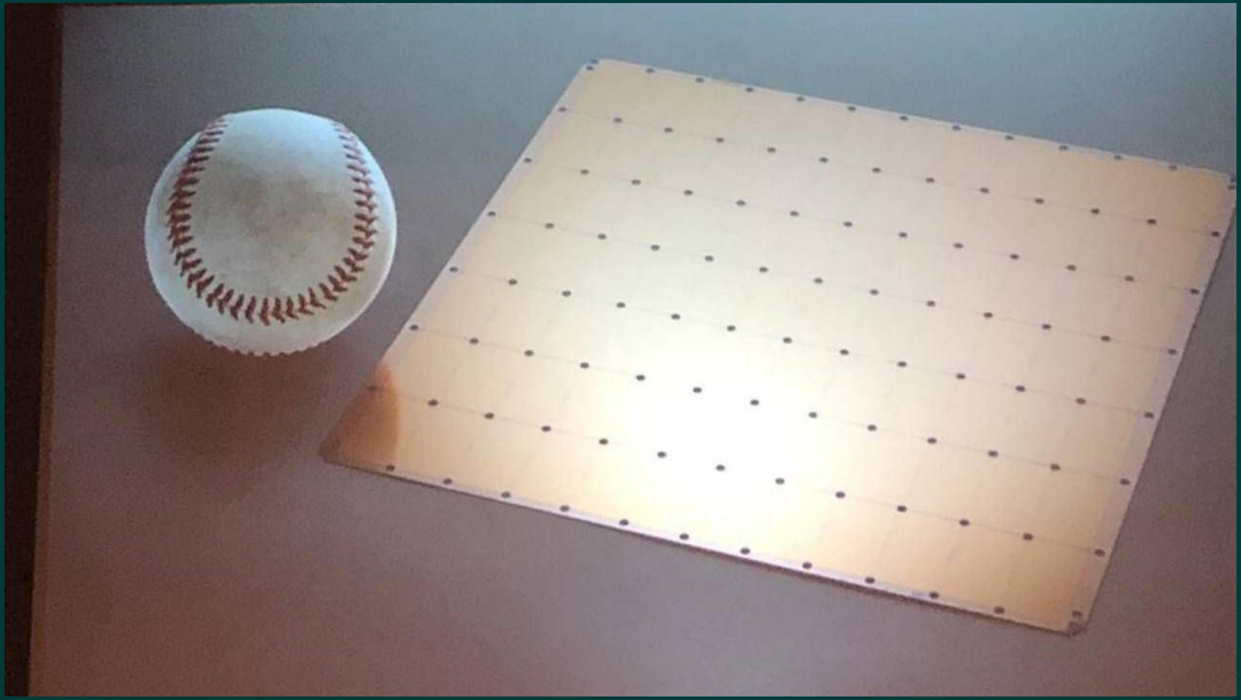
Approx. 21 billion transistors
215.25 mm^2

Big.LITTLE design

- ***8 High Performance cores/chip***
- ***8 Efficient cores/chip***

Something Different: Cerebras WSE-2

- Wafer Scale Engine (WSE-2)
- Designed specifically for AI
- 850,000 AI-optimized cores organized into 84 dies all on a single silicon die
- 40 GB of local superfast SRAM memory
- 2.6 trillion transistors
- 46,225 sq. mm. piece of silicon (from 300mm wafer)



Technology Terminology

- Other technology terms to know
  - GPU (Graphics Processing Unit)
    - processor designed for high compute intensity generally used primarily for graphics, but also vector processing
  - APU (Accelerated Processing Unit)
    - combines a CPU with a GPU on the same chip
    - used in mobile devices and gaming systems
  - SOC (System On a Chip)
    - integrates multiple components (digital and/or analog) onto single chip
  - NPU (Neural Processing Unit), TPU (Tensor Processing Unit)
    - specialized processor explicitly designed for executing machine learning algorithms. NPUs & TPUs are optimized for handling complex mathematical computations integral to artificial neural networks.

Technology Terminology

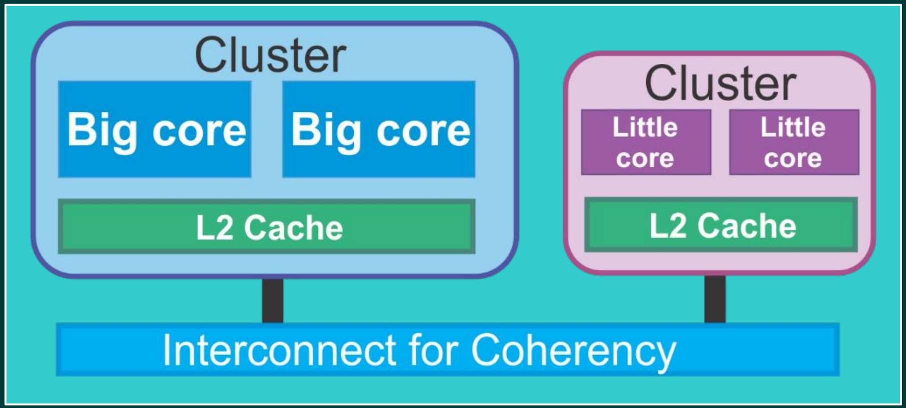
- Other technology terms to know (continued)
  - ASIC (Application Specific Integrated Circuit)
    - customized chip created for a specific use as opposed to general purpose
  - FPGA (Field Programmable Gate Array)
    - chip that can be configured after manufacture by HDL programming
  - VPU (Vision Processing Unit)
    - processor designed to accelerate machine vision tasks; may be incorporated with a NPU
  - DSP (Digital Signal Processor)
    - processing units are optimized for performing digital filtering and Fourier analysis, whether on audio or radio signals or images.

Hybrid Processor Design

- Commonly called “big.LITTLE” architecture
  - Big.LITTLE is trademarked by ARM
  - Apple adopted big.LITTLE architecture for M1 processor and subsequently the M1 Pro & M1 Max
  - Intel’s Alder Lake processors also implement the concept and they call it the performance hybrid architecture (P-cores & E-cores)
- Multi-core design with heterogenous mix of different cores capable of operating at different performance levels determined by type of tasks
  - Compute intensive execution by big/performance cores
  - Less intensive execution by little/efficient cores

big.LITTLE Characteristics

- Clusters of two types of CPU cores
  - Each core is an individual processor
  - Little processors are designed for maximum power efficiency
  - Big processors are designed to provide maximum compute performance.
- Both types of processors are coherent and share the same instruction set architecture.
- Using big.LITTLE technology, each task can be dynamically allocated to a big or little core depending on the instantaneous performance requirement of that task.



Possible Future of Hybrid Technologies

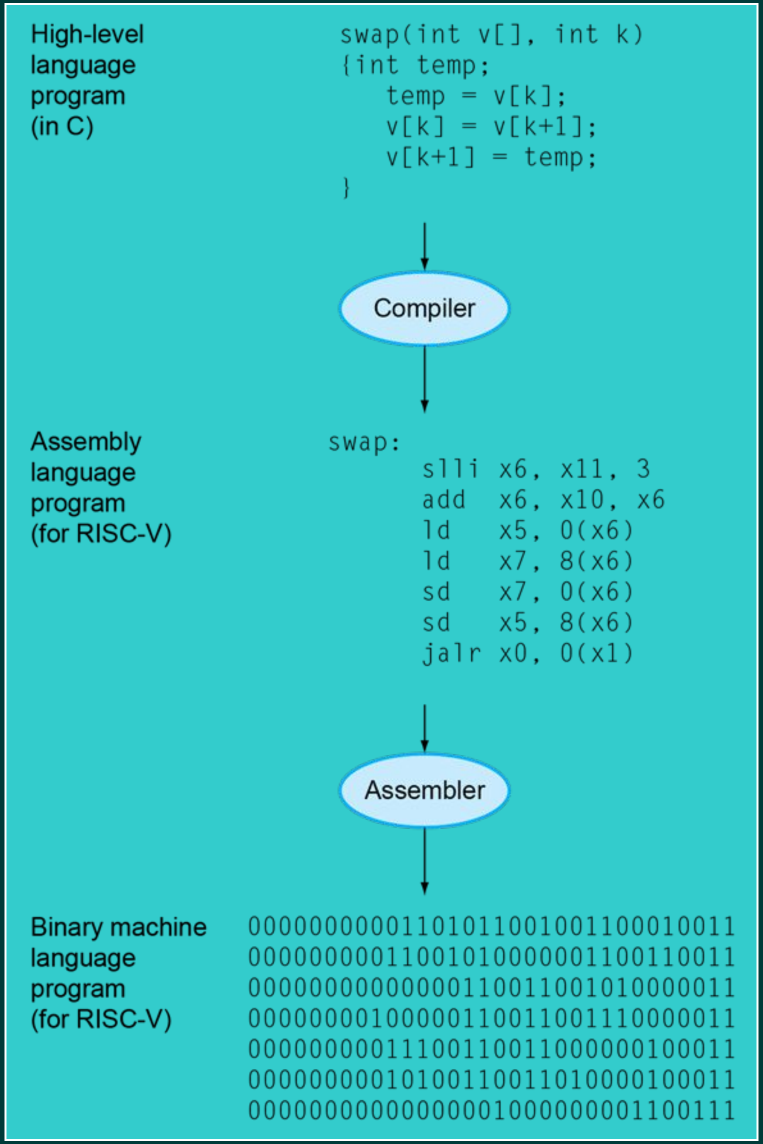
- Today's APUs contain both CPU cores and GPU cores on the same silicon chip
- The big.LITTLE approach expands on this hybrid integrated approach
- Concept of hybrid CPUs can feasibly be expanded to include integrated neural processing units (NPU) for running machine learning algorithms and vision processing units (VPUs) for vision processing
- Looking way into the future (perhaps the 2030s or beyond), QPUs (quantum processing units) could potentially be another type of processor integrated into hybrid CPUs to support complex optimizations
  - Intel, Xanadu and other companies are conducting research in silicon-based quantum possibilities

Instructions and Data

- The two basic things that need to be represented in hardware
  - Instructions - commands
  - Data - numbers, characters, etc.
- Instructions that computers understand are just numbers (collections of bits) that tell the computer to perform some operation, such as add two values.
- All information stored in a computer system whether data or instructions are stored as a binary number.
- If you just look at a number, you can't tell what it represents unless you know its context.

Levels of Program Code

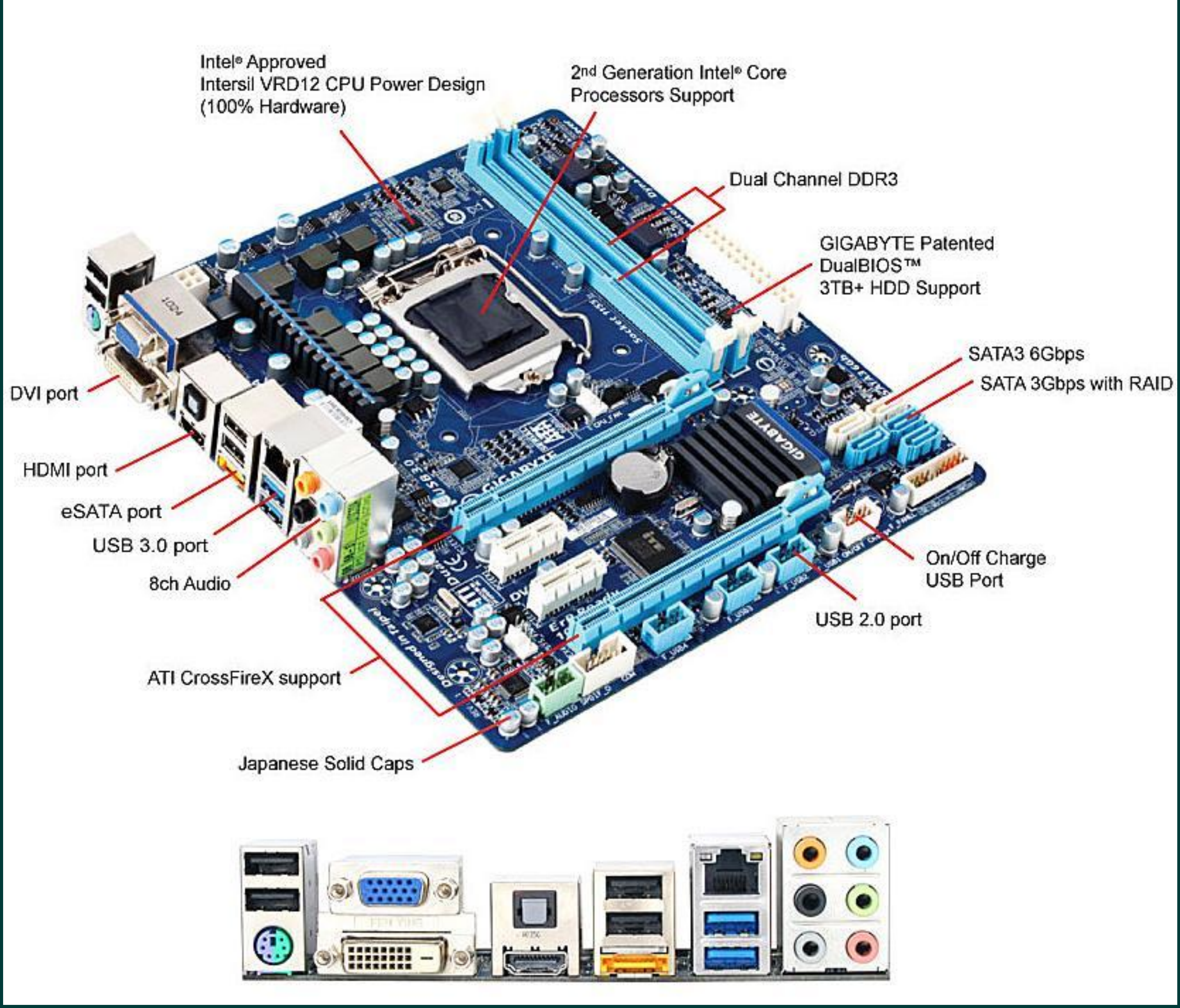
- High-level language
  - Level of abstraction closer to problem domain
  - Provides for productivity and portability
- Assembly language
  - Textual representation of instructions
- Hardware representation
  - Binary digits (bits)
  - Encoded instructions and data



Hardware

- Hardware is what the software runs on
- Computer systems generally have
  - Input only devices (keyboard, mouse, etc.)
  - Output only devices (monitor, printer, etc.)
  - Input/output devices (touch screens, disk, network, etc.)
  - Processors (manipulate information)
  - Memory (local storage)
  - Interfaces (connect all the devices)
- Under the covers
  - Electronics, wires, fans, motors, lights
- Motherboard (Mainboard)
  - Used as base for mounting electronics
  - Contains integrated circuits (chips)
    - Interface chips, memory, the processor

PC Motherboard



Mobile

Laptop

Apple iPad

Microsoft Surface

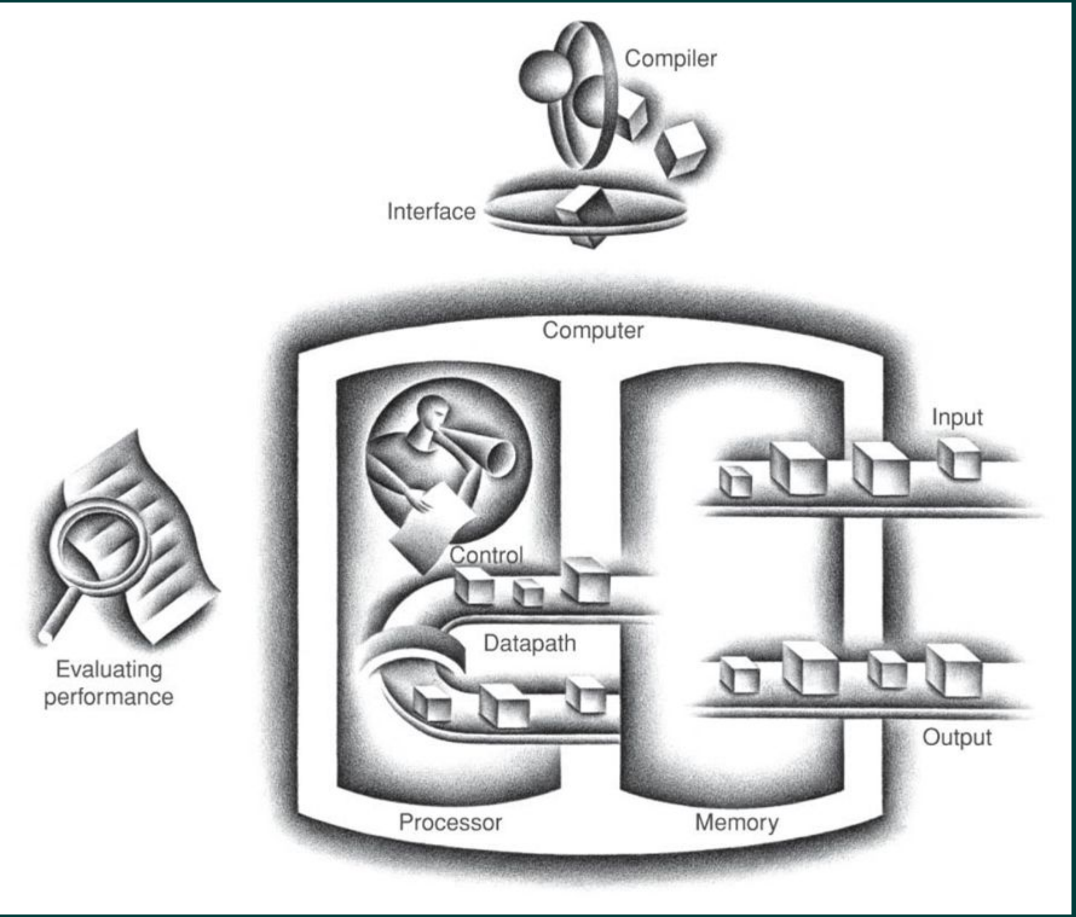
Smartphone

Hardware (continued)

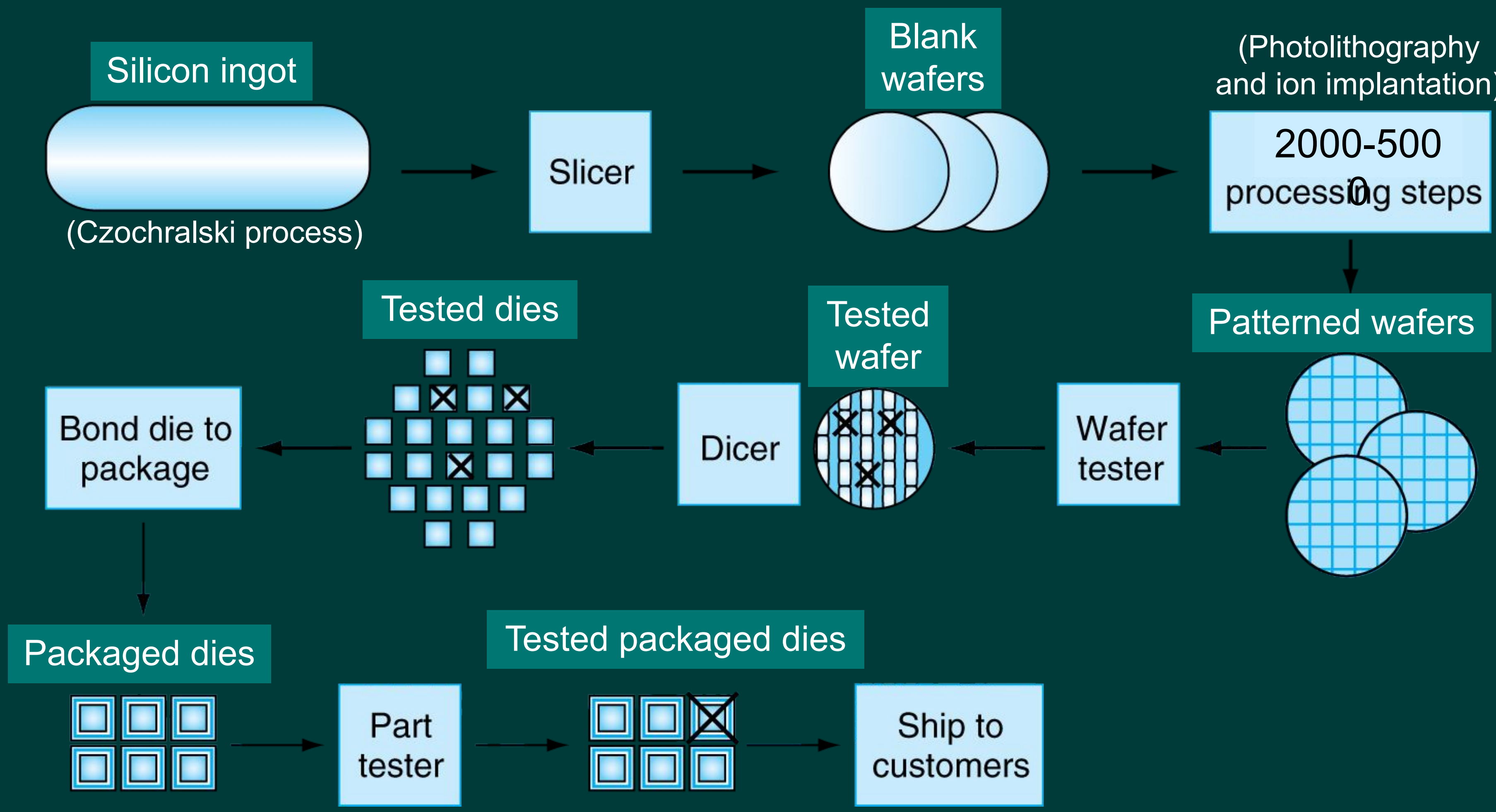
- Interface chips control connections and flow of data between components on the motherboard
- Memory is where the programs and data reside when the programs are running
- Processor is the “brain” – it follows the instructions of the program it executes
- Main processor is called a microprocessor or CPU
  - Many systems have other special purpose processors
  - Ex. Graphic Processing Unit (GPU)
  - Processor within a processor
    - GPU can be integrated into CPU
- Processor is composed of a
  - Datapath – performs arithmetic and logic operations
  - Control – tells the memory, datapath, I/O devices what to do according to the instructions it is executing

Hardware (continued)

- Computer systems contain five functional components
  - Input
  - Output
  - Memory
  - Datapath
  - Control



Making Today's Chips



https://www.youtube.com/watch?v=F2KcZGwntgg
https://www.youtube.com/watch?v=bor0qLifjz4

Summary: Major Concepts

- All computer systems
  - Process Digital Information
  - Execute Programs
  - Use microprocessors
- All data in the hardware is a binary number
- Computers are built with common functionality but the technology may differ among different systems.
- Differences between computer systems is due to internal design
  - Instruction Set Architecture
  - What the programmer needs to know to write programs for the machine
- Evolutionary design vs revolutionary design
  - CISC vs RISC

Terminology to Know

- Desktop computer
- Server computer
- Embedded computer
- microprocessor (CPU)
- datapath
- control
- computer system components
  - (input, output, memory, datapath, control)
- CISC
- RISC
- Instruction Set Architecture
- Microarchitecture
- transistor
- integrated circuit
- Moore's Law
- high level language
- assembly language
- machine language
- abstraction
- GPU
- APU
- SOC
- ASIC
- NPU
- VPU
- DSP
- FPGA
- Hybrid technologies