Computer Architecture

Course Code: CSE 360

Presented by

Dr. Md. Nawab Yousuf Ali

Professor, Dept. of CSE

Computer System Architecture

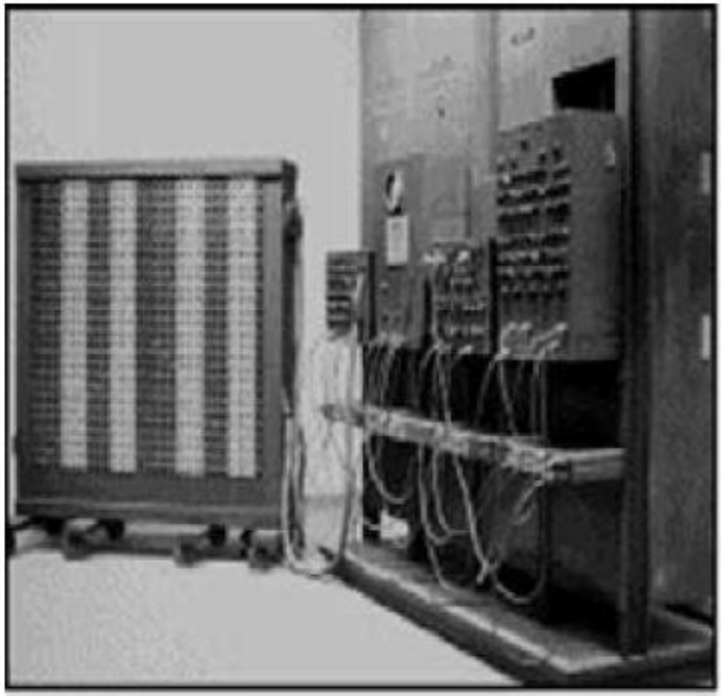
Course Code: CSE360

Lecture - 2
Computer Evolution and Performance

ENIAC – background: The First Generation

- First Generation used Vacuume Tubes
- Electronic Numerical Integrator And Computer
- Developed by Eckert and Mauchly
- At University of Pennsylvania
- Responsible for developing Trajectory tables for new weapons (without these firing tables, the new weapons and artillery were useless to gunners)
- Started 1943
- Finished 1946
- Used until 1955

ENIAC - 1946



It was U shaped, 25m long, 2.5m high and 1m wide

ENIAC - details

- The resulting machine was enormous!!!
- Weighting 30 tons
- Occupying 15,000 square feet of floor space (30x50 feet)
- Contains 18,000 vacuum tubes
- 140 kW power consumption when operating
- Capable of 5,000 additions per second
- Decimal number system (not binary); numbers were represented in decimal form and arithmetic was performed in the decimal system.
- Memory consists of 20 accumulators, each capable of holding 10 digits.

Major drawback

- Programming was done manually by plugging and unplugging cables and setting switches. Data was entered by punched cards.
- Programming for typical calculations took from half an hour to a day.

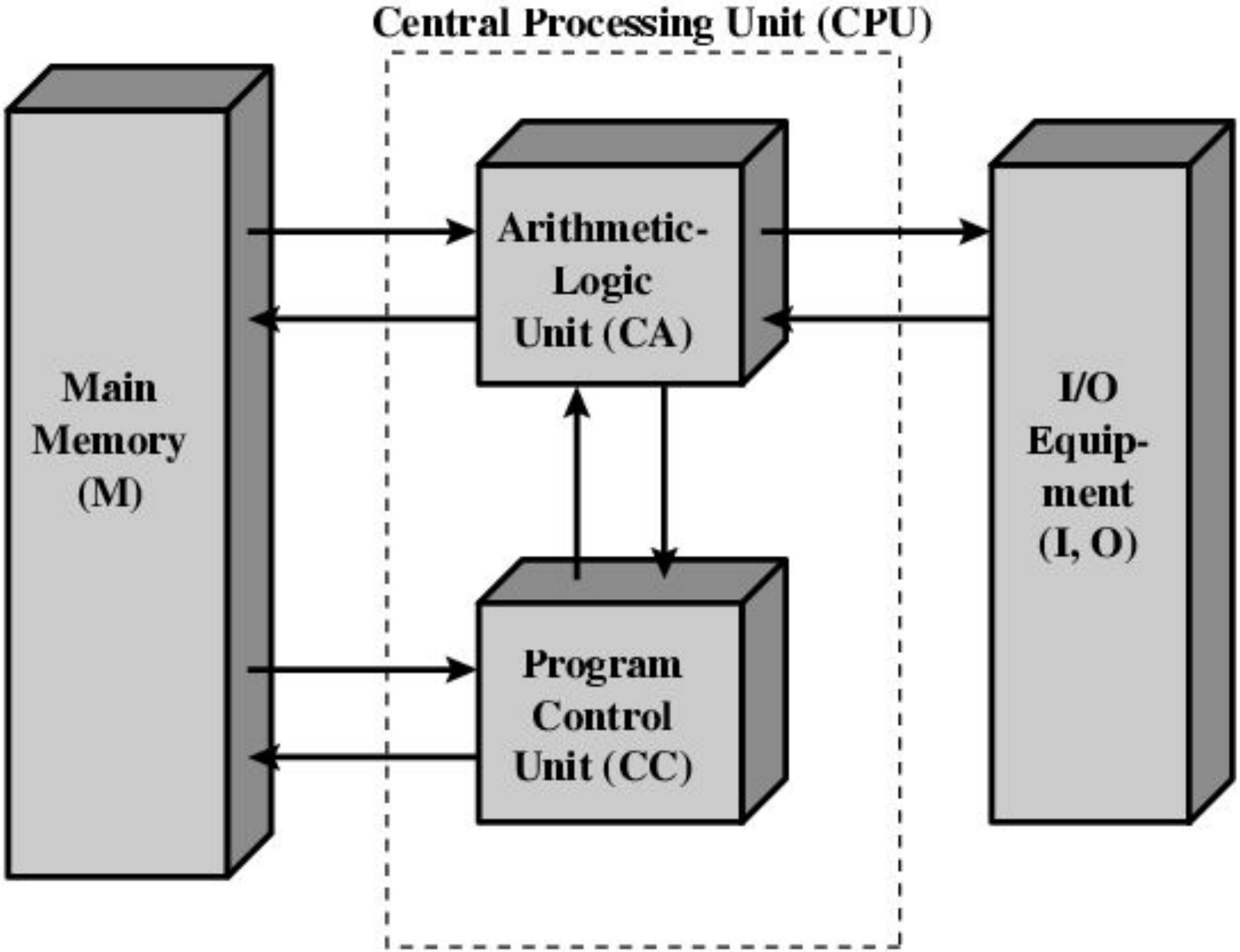
von Neumann/Turing

- Von Neumann (mathematician) was a consultant to both ENIAC and EDVAC (Electronic discrete variable computer) projects
- Proposal of Neumann, EDVAC
- In 1946, Von Neumann began to design **a new stored program computer**, referred to as ***IAS (Institute for Advanced Studies)*** *computer,* at the Princeton Institute for Advanced Studies
- IAS computer, although not completed until 1952, is prototype of all subsequent general-purpose computers.

General structure of the IAS computer:

- A main memory which stores both data and instructions
- An arithmetic-logic unit (ALU) capable of operating on binary data
- A control unit interpreting instructions in memory and causes them executing
- Input and output equipment operated by control unit

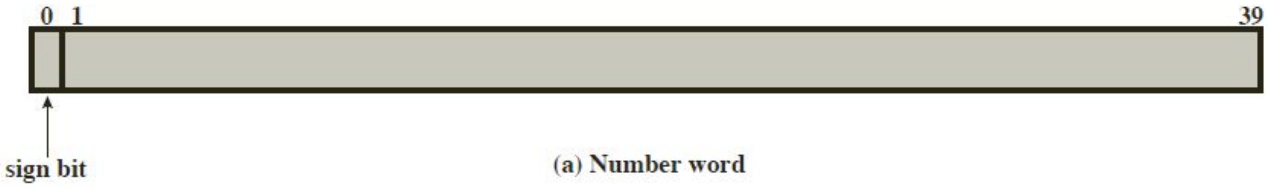
Structure of von Neumann machine


Central Processing Unit (CPU)
Main Memory (M)
Arithmetic-Logic Unit (CA)
Program Control Unit (CC)
I/O Equip-ment (I, O)

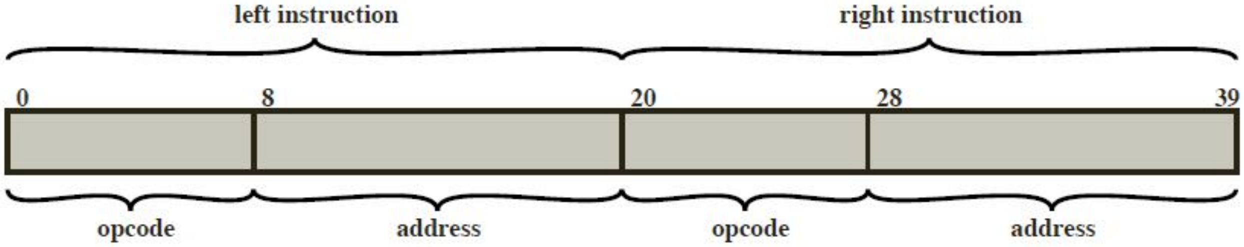
IAS - details

- The memory of the IAS consists of 1000 storage locations, called *words, of 40* binary digits (bits) each (1000 x 40 bit words).

- Both data and instructions are stored here.
- Each number is represented by a sign bit and a 39-bit value.
- A word may contain two 20-bit instructions (2 x 20 bit instructions) with each instruction consisting of an 8-bit operation code (opcode) specifying the operation to be performed and a 12-bit address designating one of the words in memory (numbered from 0 to 999).



(a) Number word

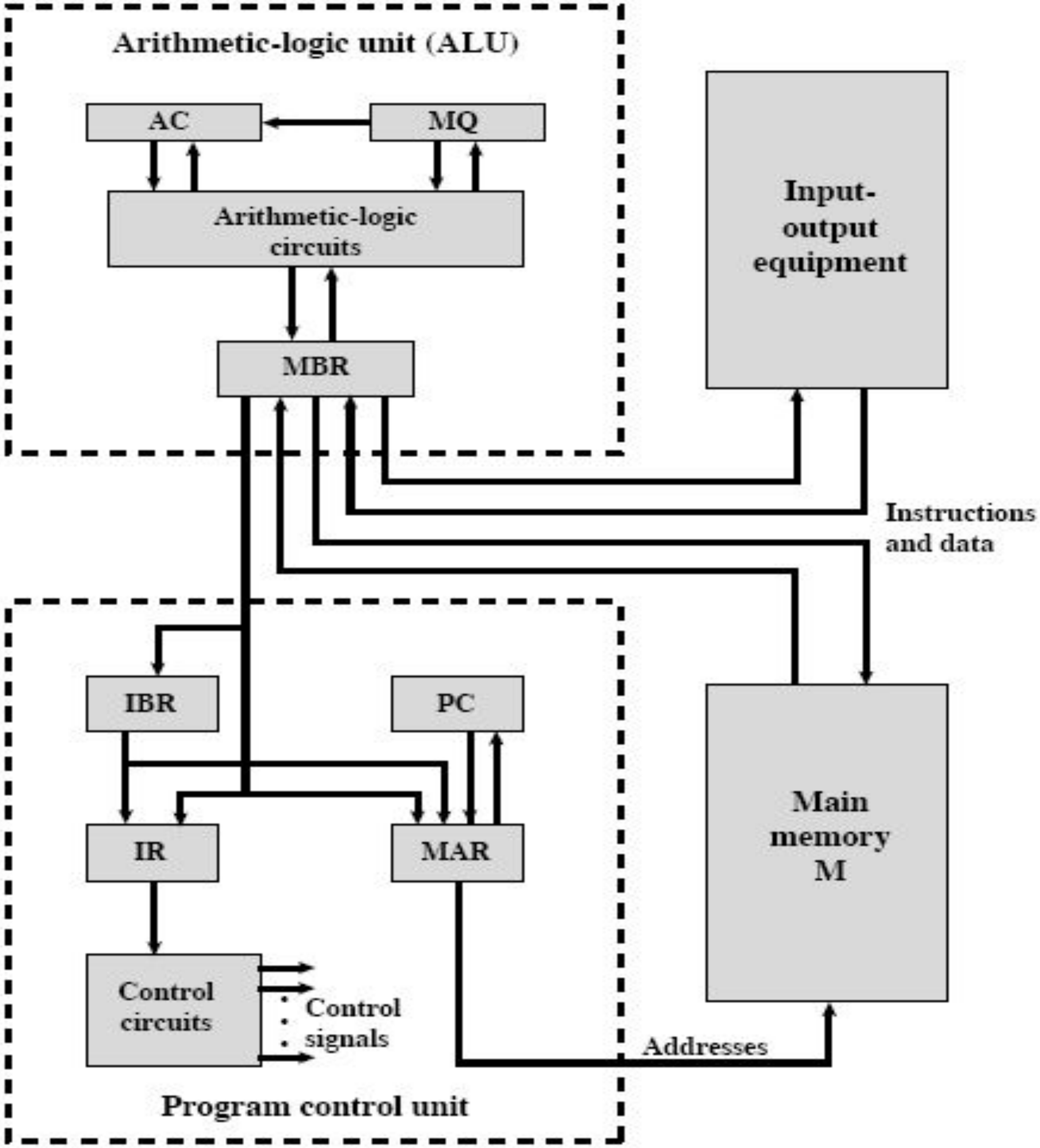


(b) Instruction word

IAS - details

- A more detailed structure diagram as follows:
- Both the control unit and the ALU contain storage locations, called *registers*
  - Memory Buffer Register (MBR): Contains a word to be stored in memory or sent to the I/O unit, or is used to receive a word from memory or from the I/O unit.
  - Memory Address Register (MAR): Specifies the address in memory of the word to be written into or read from the MBR.
  - Instruction Register (IR): Contains the 8-bit opcode instruction being executed.
  - Instruction Buffer Register (IBR): Employed to hold temporarily the right-hand instruction from a word in memory.
  - Program Counter (PC): Contains the address of the next instruction-pair to be fetched from memory.
  - Accumulator (AC) and Multiplier Quotient (MQ): Employed to hold temporarily operands and results of ALU operations, e.g. the result of multiplying two 40-bit numbers is an 80-bit number; the most significant 40 bits are stored in the AC and the least significant bit in the MQ.

Expanded Structure of IAS Computer



Commercial Computers

- In 1947 - Eckert-Mauchly Computer Corporation formed commercially.
- **First successful commercial computer was UNIVAC I** (Universal Automatic Computer), which was commissioned by the US Bureau of Census
- The Eckert-Mauchly Computer Corporation became part of Sperry-Rand Corporation
- Late 1950s - **UNIVAC II had greater memory capacity and higher performance than the UNIVAC I**.

IBM

- IBM, which was then the major manufacturer of Punched-card processing equipment, delivered its ***first electronic stored-program computer***, the **701**, in 1953.
- The 701 was intended primarily for scientific applications.
- In 1955, the IBM introduced **702** product, which had a number of hardware features that suited for business applications
- These were **the first of a long series of 700/7000 computers.**

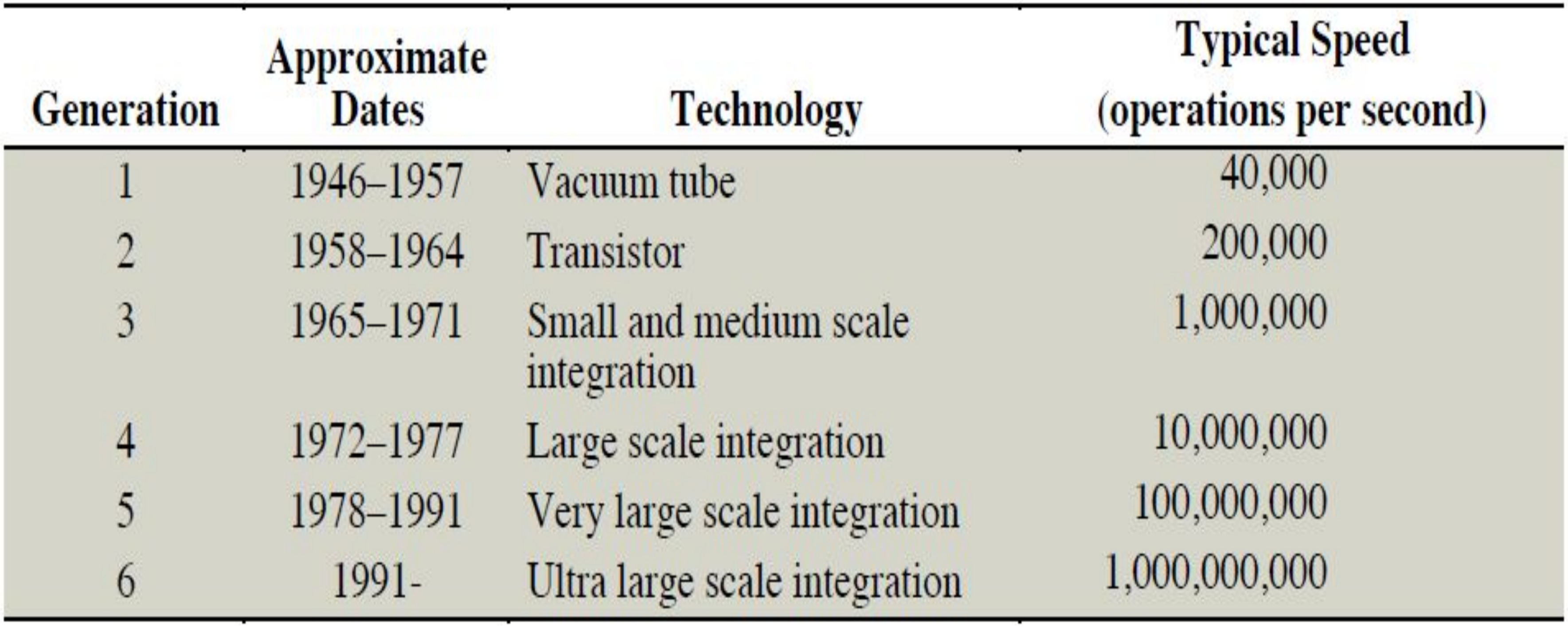
The Second Generations: Transistors

- Replacement of **vacuum tube by the transistor**
- The transistor is smaller, cheaper, and dissipates less heat
- **Transistor** is a semiconductor device is a semiconductor device used to amplify is a semiconductor device used to amplify or switch electronic signals.
- A transistor is made of a solid piece of a semiconductorA transistor is made of a solid piece of a semiconductor material, with at least three terminals for connection to an external circuit.
- A voltage or current applied to one pair of the transistor's terminals changes the current flowing through another pair of terminals.
- The transistor was invented at Bell Labs in 1947 and by the 1950s had launched an electronic revolution.
- Fully transistorized computes were commercially available, and IBM was again not the first company to deliver the new technology.
- IBM followed shortly with the 7000 series.

Transistor Based Computers

- The use of transistor defines the *second generation* of computers.
- Each new generation is characterized by greater processing performance, larger memory capacity, and smaller size than the previous one. (See Table->next slide)
- The second generation is important for the appearance of Digital Equipment Corporation (DEC) in 1957.
- DEC delivered its first computer, the PDP-1
- **Programmed Data Processor** (abbreviated *PDP*) was the name of a series of minicomputers) was the name of a series of minicomputers made by Digital Equipment Corporation.

Generations of Computer

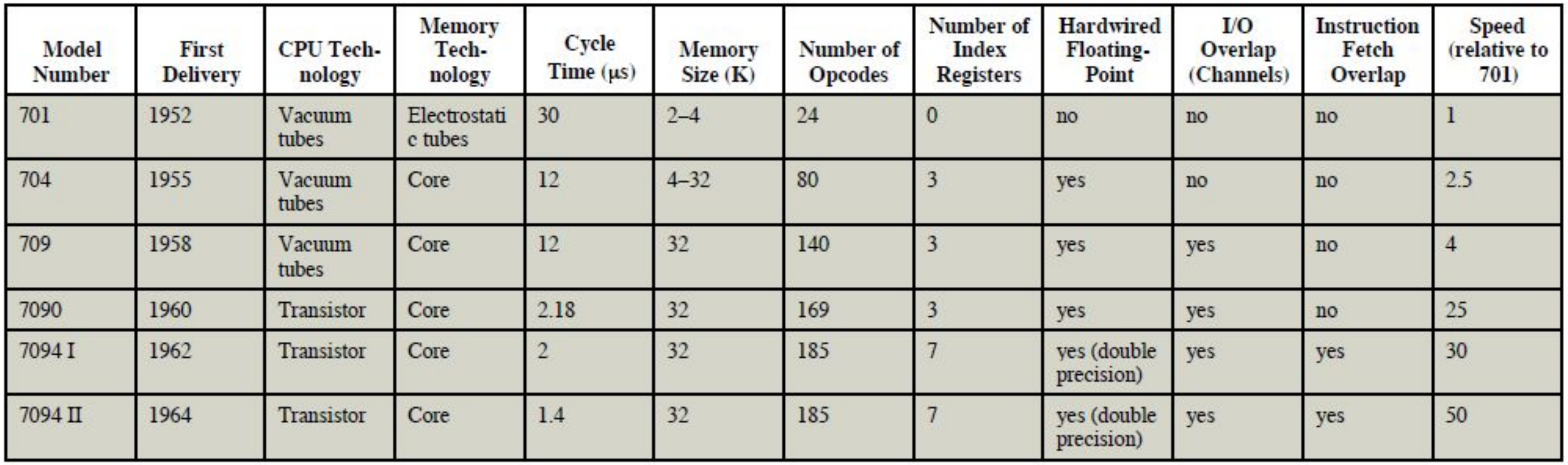


Generation	Approximate Dates	Technology	Typical Speed (operations per second)
1	1946–1957	Vacuum tube	40,000
2	1958–1964	Transistor	200,000
3	1965–1971	Small and medium scale integration	1,000,000
4	1972–1977	Large scale integration	10,000,000
5	1978–1991	Very large scale integration	100,000,000
6	1991-	Ultra large scale integration	1,000,000,000

IBM 7094

- From the introduction of the 700 series in 1952 to the introduction of the last member of the 700 series in 1964.
- Product line shows increased performance, increased capacity, and low cost.
- The size of the main memory grew from 2K to 32K words.
- The time to access one word of memory, the *memory cycle time*, fell from 30μs to 1.4μs.
- The number of opcodes grew from 24 to 185.
- Speed improvements are achieved by improved electronics (e.g., a transistor implementation is faster than a vacuum tube) and more complex circuitry.

Example members of the IBM 700/7000 Series



Model Number	First Delivery	CPU Technology	Memory Technology	Cycle Time (μs)	Memory Size (K)	Number of Opcodes	Number of Index Registers	Hardwired Floating-Point	I/O Overlap (Channels)	Instruction Fetch Overlap	Speed (relative to 701)
701	1952	Vacuum tubes	Electrostatic tubes	30	2–4	24	0	no	no	no	1
704	1955	Vacuum tubes	Core	12	4–32	80	3	yes	no	no	2.5
709	1958	Vacuum tubes	Core	12	32	140	3	yes	yes	no	4
7090	1960	Transistor	Core	2.18	32	169	3	yes	yes	no	25
7094 I	1962	Transistor	Core	2	32	185	7	yes (double precision)	yes	yes	30
7094 II	1964	Transistor	Core	1.4	32	185	7	yes (double precision)	yes	yes	50

Third Generation: Integrated Circuits

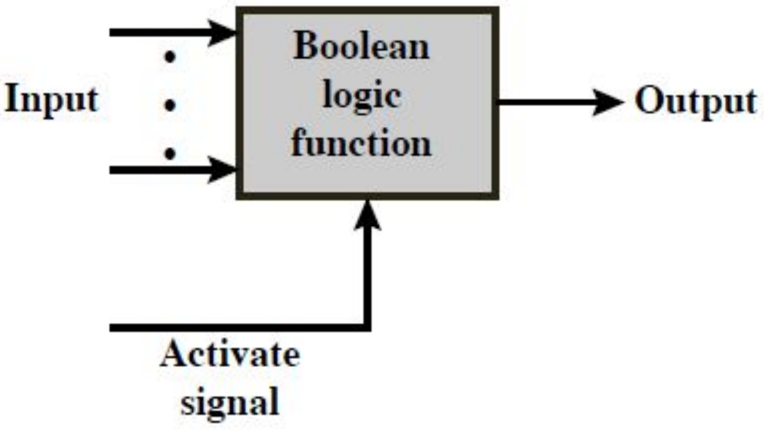
- A single, self-contained **transistor** is called a **discrete component** where **transistors are packaged individually**
- Throughout the 1950s and early 1960s, electronic equipment was composed largely of **discrete components – transistors, resistors, capacitors, and so on**
- Early second-generation computers contained about 10,000 transistors
- This figure grew to the hundreds of thousands, making the new manufacture, more powerful machines.
- In 1958, started the era of ***microelectronics*** and the invention of the integrated circuit (transistors are found in integrated circuits).
- It is the integrated circuits that defines the ***third generation of computers.***
- **Integrated circuit** also known as **IC**, **microcircuit**, **microchip**, **silicon chip**, or **chip**

Microelectronics

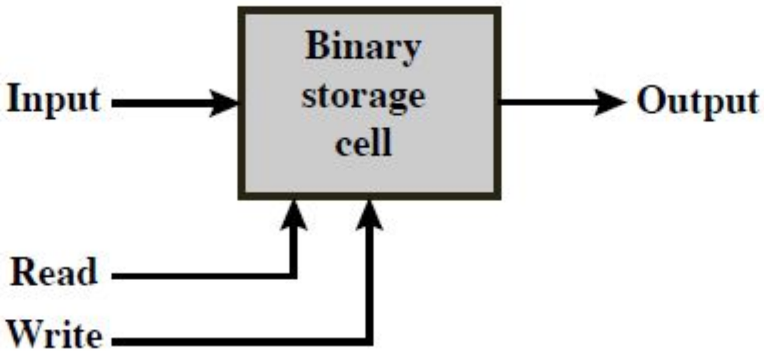
- Microelectronics means Literally - “small electronics”
- A computer is made up of **gates**, **memory cells**, and **interconnections**
- These can be manufactured on a semiconductor e.g., silicon wafer
- Trend towards the reduction in size in digital electronic circuits.

Microelectronics: Computer Elements

- Only two fundamental types of components are required: gates and memory cell
- A gate is a device that implements a simple Boolean or logical function, such as IF A AND B ARE TRUE THEN C IS TRUE (AND GATE)
- Such devices are called gates because they control data flow in the same way that canal gates do.
- The memory cell is a device that can store one bit of data that is, the device can be in one of two stable states at any time



(a) Gate



(b) Memory cell

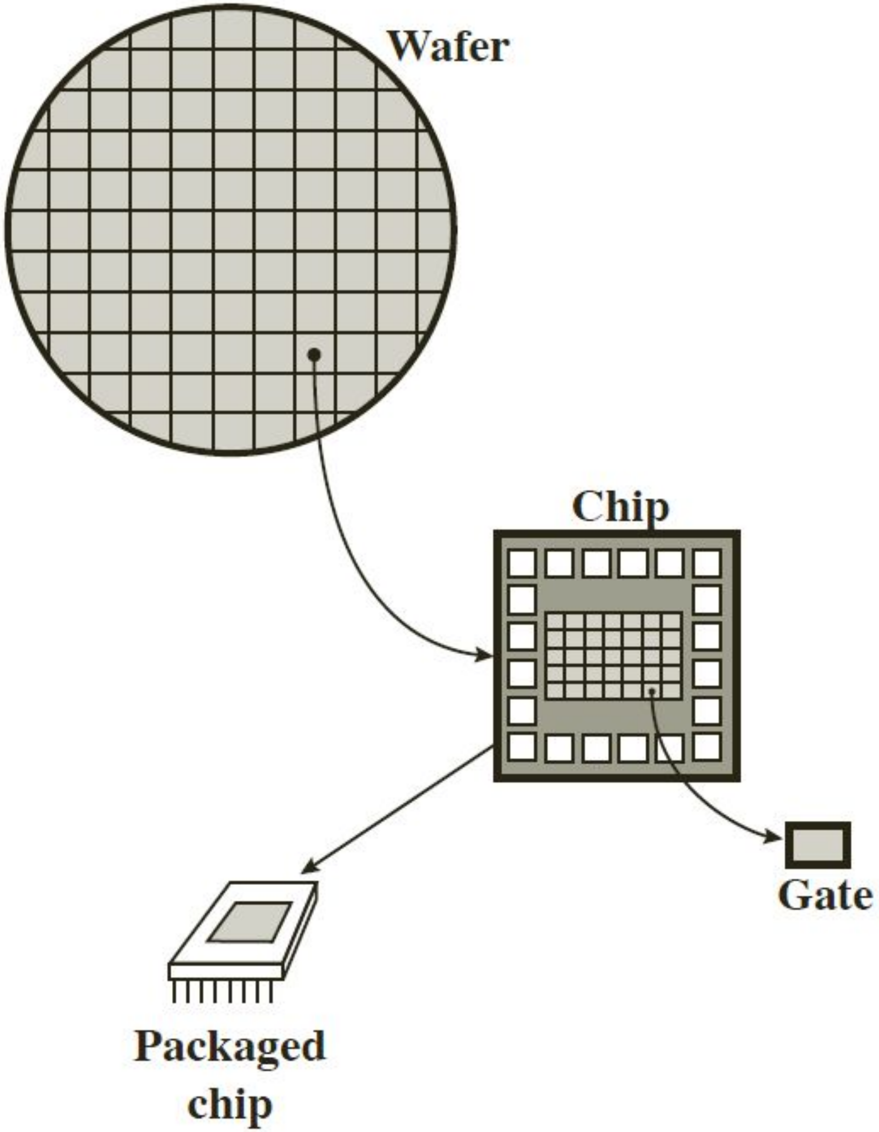
Microelectronics: Computer Elements

Four basic functions of a computer:

- **Data Storage**: Provided by memory cells
- **Data processing**: Provided by gates
- **Data movement**: The paths between components are used to move data from memory to memory and from gates to memory
- **Control**: A gate will have one or more data inputs plus a control signal input that activates the gate.
- When the control signal is ON, **the gate performs** its functions on the data inputs and produces a data output.
- Similarly, the memory cell will store the bit on its **input lead** **when the WRITE control signal is ON** and
- the memory cell will place the bit on its **output lead** **when the READ control signal is ON**.

Integrated Circuit: Wafer, Chip, and Gate

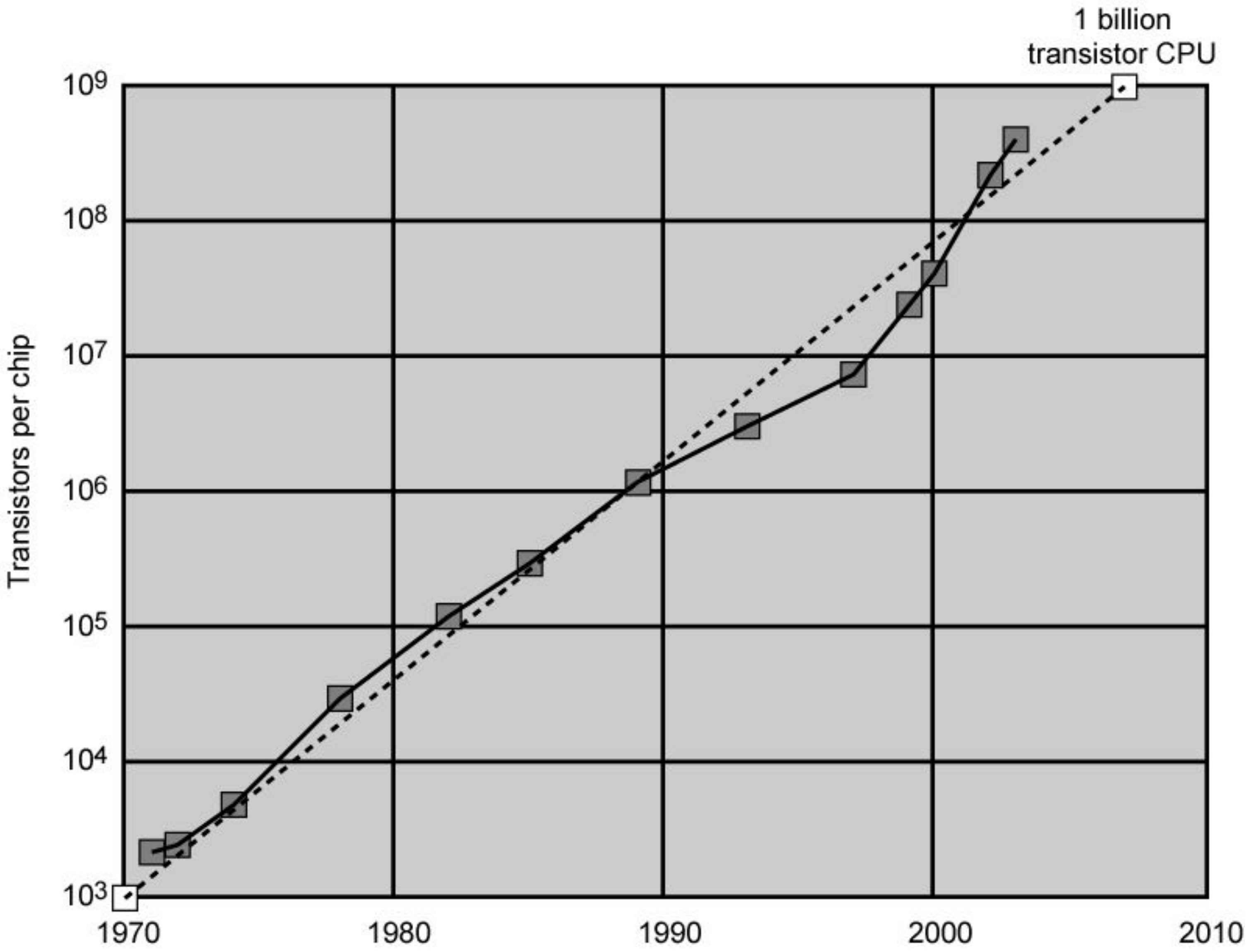
- A thin ***wafer*** of silicon is divided into a matrix of small areas, each a few millimeters square.
- The identical circuit pattern is fabricated in each area, and the wafer is broken up into ***chips***.
- Each chip consists of many ***gates*** and/or ***memory*** cells plus a number of input and output attachment points.
- The chip is then ***packaged*** that protects it and provides ***pins*** for attachment to devices beyond the chip.
- A number of these packages can then be interconnected on a printed circuit board to produce larger and more complex circuits.



Moore’s Law

- Initially, only a **few gates or memory cells** could be reliably manufactured and packaged together; these early integrated circuits are referred to as **small-scale integration (SSI).**
- It became possible to pack more and more components on the same chip.
- **Increased density of components** on chip
- Moore’s law by Gordon Moore – co-founder of Intel
- Number of transistors on a chip will double every year
- Since 1970’s development has slowed a little
  - Number of transistors doubles every 18 months
- Cost of a chip has remained almost unchanged
- **Higher packing density means shorter electrical paths, giving higher performance**
- **The computer becomes smaller**, making it more convenient to place in a variety of environments.
- **There is a reduction of power and cooling requirements**
- The interconnections on the integrated circuit are much more reliable. With more circuitry on each chip, there are fewer interchip connections.

Growth in CPU Transistor Count



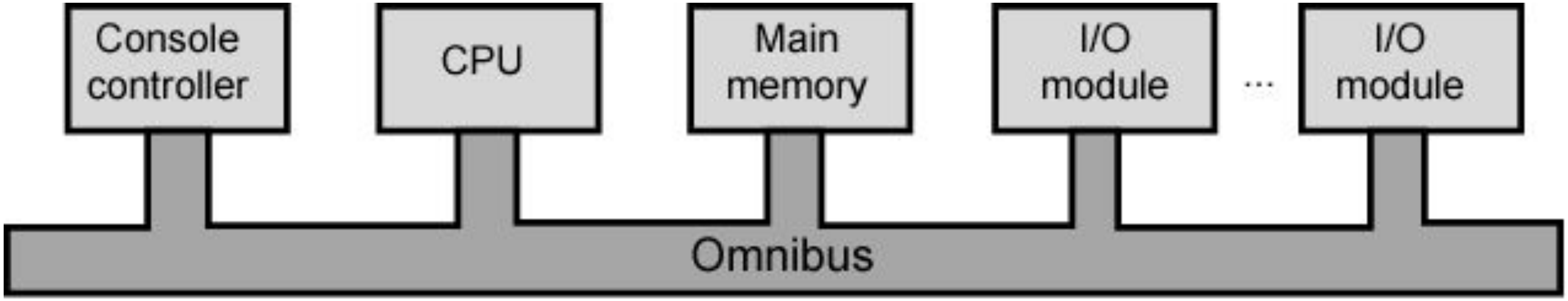
IBM System/ 360 series

- By 1964
- Replaced 7000 series
- The industry’s first planned “family” of computers
- The characteristics of a family are as follows:
  - Similar or identical instruction sets
  - Similar or identical O/S
  - **Increasing speed** (in going from lower to higher family members)
  - Increasing number of I/O ports (in going .....members)
  - **Increased memory size** (in going .... members)
  - Increased cost (in going .... members)
- Multiplexed switch structure

DEC PDP-8

- PDP-8 from Digital Equipment Corporation (DEC) in 1964
- **First minicomputer**
- Did not need air conditioned room
- Small enough to sit on a lab bench
- Cost was $16,000
- Embedded applications & OEM (Original Equipment Manufacturers)

DEC - PDP-8 Bus Structure



- **PDP-8 bus, called the Omnibus, consists of 96 separate signal paths, used to carry control, address, and data signals.**
- All system components share a common set of signal paths, their use **must be controlled by the CPU**.
- This architecture is highly flexible, allowing modules to be plugged into the bus to create various configurations.

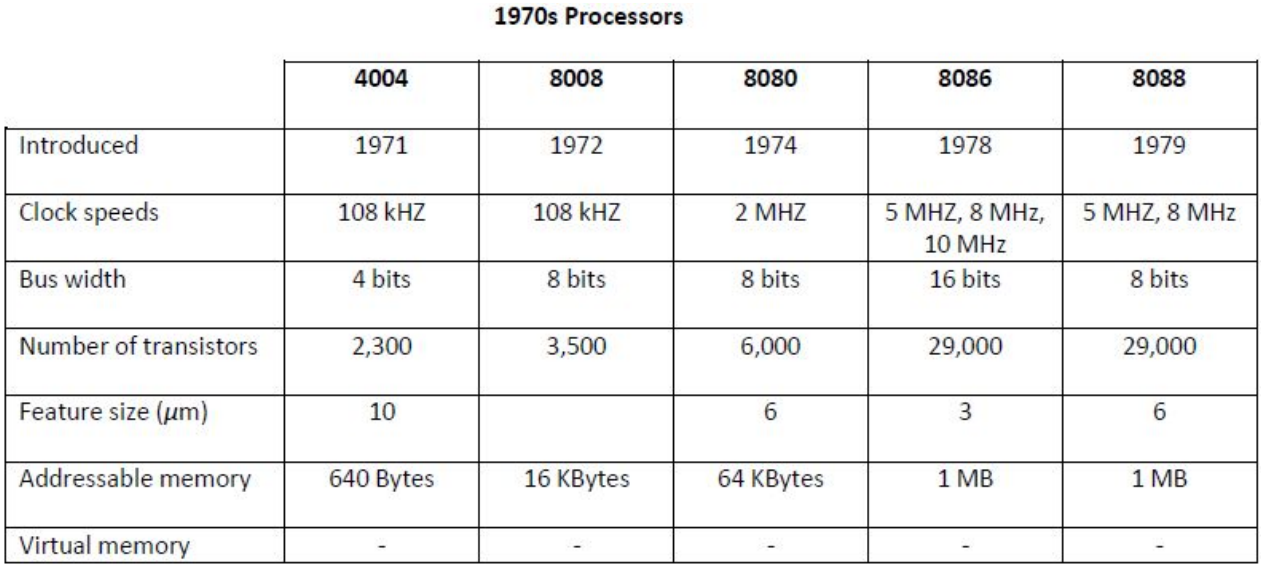
Later Generations

- For large scale integration (LSI), more than ***1000 components*** can be placed on a single integrated circuit chip.
- For very large scale integration (VLSI), more than ***10,000 components*** per chip.
- For ultra large scale integration (ULSI), more than ***one million components***.

Semiconductor Memory

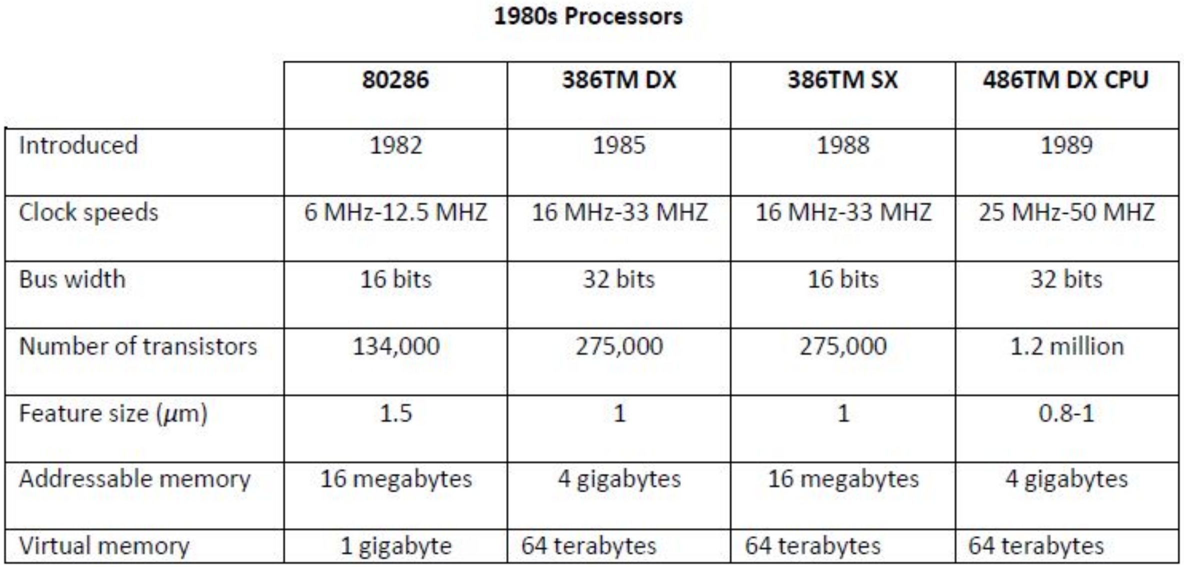
- In 1970, Fairchild produced the **first relatively capacious semiconductor memory**.
- This chip, could hold 256 bits of memory.
- It was much **faster than core**
- It took only 70 billionths of a second to read a bit.
- However, cost per bit was higher than for that of core.
- Since 1970, semiconductor memory has been through 11 generations: **1K, 4K, 16K, 64K, 256K, 1M, 4M, 16M, 64M, 256M, and 1G.**
- **Each generation has provided four times the storage density of the previous generations, accompanied by declining cost per bit and declining access time**.

Evaluation of Intel Microprocessors



1970s Processors

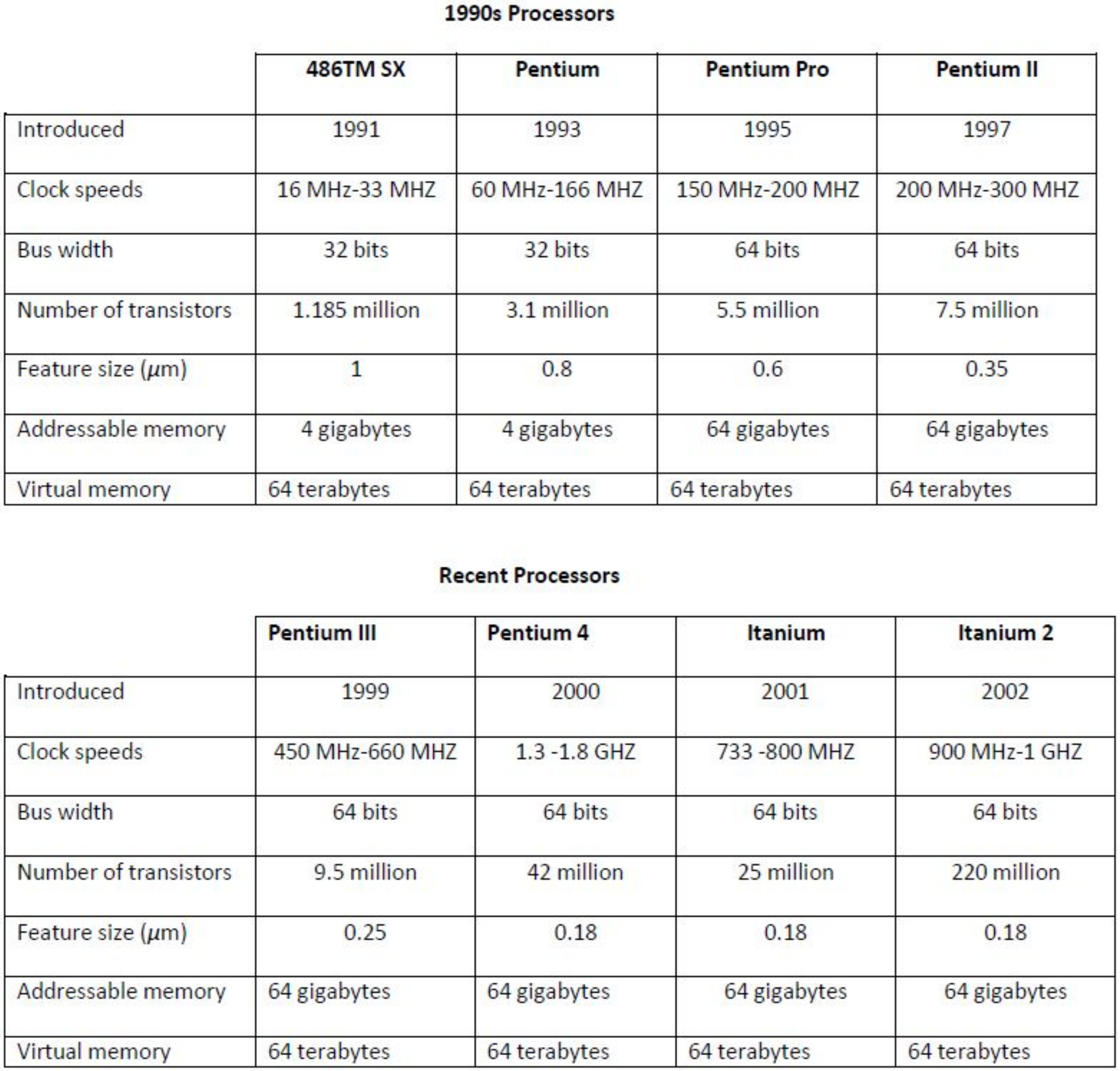
	4004	8008	8080	8086	8088
Introduced	1971	1972	1974	1978	1979
Clock speeds	108 kHZ	108 kHZ	2 MHZ	5 MHZ, 8 MHz, 10 MHz	5 MHZ, 8 MHz
Bus width	4 bits	8 bits	8 bits	16 bits	8 bits
Number of transistors	2,300	3,500	6,000	29,000	29,000
Feature size (μm)	10		6	3	6
Addressable memory	640 Bytes	16 KBytes	64 KBytes	1 MB	1 MB
Virtual memory	-	-	-	-	-



1980s Processors

	80286	386TM DX	386TM SX	486TM DX CPU
Introduced	1982	1985	1988	1989
Clock speeds	6 MHz-12.5 MHZ	16 MHz-33 MHZ	16 MHz-33 MHZ	25 MHz-50 MHZ
Bus width	16 bits	32 bits	16 bits	32 bits
Number of transistors	134,000	275,000	275,000	1.2 million
Feature size (μm)	1.5	1	1	0.8-1
Addressable memory	16 megabytes	4 gigabytes	16 megabytes	4 gigabytes
Virtual memory	1 gigabyte	64 terabytes	64 terabytes	64 terabytes

Evaluation of Intel Microprocessors



1990s Processors

	486TM SX	Pentium	Pentium Pro	Pentium II
Introduced	1991	1993	1995	1997
Clock speeds	16 MHz-33 MHZ	60 MHz-166 MHZ	150 MHz-200 MHZ	200 MHz-300 MHZ
Bus width	32 bits	32 bits	64 bits	64 bits
Number of transistors	1.185 million	3.1 million	5.5 million	7.5 million
Feature size (μm)	1	0.8	0.6	0.35
Addressable memory	4 gigabytes	4 gigabytes	64 gigabytes	64 gigabytes
Virtual memory	64 terabytes	64 terabytes	64 terabytes	64 terabytes

Recent Processors

	Pentium III	Pentium 4	Itanium	Itanium 2
Introduced	1999	2000	2001	2002
Clock speeds	450 MHz-660 MHZ	1.3 -1.8 GHZ	733 -800 MHZ	900 MHz-1 GHZ
Bus width	64 bits	64 bits	64 bits	64 bits
Number of transistors	9.5 million	42 million	25 million	220 million
Feature size (μm)	0.25	0.18	0.18	0.18
Addressable memory	64 gigabytes	64 gigabytes	64 gigabytes	64 gigabytes
Virtual memory	64 terabytes	64 terabytes	64 terabytes	64 terabytes

Microprocessor Speed

- **Pipelining:** A processor organization in which the processor consists of a number of stages, **allowing multiple instructions to be executed concurrently**.
- **On board cache:** A special buffer storage, smaller and faster than main storage, that is used to contain a copy of instructions and data from main memory that are to be needed next by the processor (to reduce the average time to access the memory).
- **On board L1 & L2 cache**
- **Branch prediction:** A mechanism used by the processor to predict which branches or groups of instructions are likely to be processed next.

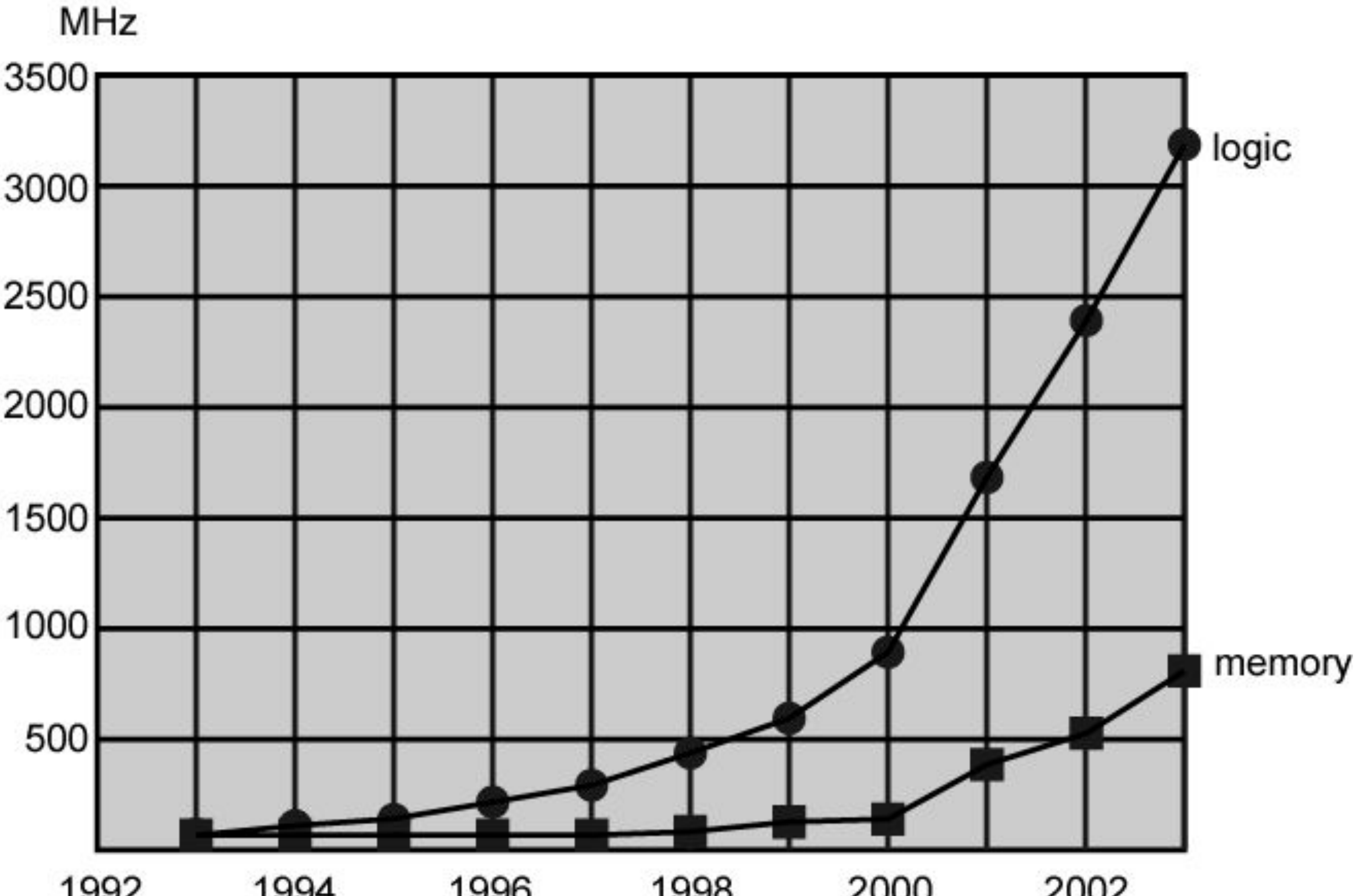
Microprocessor Speed

- **Data flow analysis:** The processor analyses which instructions are dependent on each other’s results, or data, to create an optimized schedule of instructions. Instructions are scheduled to be executed when ready, independent of the original program order. This prevents unnecessary delay.
- **Speculative execution:** Using branch prediction and data flow analysis, some processors speculatively execute instructions ahead of their actual appearance in the program execution, holding the results in temporary locations. This enables the processor to keep **its execution engines as busy as possible** by executing instructions that likely to be needed.

Performance Balance

- Processor speed increased
- Memory capacity increased
- Memory speed lags behind processor speed

Logic and Memory Performance Gap



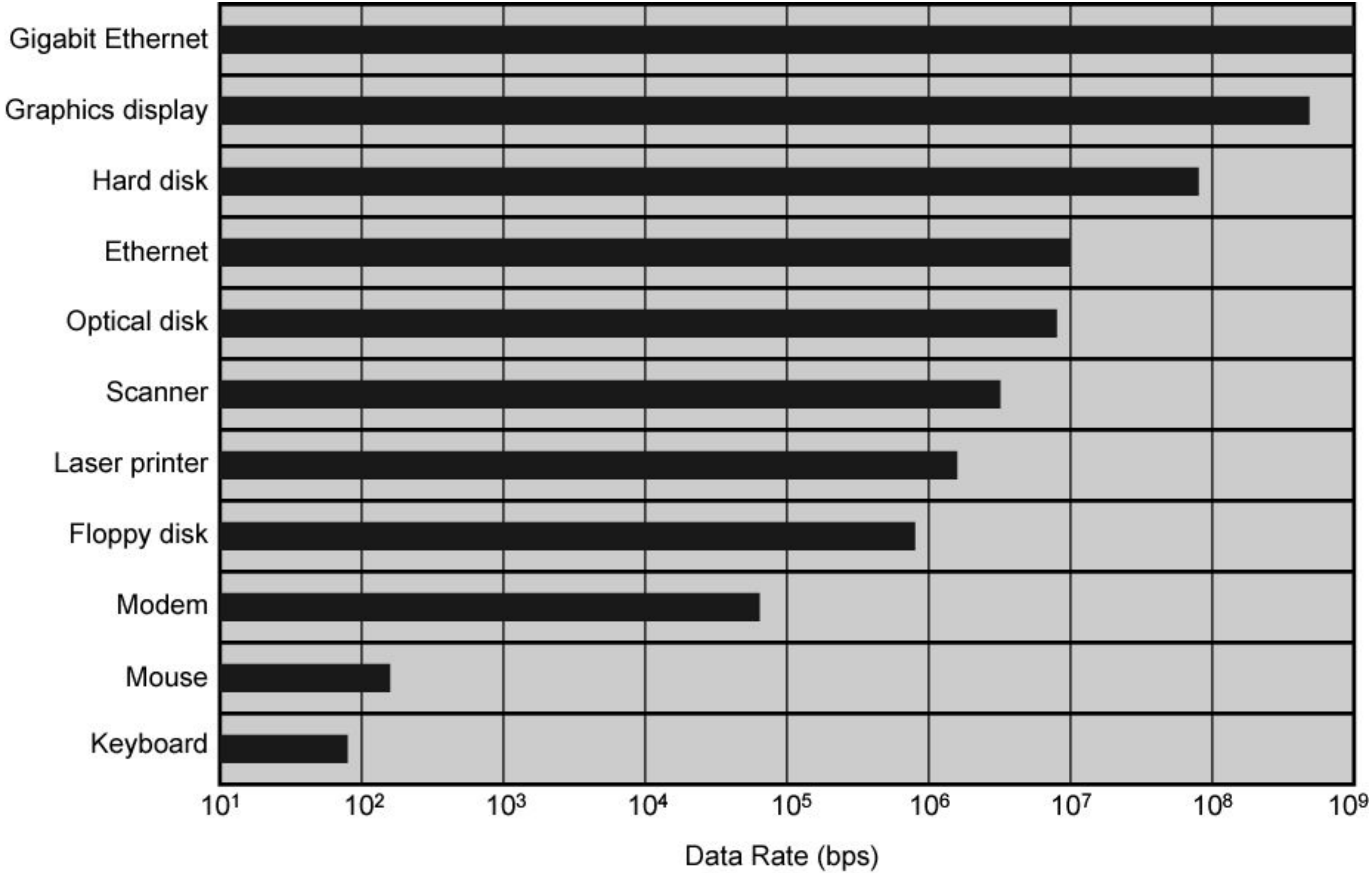
Solutions

- Increase the number of bits retrieved at one time by making DRAM (Dynamic Random Access memory) "wider" rather than "deeper" and by using wide bus data paths
- Change the DRAM interface
  - Include "Cache" on the DRAM chip
- Reduce the frequency of memory access by incorporating
  - More complex and efficient cache between the processor and main memory.
- Increase the interconnection bandwidth between processors and memory by using
  - High speed buses

Handling of I/O Devices

- Computers become faster and more capable, more sophisticated applications are developed that support the use of peripherals with intensive I/O demands
- **I/O devices create large data throughput demands**
- **Processors can handle the data pumped out by these I/O devices**
- **Problem is getting data moved between processor and peripheral.**
- Solutions to satisfy I/O demands:
  - Caching
  - Buffering
  - Higher-speed interconnection buses
  - More elaborate bus structures
  - Multiple-processor configurations

Typical I/O Device Data Rates



Key is Balance

Designers constantly strive to balance the throughput and processing demands of the:

- **Processor components**
- **Main memory**
- **I/O devices**
- **Interconnection structures**

The idea is to cope with the following two constantly evolving factors:

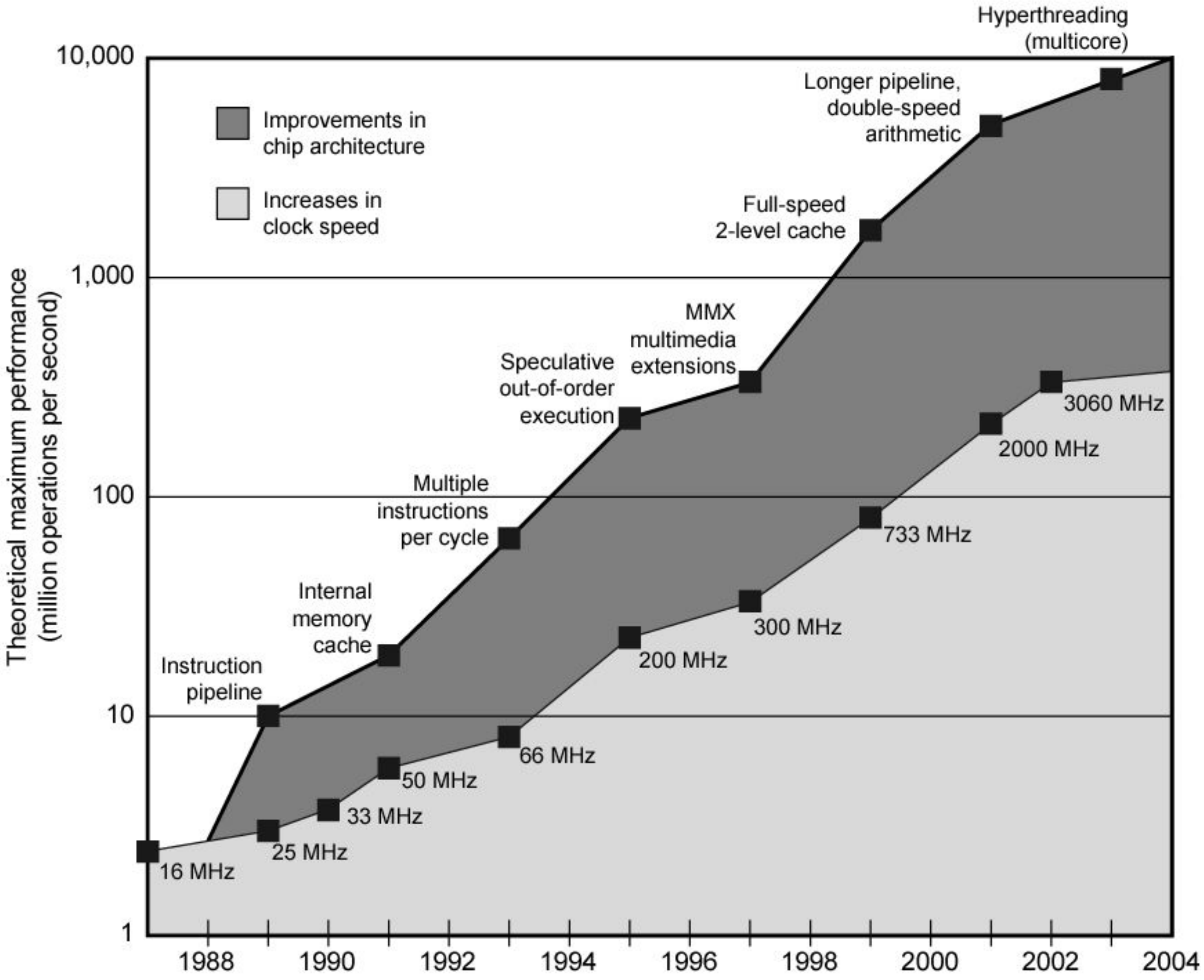
- ❑ The rate at which performance is changing in the various technology areas (processors, buses, memory, peripherals) differs greatly from one type of element to another
- ❑ New applications and new peripheral devices constantly change the nature of the demand

Improvements in Chip Organization and Architecture

- **Increase speed of processor**
  - Fundamentally due to shrinking logic gate size
  - **More gates, packed more tightly, increasing clock rate**
  - With gates closer together, **propagation time for signals reduced, that speeds up the processor**)

- **Increase size and speed of caches between processors and main memory**
  - Cache access times drop significantly
- **Change to processor organization and architecture that**
  - Increase the effective speed of instruction execution
  - **Parallelism**

Intel Microprocessor Performance



Strategies to increase performance – Since 1980

Firstly, Increased Cache Capacity

- Typically two or three levels of cache between processor and main memory
- Chip density increased
  - More cache memory on chip
  - Faster cache access
- Pentium chip devoted about 10% of on-chip area to a cache
- Pentium 4 devotes about 50%

Strategies to increase performance – Since 1980

Secondly, More Complex Execution Logic

- Enable parallel execution of instructions
- Pipeline works:
  - Different stages of execution of **different instructions at same time along pipeline**
- Superscalar allows multiple pipelines within single processor
  - Instructions that do not depend on one another can be executed in parallel

Pentium Evolution (1)

- 8080
  - The world’s first general purpose microprocessor
  - 8 bit machine, 8-bit data path and addresses 1MB of memory using 20-bit address bus
  - Used in first personal computer – Altair
- 8086
  - much more powerful,16 bit machine and 16-bit data path and addresses 1 MB of memory using 20-bit address bus
  - instruction cache, prefetch few instructions before they are executed
- 80286
  - 16 bit machine, 16-bit data path and addresses 16MB of memory using 24-bit address bus
- 80386
  - 32 bit machine, 32-bit data path
  - Support for multitasking

**High speed memory internal to the CPU*

Pentium Evolution (2)

- 80486
  - Sophisticated and powerful cache and sophisticated instruction pipelining
  - built in math co-processor, relieving of complex math operations from the main CPU
- Pentium
  - Intel introduces the Superscalar technology
  - Allows multiple instructions executed in parallel
- Pentium Pro
  - Increased superscalar organization
  - Aggressive use of register renaming
  - branch prediction
  - data flow analysis
  - speculative execution

Pentium Evolution (3)

- Pentium II
  - Incorporated Intel MMX (Multimedia Extension) technology
  - designed to process graphics, video & audio data efficiently
- Pentium III
  - Incorporates additional floating point instructions for 3D graphics software
- Pentium 4
  - Incorporated additional floating point and multimedia enhancements
- Itanium
  - 64 bit
- Itanium 2
  - Hardware enhancements to increase speed

PowerPC Family (1)

- The following are the principal members of the PowerPC family:
- 601:
  - 32-bit machine, quickly to market
- 603:
  - 32-bit, Low-end desktop and portable
  - Comparable machine in performance with 601
  - Lower cost and more efficient implementation
- 604:
  - 32-bit machine, Desktop and low-end servers
  - Much more advanced superscalar design
  - Greater performance
- 620:
  - 64-bit architecture, including 64-bit registers and data paths
  - Desktop and low-end servers

PowerPC Family (2)

- 740/750:
  - Also known as G3 processor
  - Integrates two levels of cache on main processor chip
- G4:
  - Increases parallelism and internal speed of the processor chip
- G5:
  - Improvements in parallelism and internal speed
  - 64-bit organization

Key Points

- The evaluation of computers has been characterized by increasing processor speed, decreasing computer size, increasing memory size, and increasing I/O capacity and speed.
- One factor responsible for the great increase in processor speed is the shrinking size of microprocessor components; this reduces the distance between components and hence increases speed.
- However, the true gains in speed in recent years have come from pipelining, parallel execution techniques, and use of speculative execution techniques. The reason is keep the processor busy as much of the time as possible.
- A critical issue in computer system design is balancing the performance. As we see, the processor speed has increased more rapidly than memory access time. A variety of techniques is used to compensate for this mismatch, including caches, wider data paths from memory to processor, and more intelligent memory chips.