COMPUTER ORGANIZATION AND ARCHITECTURE (IT 2202)

Lecture 1-2

Prof Hafizur Rahaman

Indian Institute of Engineering Science and Technology, Shibpur
(Formerly, Bengal Engineering and Science University, Shibpur)
Howrah 711 103, West Bengal, India

Course Contents

- Introduction: Review of basic computer architecture

- Data Path Design

- Quantitative techniques in computer design, measuring and reporting performance

- Pipelining Design: instruction and arithmetic pipeline, Hazards, techniques for handling hazards. Pipeline optimization techniques

- Memory technology: Inclusion, Coherence and locality properties; Cache memory organizations, Techniques for reducing cache misses; Virtual memory organization and management techniques, memory replacement policies.

References / Books

1. Computer Architecture and Organization Design Principles and Applications: B. Govindarajalu: TMH
2. Computer Organization and Architecture Designing for Performance: William Stallings: Pearson
3. Computer Architecture A Quantitative Approach: John L. Hennessy and David A. Patterson: ELSEVIER
4. Computer Systems Architecture A Networking Approach: Rob Williams: 2nd Ed: PEARSON
5. Computer Organization and Design The Hardware Software Interface ARM Edition: David A. Patterson and John L. Hennessy: MK

Introduction

- Computer is basic element of human life
  - They are everywhere (embedded Systems?)
  - Laptops, tablets, mobile phones, intelligent applications
- It is needed to understand how a computer works
  - What are the components inside a computer?
  - How does it work?

Course Objectives

We will learn –

- How computers work, basic principles
- How to analyze their performance
- How computers are designed and built
- Issues affecting modern processors (caches, pipelines, etc.)

Course objective is to gain the knowledge required to design and analyze high-performance computer systems.

Course Motivation

Knowledge of CA will be useful if we need to

- design/build a new computer
- design/build a new version of a computer
- improve software performance
- purchase a computer
- provide a solution with an embedded computer

Historical Background

- Constant research on building automatic Computing machines has resulted the development of computers
- Initial efforts: mechanical devices like pulleys, levers and gears
- During World War II: mechanical relays to carry out computations
- Vacuum tube developed: first electronic computer called ENIAC (Electrical Numerical Integrator and Calculator).
- Semiconductor transistors developed and journey of miniaturization started

Historical Background

- Two major stages of development
  - Mechanical : prior to 1945
  - Electronic: after 1945

- Mechanical
  - Abacus: back to 500BC
  - Mechanical adders/subtractor by Blaise Pascal in France (1642)
  - Difference Engine by Charles Babbage in England (1927)
  - Binary mechanical computer by Konard Zuse in Germany (1941)
  - Electromechanical decimal computer by Howard Alken (1944)- Harvard Mark 1 by IBM

Historical Background

- Mechanical
  - Abacus
  - The abacus, developed in china in 500 BC was used for the purpose of calculation of numbers

Historical Background

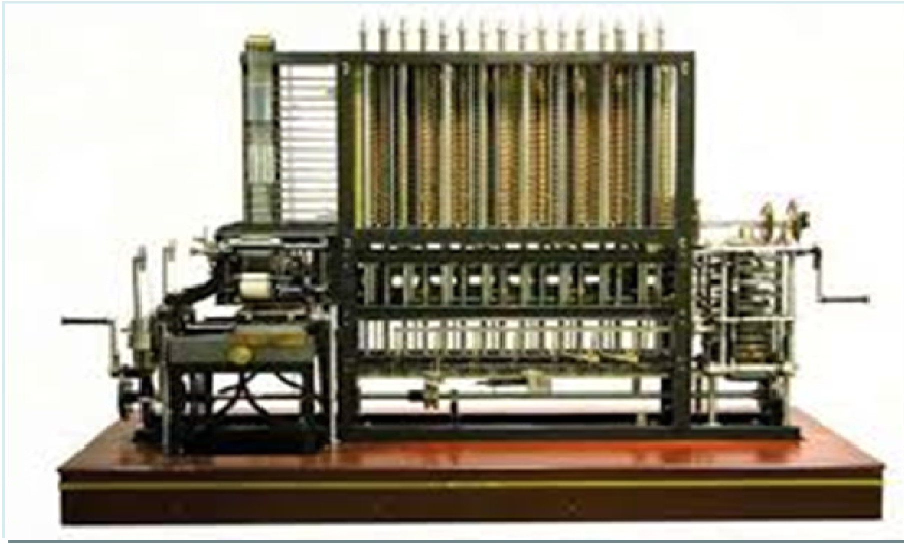
PASCALINE (1642)

- Mechanical calculator invented to add/subtract by Blaise Pascal in France (1642)

- Could add and subtract two numbers directly and multiply and divide by repetition

Historical Background

- First automatic computing engine was designed by Charles Babbage in England in the 19th century, but he could not build it.

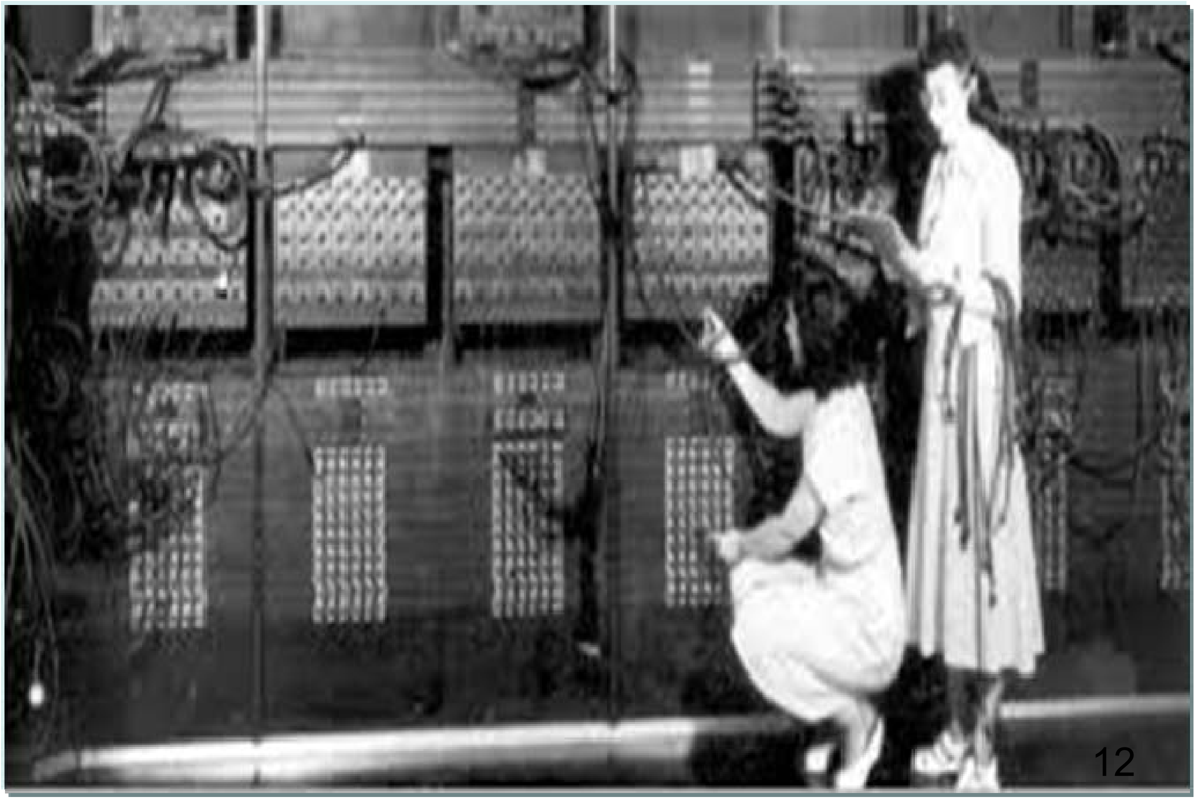


- First complete Babbage engine was built after one century, it was designed
  - 8000 parts
  - Weighted 5 tons
  - 11 feet in length

Historical Background

ENIAC (Electrical Numerical Integrator and Calculator)

- Developed at the University of Pennsylvania
- Used 18,000 vacuum tubes, weighted 30 tons, and occupied a 30ft x 50 ft space.
- First electronic general-purpose computer. It was complete digital and able to solve "a large class of numerical problems" through reprogramming.



Historical Background

Harvard Mark 1

- Built at the University of Harvard in 1944 with support from IBM
- Used mechanical relays (switches) to represent data.
- It weighted 35 tons, and required 500 miles of wiring.

Historical Background

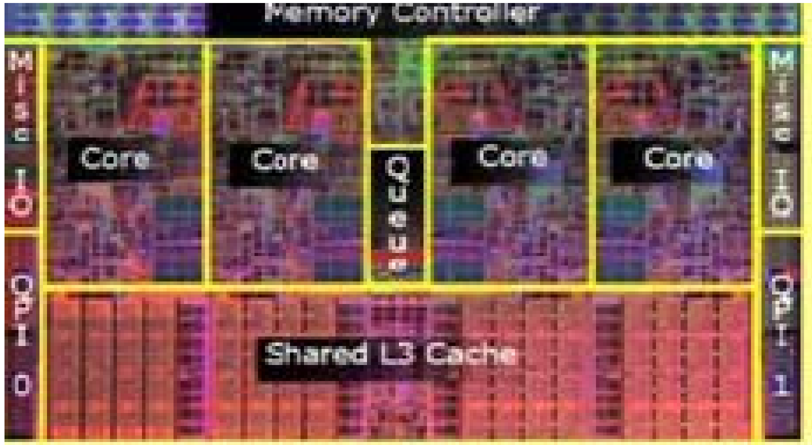
IBM System/360

- Very popular mainframe computer of the 60s and 70s
- Introduced many advanced architectural concepts that appeared in microprocessors several decade later.

Historical Background

Intel Core i7

- A modern processor chip, that comes in dual-core, quad-core and 6-core variants.
- 64 bit processor that comes with various microarchitectures



Five Generations of Electronic Computers

- History of electronic computer development divided into 5 generations
- Each generation characterized by a major technological development
- Fundamental changes in terms of
  - Size
  - Cost
  - Power
  - Efficiency
  - Reliability

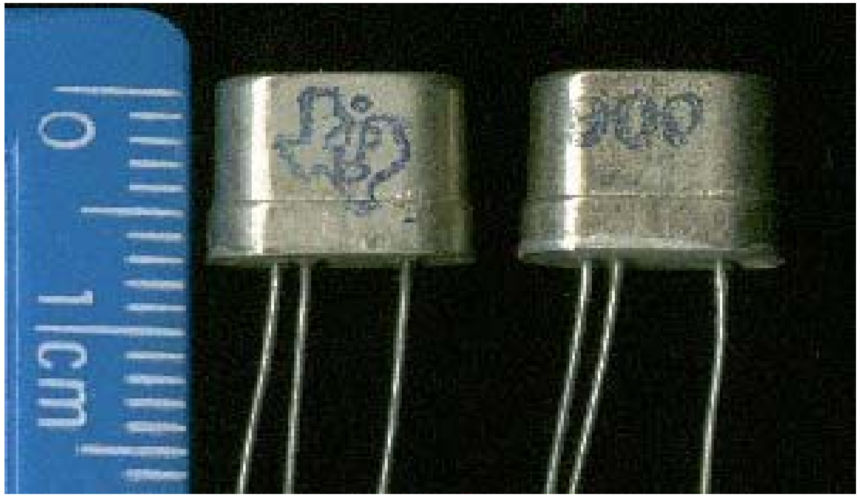
First Generation – (1945-54)

- Used vacuum tubes and relay memories
- expensive, bulky, unreliable, power guzzlers
- Used punched cards/tapes, magnetic drum memories,

- Single user system using machine/assembly language
- ENIAC (Electronic Numerical Integrator And Computer)
- IAS machine- first electronic computer built at the Institute for Advanced Study in Princeton, New Jersey.
- IBM introduced the first electronic computer IBM 701 in 1953.

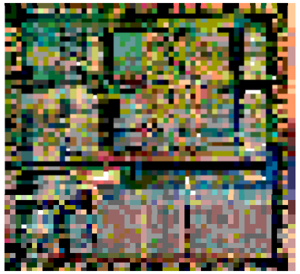
Second Generation (1955-64)

- Used transistors, diodes, magnetic ferrite cores
- smaller, faster, cheaper, more energy-efficient and more reliable as compared to vacuum tubes



- Nobel Prize: J. Bardeen, W. Brattain, W. Shockley
- Assembly languages,
- HLL with compilers early versions of FORTRAN and COBOL
- IBM 7090, CDC 1604, Univac LARC

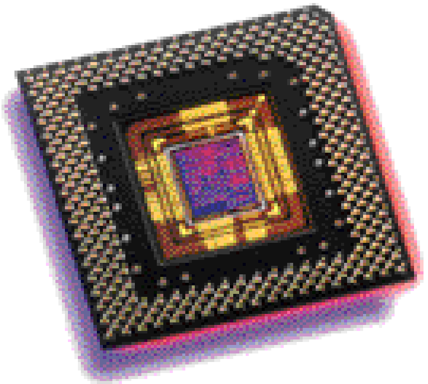
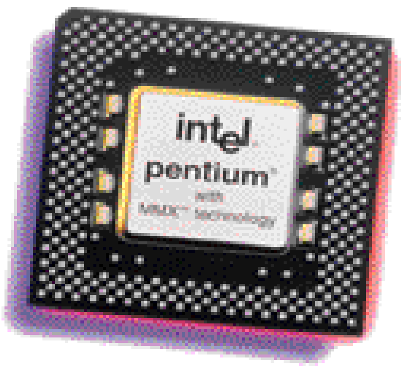
3rd Generation (1965-74)

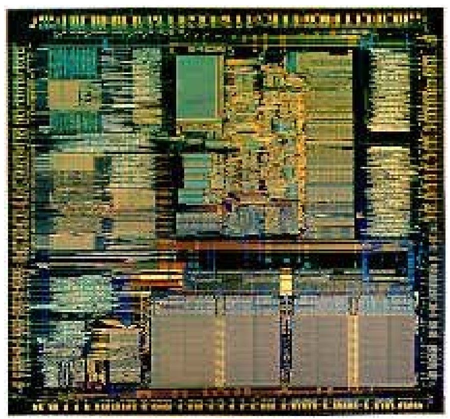
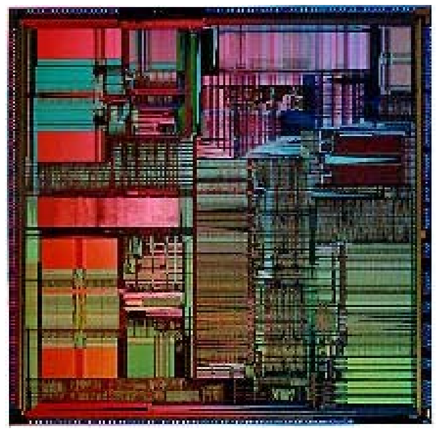


- Used integrated circuits (SSI and MSI)
- SSI (<12 transistors),
- MSI(>12 but <100 transistors)
- LSI (100 or more transistors)
- speed and efficiency drastically increased
- keyboards and monitors
- Multiprogramming and time-sharing operating systems
- IBM 360/370, CDC 6600, TI-ASC, PDP-8

4th Generation (1975-90)

- Used VLSI circuits (LSI and VLSI)
- Multiprocessor OS, HLL, parallel processing
- IBM 3090, VAX 9000, Cray X-MP

5^{th} Generation – (1991-Present)

- ULSI/SOC Techniques have been introduced.
- Massively Parallel Processing, heterogeneous Processing have been implemented.
- Voice recognition as input/output
- Research on natural language
- Parallel processing and scalability
- Video and image processing

Computer Generations

Generation	Main Technology	Representative Systems
First (1945-54)	Vacuum tubes, relays	Machine & assembly language ENIAC, IBM-701
Second (1955-64)	Transistors, memories, I/O processors	Batch processing systems, HLL IBM-7090
Third (1965-74)	SSI and MSI integrated circuits Microprogramming	Multiprogramming / Time sharing IBM 360, Intel 8008
Fourth (1975-84)	LSI and VLSI integrated circuits	Multiprocessors Intel 8086, 8088
Fifth (1984-90)	VLSI, multiprocessor on-chip	Parallel computing, Intel 486
Sixth (1990 onwards)	ULSI, scalable architecture, post-CMOS technologies	Massively parallel processors Pentium, SUN Ultra workstations

Relative Performance Per Unit Cost

- Year Technology Performance/cost
- 1951 Vacuum tube 1
- 1965 Transistor 35
- 1975 Integrated circuit 900
- 1995 VLSI 2,400,000

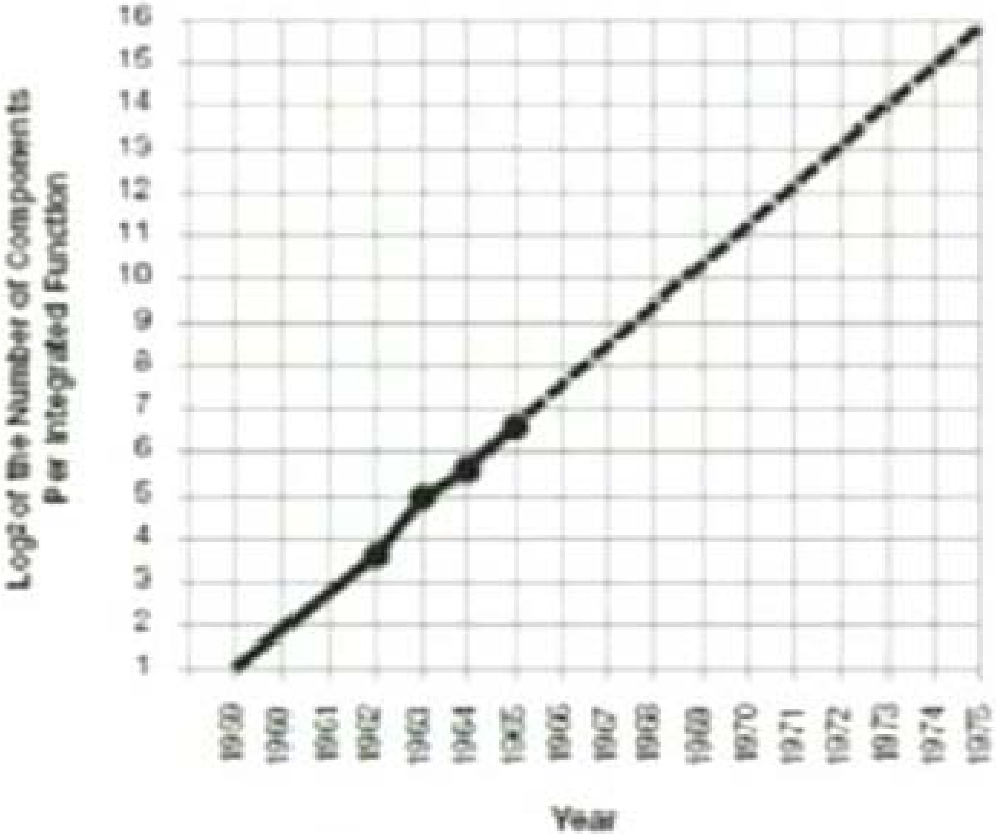
Moore’s Law

- Computer Performance has been increasing phenomenally over the last 6 decades
- Brought out by Moore’s Law:
    - Transistors per square inch roughly double every eighteen months
- Moore’s law is not exactly a law:
    - But has held good for nearly 50 years

Moore’s Law

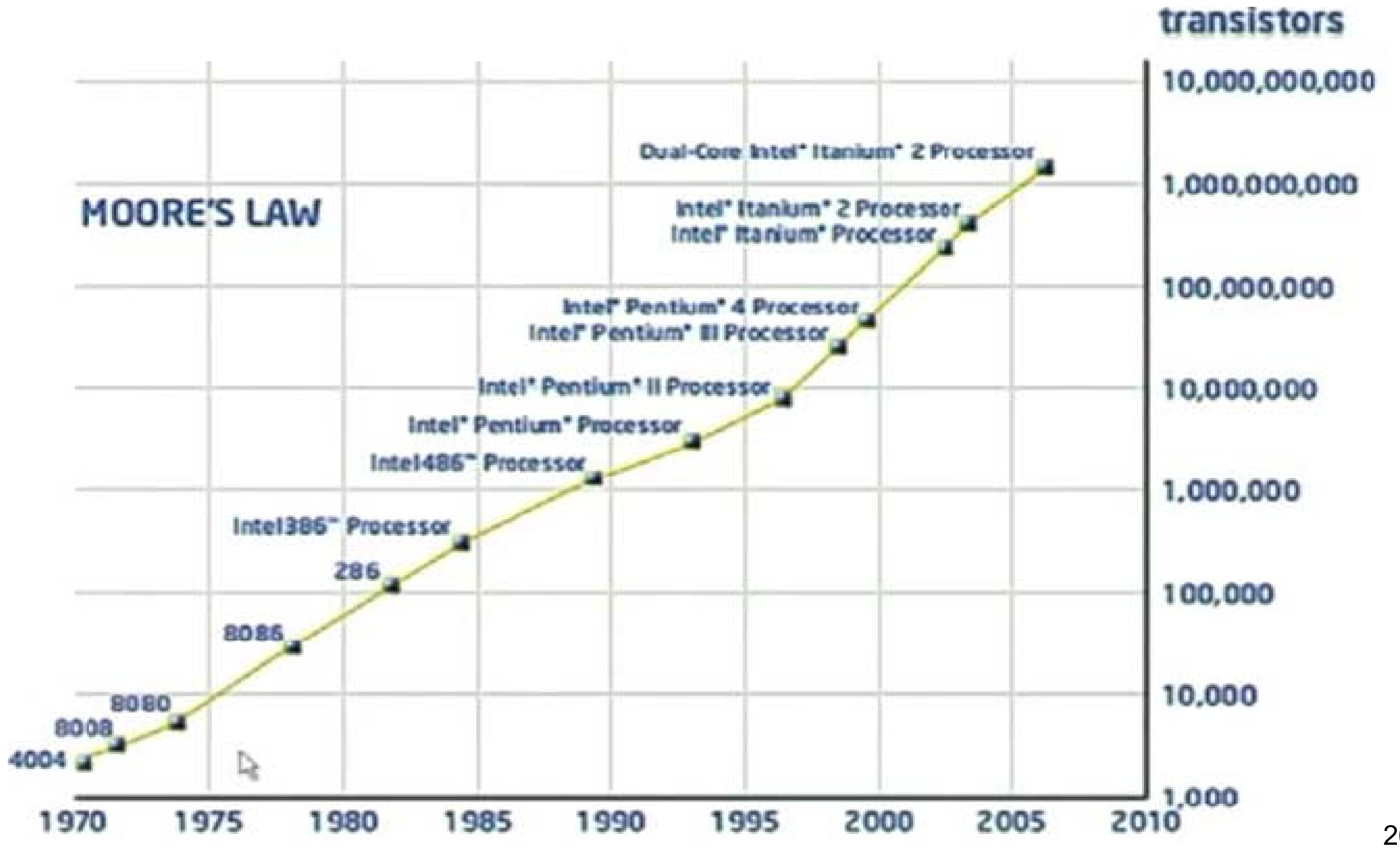
Gordon Moore (co-founder of Intel) predicted in 1965: “Transistor density of minimum cost semiconductor chips would double roughly every 18 months.”

Transistor density is *correlated* to processing speed.



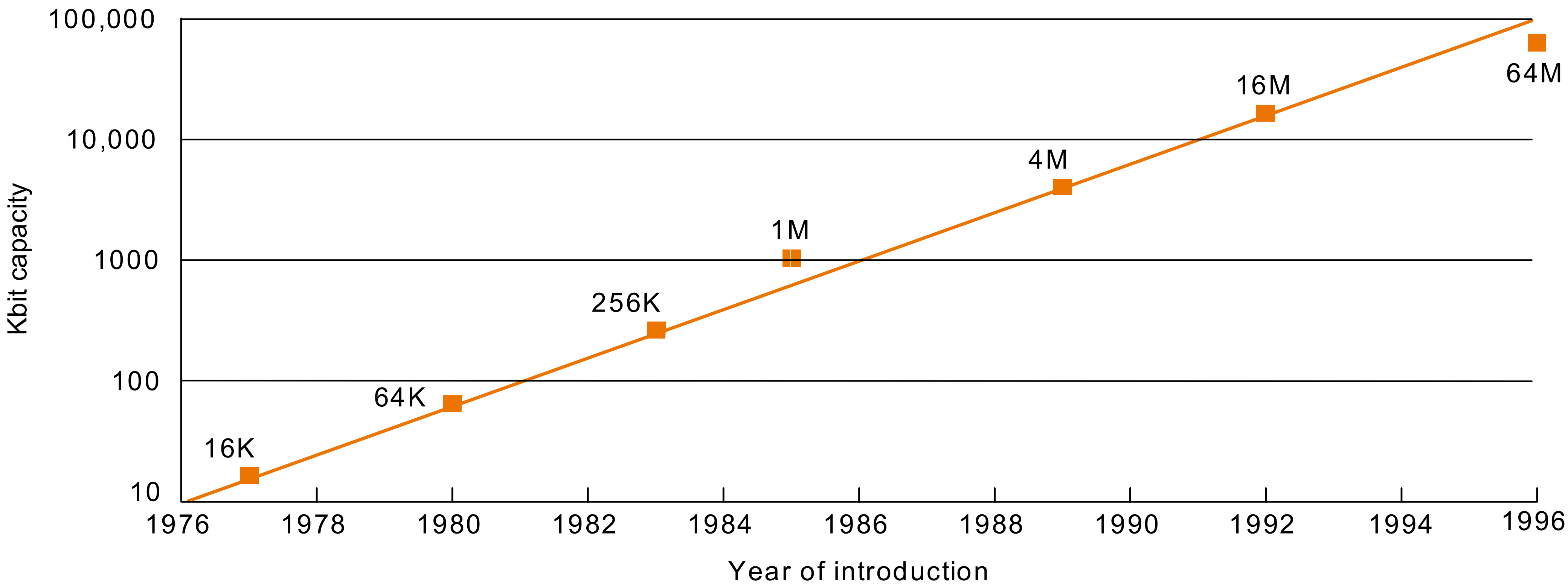
“Cramming More Components onto Integrated Circuits” in the April 19, 1995 issue of the Electronics Magazine

Moore's Law

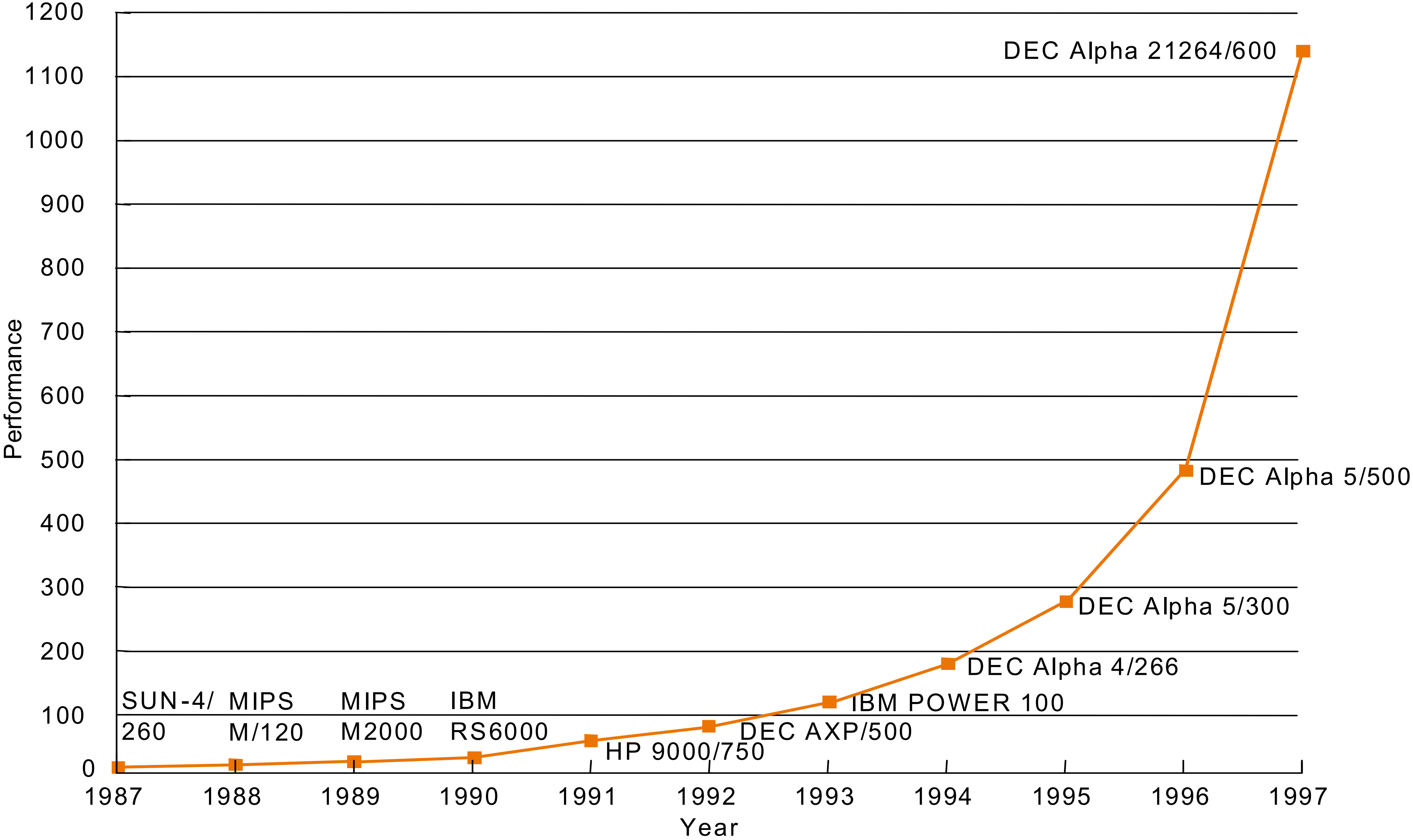


transistors
MOORE'S LAW
10,000,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
Dual-Core Intel® Itanium® 2 Processor
Intel® Itanium® 2 Processor
Intel® Itanium® Processor
Intel® Pentium® 4 Processor
Intel® Pentium® III Processor
Intel® Pentium® II Processor
Intel® Pentium® Processor
Intel486™ Processor
Intel386™ Processor
286
8086
8080
8008
4004
1970
1975
1980
1985
1990
1995
2000
2005
2010

Growth in DRAM Capacity



Increase in Workstation Performance



Embedded Computers

- Processor as an intelligent electronic component rather than Processor as a computing engine
- Real time operation
- Requires hardware-software co-design
- Highly customized

Why different Processors?

- What is the difference between processors used in desk-tops, lap-tops, mobile phones, washing machines etc.?
  - Performance / speed
  - Power consumption
  - Cost
  - General purpose / special purpose

Computer Architecture

- Architecture refers to those attributes of a system visible to a programmer, or those attributes that have a direct impact on the logical execution of a program.

- Architectural attributes include the instruction set, the number of bits used to represent various data types (e.g., numbers, characters), I/O mechanisms, and techniques for addressing memory.

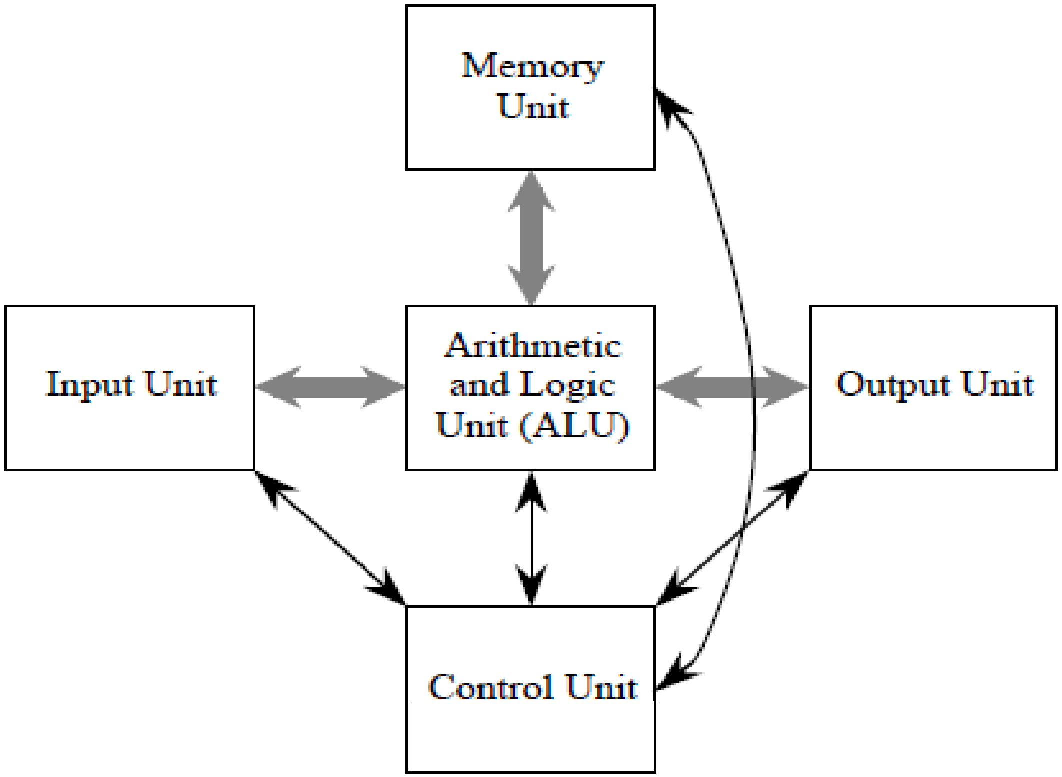
Architecture Versus Organization

- What is the difference between:
  - computer architecture and computer organization?
  - Architecture:
  - Also known as Instruction Set Architecture (ISA)
  - Programmer view of a processor: instruction set, registers, addressing modes, etc.
- Organization:
  - High-level design: how many caches? how many arithmetic and logic units? What type of pipelining, control design, etc.
  - Sometimes known as micro-architecture

Computer Architecture

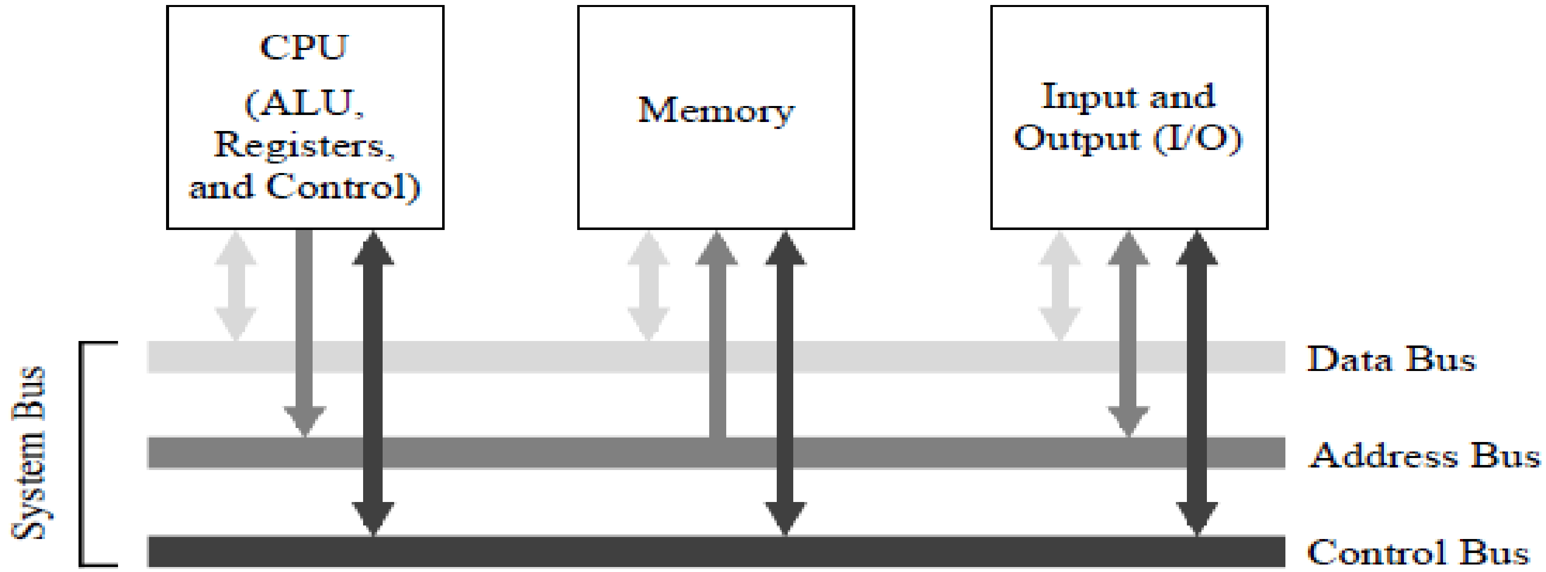
Von Neumann Model: Von Neumann model consists of five major components:

1. Input unit
2. Output unit
3. Arithmetic Logic Unit
4. Memory Unit
5. Control Unit

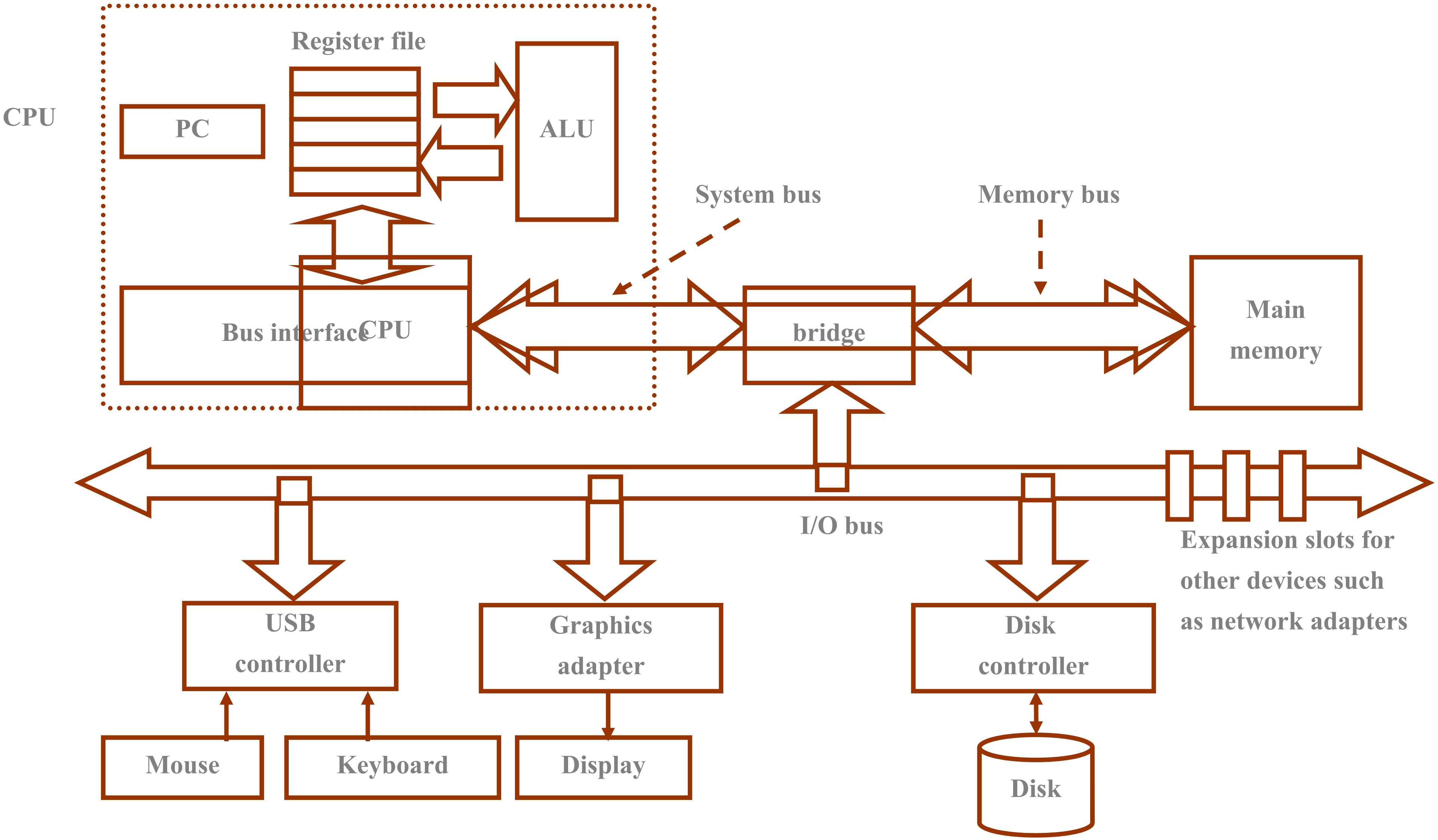


System Bus Model

- A refinement of Von Neumann model consists of CPU (ALU and Control), memory and Input/Output Unit
- Communication among components is handled by a shared pathway called system bus consisting of data bus, address bus and control bus
- There is also a power bus
- Some architectures may also have a separate I/O bus



Inside the Processor



Inside the Processor

- A processor is also called a Central Processing Unit; it consists of a Control Unit and an Arithmetic Logic Unit.
- All the calculations happens inside the CPU.
- All arithmetic operations are performed inside the ALU.
- Control unit basically generates the sequence of control signals to carry out all operations.
- The control unit generates those sequences of control signals.

Inside the Processor

- Finally, you have to store the result will be stored or will be displayed.

- All these steps that we are instructing to a computer are generated by the control unit.
- In a processor an instruction actually specify the exact operation that is to be performed.
- A program is a set of instructions that constitute a program.

Inside the Processor

- ALU – a processor unit -performs arithmetic and logical operation inside the computer.
- ALU consists of several registers;
  - General Purpose Registers;
  - Special Purpose Registers;
  - Temporary storage
- Registers are storage unit, and these registers are used to store data
- Some operations are computed and again the results will be stored into it.
- We store data for computation; and after the computation is performed, we also store back the data.
- It contains circuitry to carry out the logic operations like AND, OR, NOT, SHIFT, COMPARE etc

Inside the Processor

- **ALU also contains circuitry to carry out arithmetic operations like addition, subtraction, multiplication, division.**
- **During execution, the data or the operands are brought in and stored in some register, the desired operation is carried out and the result is stored back in some register or memory.**
- **During an instruction execution, the instructions and data are stored in some memory locations.**
- **So, we have to bring the data from those memory locations into some of the registers,**
- **We perform the operation and then we store back the data into either register or into those memory locations.**

Inside the Processor

- **ALU also contains circuitry to carry out arithmetic operations like addition, subtraction, multiplication, division.**
- **During execution, the data or the operands are brought in and stored in some register, the desired operation is carried out and the result is stored back in some register or memory.**
- **During an instruction execution, the instructions and data are stored in some memory locations.**
- **So, we have to bring the data from those memory locations into some of the registers,**
- **We perform the operation and then we store back the data into either register or into those memory locations.**

Inside the Processor (Role of control unit)

- **The control unit generates the signals that is necessary to perform the task. So, it acts as a nerve center that senses the states of various functional units and sent control signals to control the states.**

- **To carry out a specific operation (say, R1← R2+R3), the control unit must generate control signal in a specific sequence.**
  - **Enable the outputs of registers R2 and R2.**
  - **Select the addition operation.**
  - **Store the output of the adder circuit into register R1.**
- **When an instruction is fetched from memory, the**
- **operation (called opcode) is decoded unit, and the control signals is issued.**

Inside the Processor (Memory Unit)

- **Now, consider the memory unit. There are two types of memory subsystems, two main types.**
  - **Primary or Main memory, which stores the active instructions and data for program being executed on the processor.**
  - **Secondary memory which is used as backup and stores all active and inactive programs and data typically the files.**
- **The processor only has a direct access to primary memory.**
- **The programs and data are stored in the primary memory and whenever it is required, the processor takes it from the primary memory and not from your secondary memory.**
- **In reality, the memory system is implemented as a hierarchy of several levels.**
- **we have L1 cache, we have L2 cache, we have L3 cache, primary memory and secondary memory.**

Inside the Processor (Memory Unit)

- Objective is to provide faster memory access at affordable cost

- We store instructions and data in the memory unit and for processing of those instructions on data, we have to bring those instructions and data to the processors to execute it.

Inside the Processor (Memory Unit)

- We have various different types of memory:- Random Access Memory, Read Only Memory, Magnetic Disk, Flash Memories etc.
- Random Access Memory (RAM) is used for cache and primary memory. Read and Write access times are independent of the location being accessed. This means, if we either access location 1 or last location or the middle location, the access time will be same.
- Read Only Memories (ROM) are used as a part of primary memory to store some fixed data that can not be changed.

Inside the Processor (Memory Unit)

- Magnetic Disk uses direction of magnetization of tiny magnetic particles on a metallic surface to store the data. The access time varies depending on the location being accessed, and these are used in secondary memory.

- Presently, Flash Memories is replacing this magnetic disk as secondary memory. They are much faster and smaller in size, and they do not have any movable part.

Inside the Processor (Memory Unit)

- Magnetic Disk uses direction of magnetization of tiny magnetic particles on a metallic surface to store the data. The access time varies depending on the location being accessed, and these are used in secondary memory.

- Presently, Flash Memories is replacing this magnetic disk as secondary memory. They are much faster and smaller in size, and they do not have any movable part.

- Flash memory is an electronic non-volatile memory storage medium that can be electrically erased and reprogrammed.

- Two main types of flash memory:- NOR flash and NAND flash based on NOR and NAND logic gates

- NOR and NAND flash use MOSFET Technology

Inside the Processor (Input Unit)

- is used to feed data to the computer system.
- The commonly used devices are
  - Keyboards,
  - Mouse,
  - Joystick
  - Camera.

Inside the Processor (Output Unit)

- is used to send the result of some computation to outside world
  - Common output devices are
  - Printer is used to print the data
  - LCD screen/LED screen is used to see the output on the screen
  - Speaker is used to hear the voice
  - Projection systems are also used as an output unit