BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Computer Organization and Software Systems

CONTACT SESSION 2

Mr. Vaibhav Jain

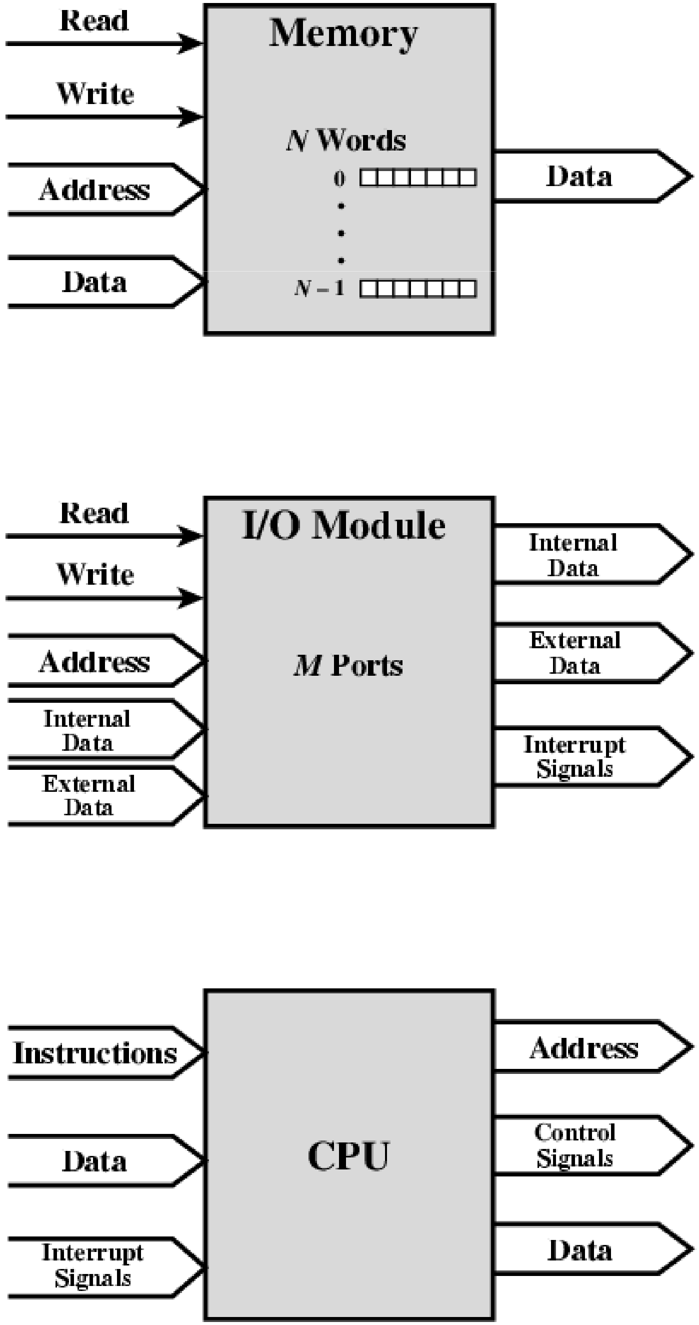
WILP – BITS Pilani

Computer Organization and Software Systems (SS ZG516)

Session _2

- Interconnection Structures
- Bus Interconnection
- Performance Assessment: CPI, MIPS Rate
- Amdahl’s Law
- Computer Memory System Overview

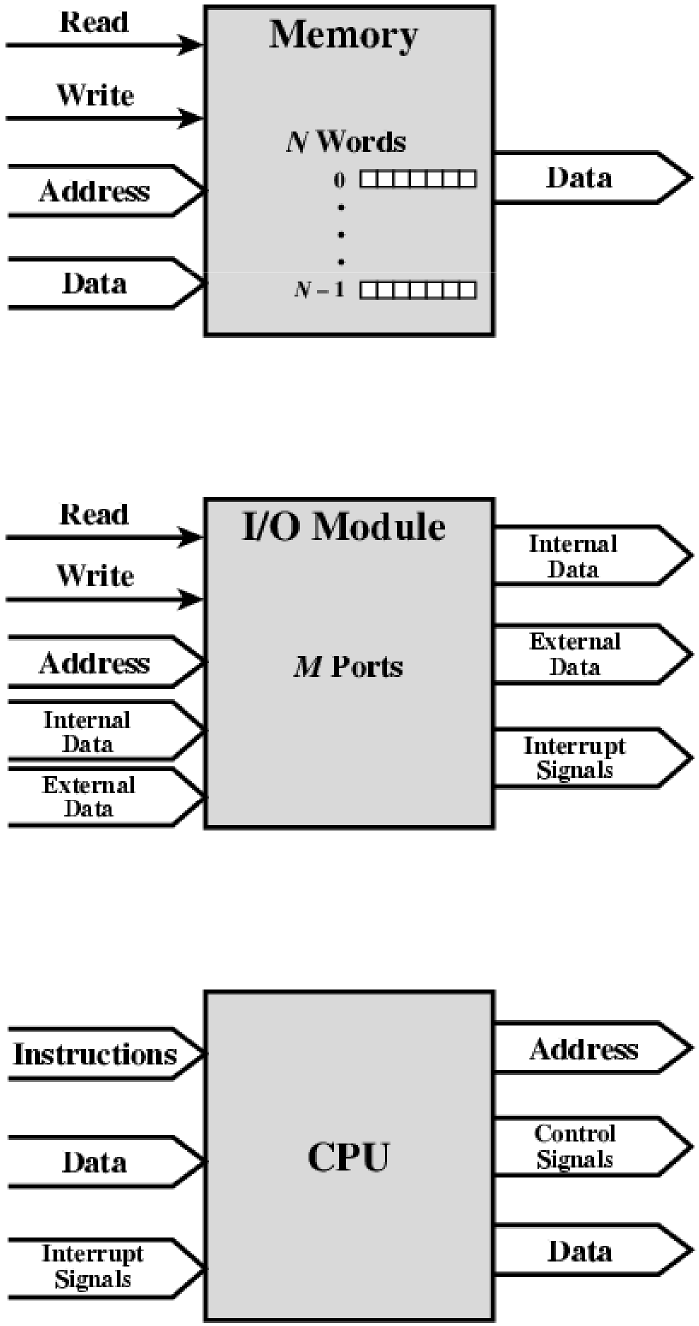
Computer Modules



Memory Connection

- Receives and sends data
- Receives addresses (of locations)
- Receives control signals
  - Read
  - Write
  - Timing

Computer Modules



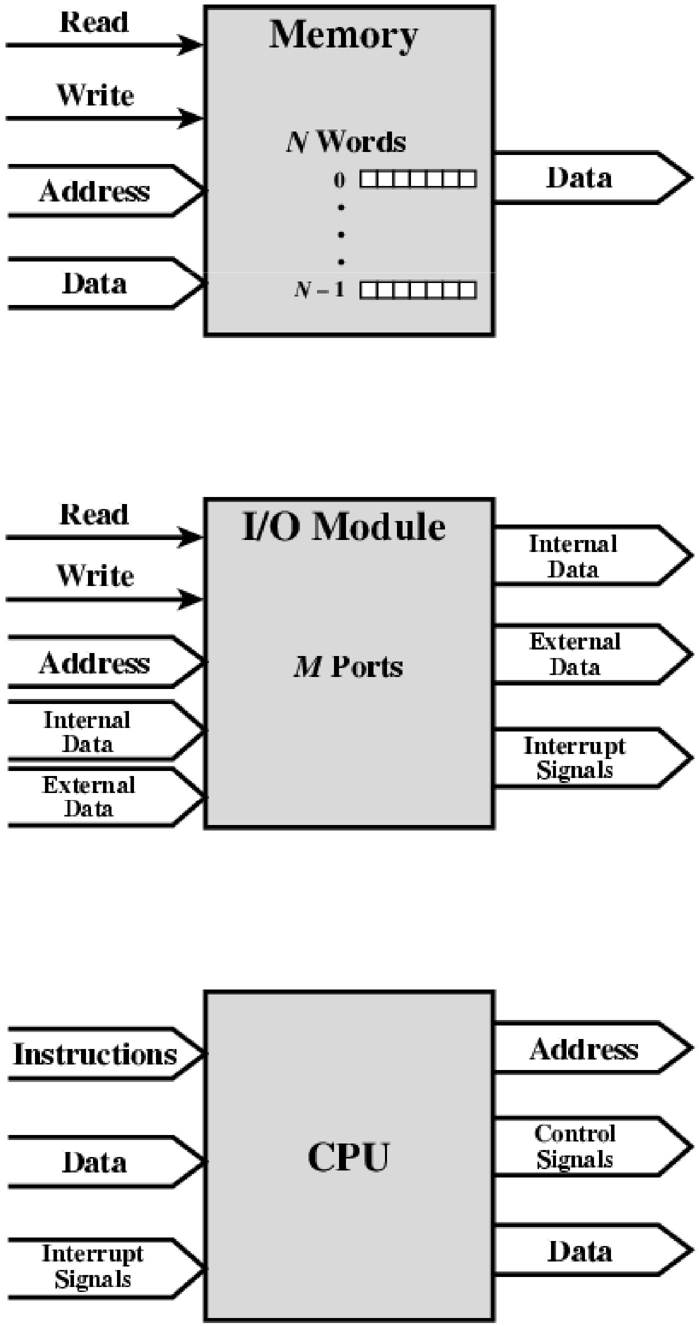
Input/Output Connection(1)

- Similar to memory from computer's viewpoint
- Output
  - Receive data from computer
  - Send data to peripheral
- Input
  - Receive data from peripheral
  - Send data to computer

Input/Output Connection(2)

- Receive control signals from computer
- Send control signals to peripherals
  - e.g. spin disk
- Receive addresses from computer
  - e.g. port number to identify peripheral
- Send interrupt signals (control)

Computer Modules



CPU Connection

- Reads instruction and data
- Writes out data (after processing)
- Sends control signals to other units
- Receives (& acts on) interrupts

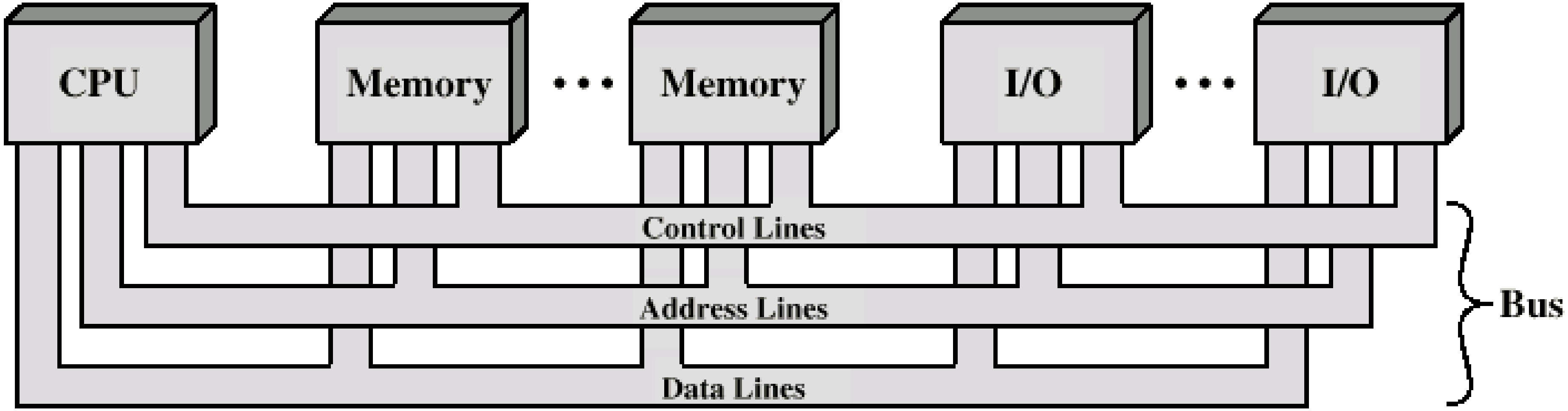
Interconnection Structures

- Three modules of any system
  - Processor, memory, I/O
  - The interconnection structure must support the following types of transfers:
    - Memory to processor
    - Processor to memory
    - I/O to processor
    - Processor to I/O
    - I/O to or from memory

What is a Bus?

- A communication pathway connecting two or more devices
- Shared transmission medium
- Broadcast (*signal transmitted by any one device is available for reception by all other devices attached to the bus*)
- Often grouped
  - A number of channels in one bus
  - e.g. 32 bit data bus is 32 separate single bit channels
- A bus that connects major computer components (processor, memory, I/O) is called a system bus.

Bus Interconnection Scheme



Data Bus

- Carries data
  - Remember that there is no difference between “data” and “instruction” at this level
- Width is the number of lines
  - 8, 16, 32, 64 bit
- Each line can carry only 1 bit at a time.

Address bus

- Identify the source or destination of data
- e.g. CPU needs to read an instruction (data) from a given location in memory
- Bus width determines maximum memory capacity of system
  - e.g. 8080 has 16 bit address bus giving 64k address space
  - $2^{16} = 2^{10} \times 2^{6}$

    $= 2^{6} \times 2^{10}$

    $= 64\text{ k}$

Control Bus

- The data and address lines are shared by all components, there must be a means of controlling their use.
- The various control lines are –
  - Memory read/write signal
  - I/O read/write signal
  - Acknowledgement
  - Interrupt request
  - Clock signals
  - Reset Signal

Control Bus – Control Lines

- **Memory write**: Causes data on the bus to be written into the addressed location
- **Memory read**: Causes data from the addressed location to be placed on the bus
- **I/O write**: Causes data on the bus to be output to the addressed I/O port
- **I/O read**: Causes data from the addressed I/O port to be placed on the bus
- **Transfer ACK**: Indicates that data have been accepted from or placed on the bus

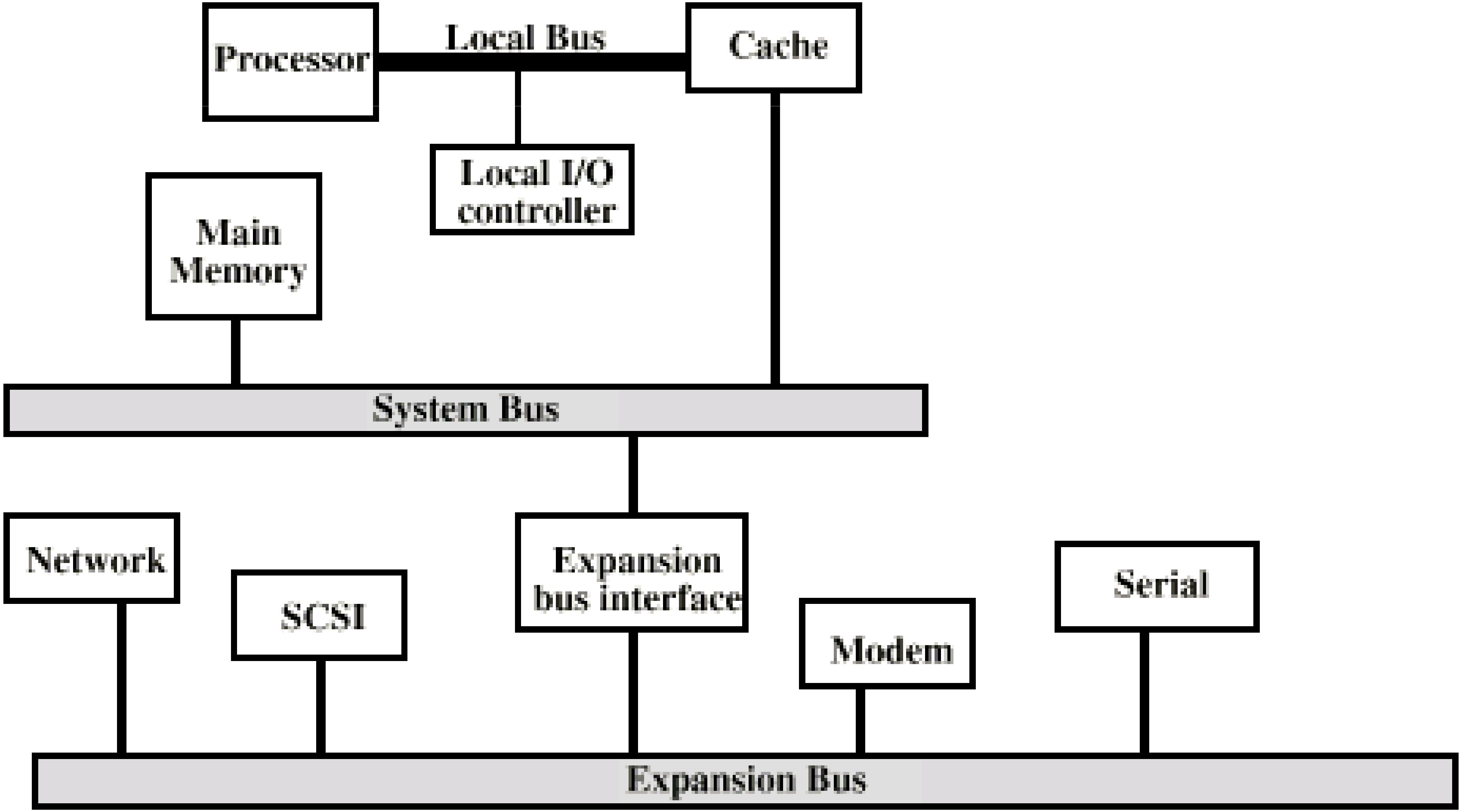
Control Bus – Control Lines

- **Bus request**: Indicates that a module needs to gain control of the bus
- **Bus grant**: Indicates that a requesting module has been granted control of the bus
- **Interrupt request**: Indicates that an interrupt is pending
- **Interrupt ACK**: Acknowledges that the pending interrupt has been recognized
- **Clock**: Is used to synchronize operations
- **Reset**: Initializes all modules

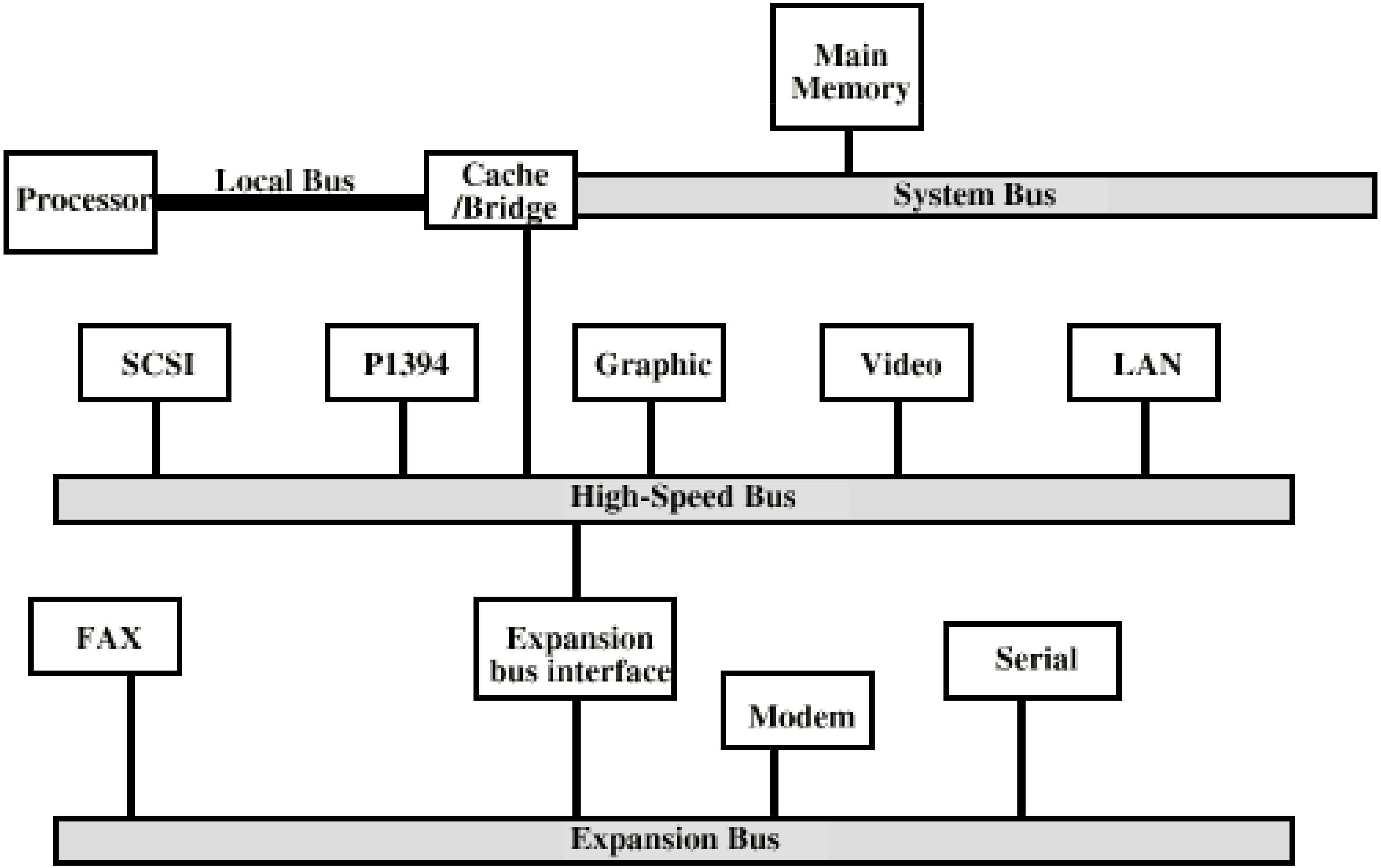
Single Bus Problems

- Lots of devices on one bus leads to:
  - Propagation delays
    - More devices, the greater the bus length and hence the greater the propagation delay. When control of the bus passes from one device to another frequently, these propagation delays can noticeably affect performance.
  - Data transfer demand
    - The bus may become a bottleneck as the aggregate data transfer demand approaches the capacity of the bus. This problem can be countered to some extent by increasing the data rate that the bus can carry and by using wider buses (e.g., increasing the data bus from 32 to 64 bits).
- Most systems use multiple buses to overcome these problems

Traditional (ISA – *Industry Standard Architecture*) (with cache)



High Performance Bus



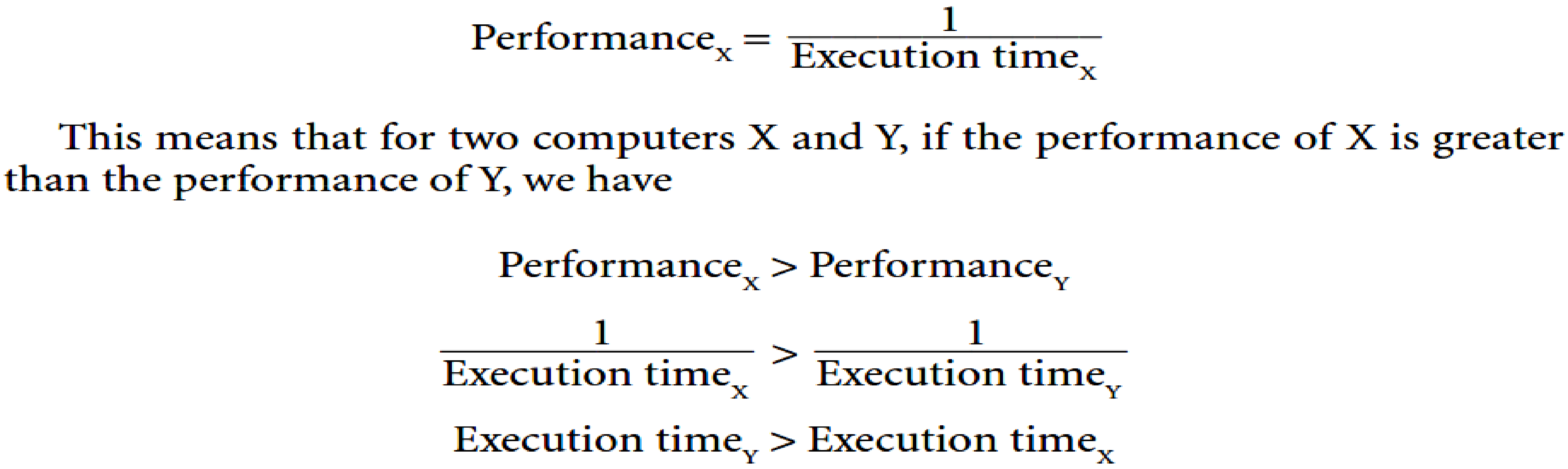
Performance Assessment

- If you were running a program on two different desktop computers, you’d say that the faster one is the desktop computer that gets the job done first.

- If you were running a datacenter that had several servers running jobs submitted by many users, you’d say that the faster computer was the one that completed the most jobs during a day.

- As an individual computer user, you are interested in reducing response time—the time between the start and completion of a task—also referred to as execution time.

Performance Assessment – Clock Speed and Instructions per Second



$$\text{Performance}_X = \frac{1}{\text{Execution time}_X}$$

This means that for two computers X and Y, if the performance of X is greater than the performance of Y, we have

$$\text{Performance}_X > \text{Performance}_Y$$

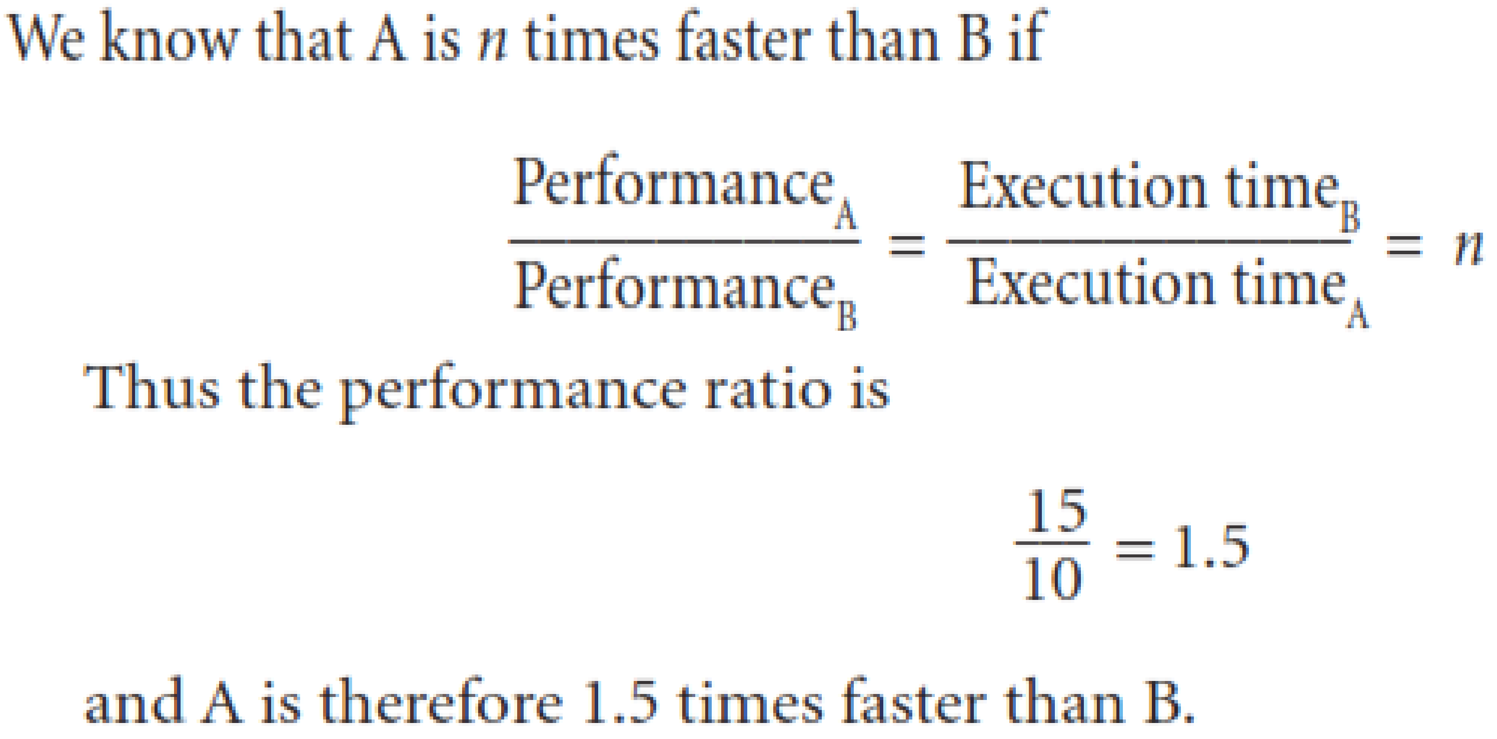
$$\frac{1}{\text{Execution time}_X} > \frac{1}{\text{Execution time}_Y}$$

$$\text{Execution time}_Y > \text{Execution time}_X$$

- **Almost all computers are constructed using a clock that determines when events take place in the hardware.**
- **These discrete time intervals are called clock cycles.**
- **The rate of pulses is known as the *clock rate*, or clock speed. (e.g., 4 gigahertz, or 4 GHz), which is the inverse of the clock period.**

Questions -

- If computer A runs a program in 10 seconds and computer B runs the same program in 15 seconds, how much faster is A than B?.



We know that A is n times faster than B if

$$\frac{\text{Performance}_A}{\text{Performance}_B} = \frac{\text{Execution time}_B}{\text{Execution time}_A} = n$$

Thus the performance ratio is

$$\frac{15}{10} = 1.5$$

and A is therefore 1.5 times faster than B.

Performance Assessment

CPU Performance and Its Factor

$$\begin{matrix}\text{CPU execution time} \\ \text{for a program}\end{matrix} = \begin{matrix}\text{CPU clock cycles} \\ \text{for a program}\end{matrix} \times \text{Clock cycle time}$$

Alternatively, because clock rate and clock cycle time are inverses,

$$\begin{matrix}\text{CPU execution time} \\ \text{for a program}\end{matrix} = \frac{\text{CPU clock cycles for a program}}{\text{Clock rate}}$$

Instruction Performance

$$\text{CPU clock cycles} = \text{Instructions for a program} \times \begin{matrix}\text{Average clock cycles} \\ \text{per instruction}\end{matrix}$$

The term clock cycles per instruction, which is the average number of cycles each instruction takes to execute, is often abbreviated as CPI.

Performance Assessment – CPU Performance Equation

$$\text{CPU time} = \text{Instruction count} \times \text{CPI} \times \text{Clock cycle time}$$

or, since the clock rate is the inverse of clock cycle time:

$$\text{CPU time} = \frac{\text{Instruction count} \times \text{CPI}}{\text{Clock rate}}$$

Questions -

- Our favorite program runs in 10 seconds on computer A, which has a 2 GHz clock. We are trying to help a computer designer build a computer, B, which will run this program in 6 seconds. The designer has determined that a substantial increase in the clock rate is possible, but this increase will affect the rest of the CPU design, causing computer B to require 1.2 times as many clock cycles as computer A for this program. What clock rate should we tell the designer to target?

Let's first find the number of clock cycles required for the program on A:

$$\text{CPU time}_A = \frac{\text{CPU clock cycles}_A}{\text{Clock rate}_A}$$

$$10 \text{ seconds} = \frac{\text{CPU clock cycles}_A}{2 \times 10^9 \dfrac{\text{cycles}}{\text{second}}}$$

$$\text{CPU clock cycles}_A = 10 \text{ seconds} \times 2 \times 10^9 \frac{\text{cycles}}{\text{second}} = 20 \times 10^9 \text{ cycles}$$

CPU time for B can be found using this equation:

$$\text{CPU time}_B = \frac{1.2 \times \text{CPU clock cycles}_A}{\text{Clock rate}_B}$$

$$6 \text{ seconds} = \frac{1.2 \times 20 \times 10^9 \text{ cycles}}{\text{Clock rate}_B}$$

$$\text{Clock rate}_B = \frac{1.2 \times 20 \times 10^9 \text{ cycles}}{6 \text{ seconds}} = \frac{0.2 \times 20 \times 10^9 \text{ cycles}}{\text{second}} = \frac{4 \times 10^9 \text{ cycles}}{\text{second}} = 4 \text{ GHz}$$

To run the program in 6 seconds, B must have twice the clock rate of A.

Performance Assessment – Question

A compiler designer is trying to decide between two code sequences for a particular computer. The hardware designers have supplied the following facts:

Instruction Class	CPI
A	1
B	2
C	3

For a particular high-level language statement, the compiler writer is considering two code sequences that require the following instruction counts:

Code sequence	IC for instruction Class		
	A	B	C
C1	2	1	2
C2	4	1	1

Which code sequence executes the most instructions?
Which will be faster?
What is the CPI for each sequence?

Performance Assessment –
Millions of Instructions per Second (MIPS) Rate

- The rate at which instructions are executed

$$\text{MIPS} = \frac{\text{Instruction count}}{\text{Execution time} \times 10^6}$$

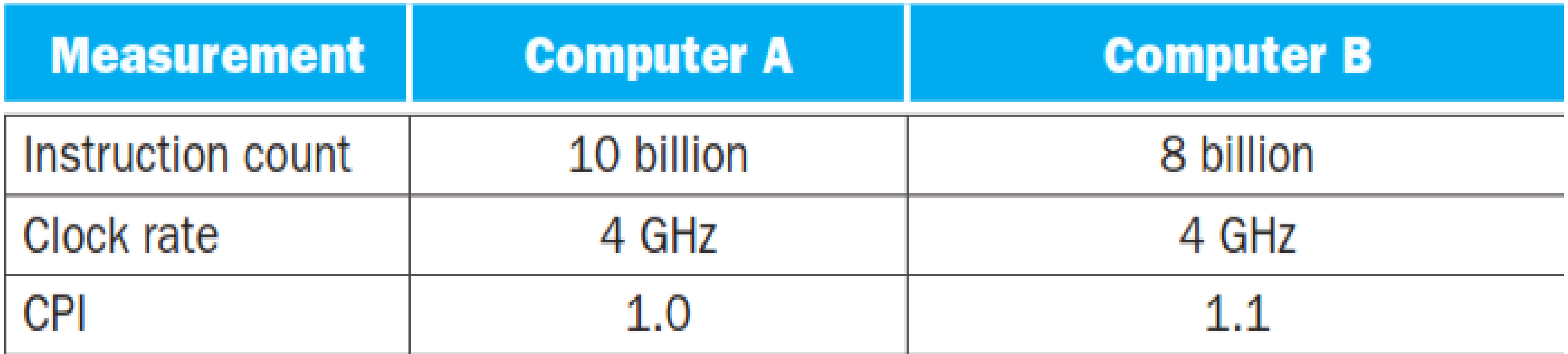
- Since MIPS is an instruction execution rate, MIPS specifies performance inversely to execution time; faster computers have a higher MIPS rating.

$$\text{MIPS} = \frac{\text{Instruction count}}{\dfrac{\text{Instruction count} \times \text{CPI}}{\text{Clock rate}} \times 10^6} = \frac{\text{Clock rate}}{\text{CPI} \times 10^6}$$

- Consider the following performance measurements for a program:

A) Which computer has the higher MIPS rating? ------Computer A

B) Which computer is faster? ------Computer B



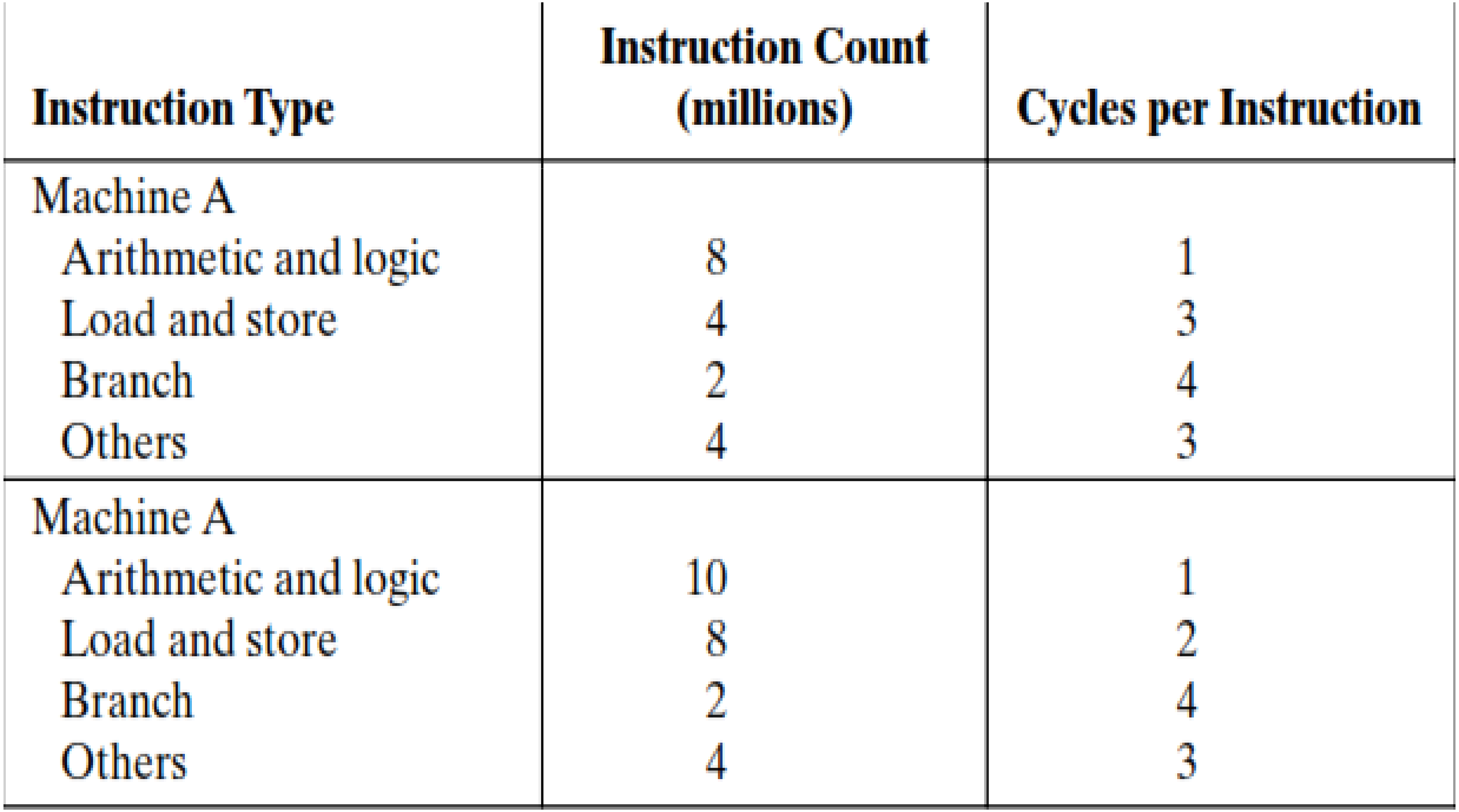
Measurement	Computer A	Computer B
Instruction count	10 billion	8 billion
Clock rate	4 GHz	4 GHz
CPI	1.0	1.1

Questions -

- Consider two different machines, with two different instruction sets, both of which have a clock rate of 200 MHz. The following measurements are recorded on the two machines running a given set of benchmark programs:

- a. Determine the effective CPI, MIPS rate, and execution time for each machine.
- b. Comment on the results.

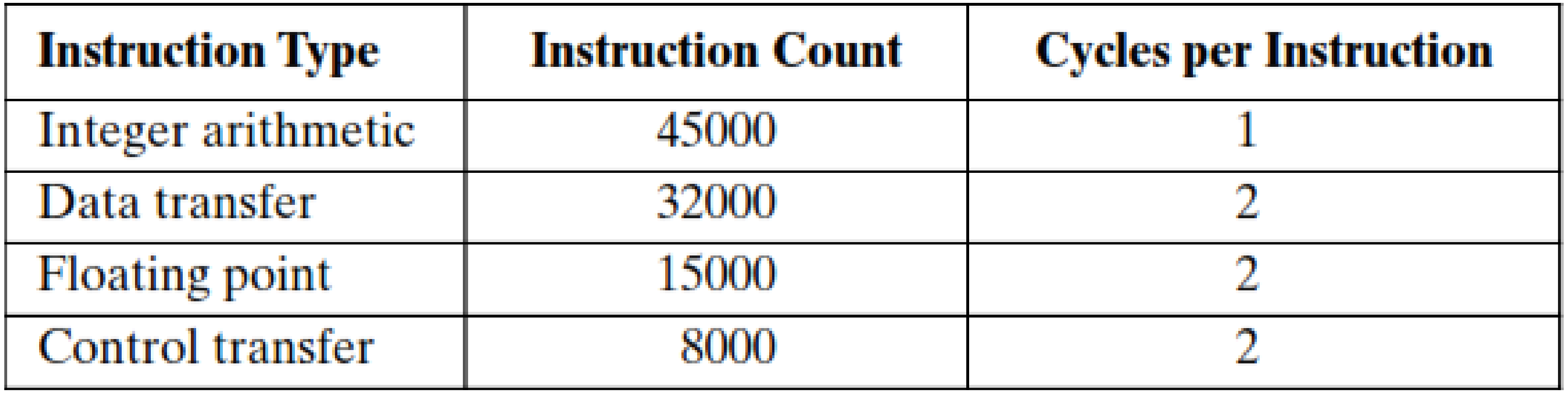
Questions -



Instruction Type	Instruction Count (millions)	Cycles per Instruction
Machine A		
Arithmetic and logic	8	1
Load and store	4	3
Branch	2	4
Others	4	3
Machine A		
Arithmetic and logic	10	1
Load and store	8	2
Branch	2	4
Others	4	3

Questions -

- A benchmark program is run on a 40 MHz processor. The executed program consists of 100,000 instruction executions, with the following instruction mix and clock cycle count.



Instruction Type	Instruction Count	Cycles per Instruction
Integer arithmetic	45000	1
Data transfer	32000	2
Floating point	15000	2
Control transfer	8000	2

- Determine the effective CPI, MIPS rate, and execution time for this program.

Performance Assessment – Amdahl’s Law

- When considering system performance, designers look for ways to improve performance by improvement in technology or change in design.

- Examples include the use of parallel processors, the use of a memory cache hierarchy, and speedup in memory access time and I/O transfer rate due to technology improvements.

Performance Assessment – Amdahl’s Law

- In all, it is important to note that a speedup in one aspect of the technology or design does not result in a corresponding improvement in performance. This limitation is succinctly expressed by Amdahl’s law.

- Gene Amdahl [AMDA67]

- Potential speed up of program using multiple processors

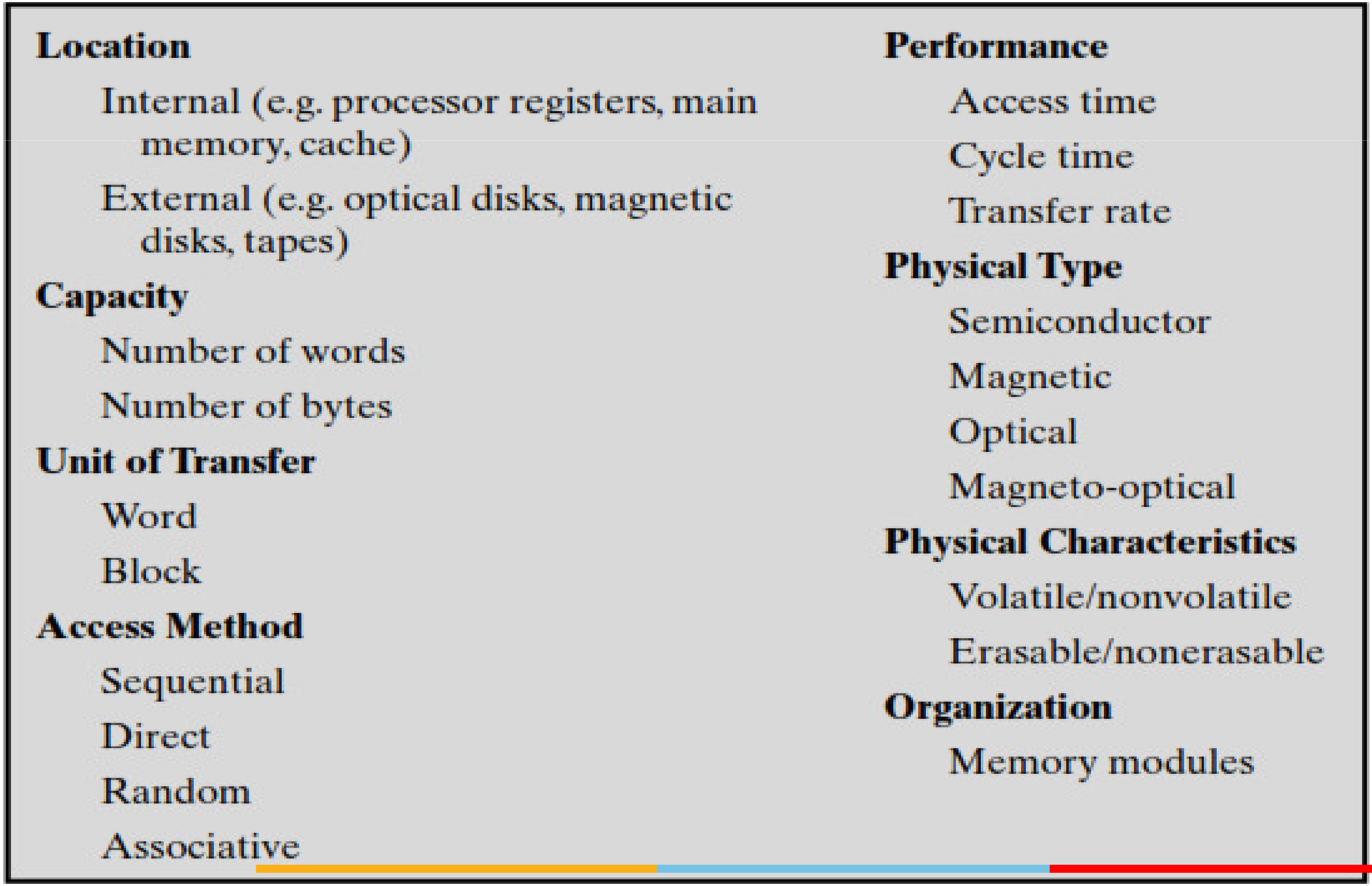
Amdahl’s Law Formula

- For program running on single processor
  - Fraction *f* of code infinitely parallelizable with no scheduling overhead
  - Fraction (1-*f*) of code inherently serial
  - T is total execution time for program on single processor
  - N is number of processors that fully exploit parrallel portions of code

$$Speedup = \frac{\text{time to execute program on a single processor}}{\text{time to execute program on } N \text{ parallel processors}} = \frac{T(1-f)+Tf}{T(1-f)+\dfrac{Tf}{N}} = \frac{1}{(1-f)+\dfrac{f}{N}}$$

- **Conclusions**
  - ***f*** **small, parallel processors has little effect**
  - ***N*** **->∞, speedup bound by 1/(1 –** ***f*)**
    - **Diminishing returns for using more processors**

MEMORY - Characteristics



Location

- Internal (e.g. processor registers, main memory, cache)
- External (e.g. optical disks, magnetic disks, tapes)

Capacity

- Number of words
- Number of bytes

Unit of Transfer

- Word
- Block

Access Method

- Sequential
- Direct
- Random
- Associative

Performance

- Access time
- Cycle time
- Transfer rate

Physical Type

- Semiconductor
- Magnetic
- Optical
- Magneto-optical

Physical Characteristics

- Volatile/nonvolatile
- Erasable/nonerasable

Organization

- Memory modules

Capacity (MxN)

- Word size (N)
  - The natural unit of organisation
  - This is typically expressed in terms of bytes (1 byte = 8 bits) or words.
  - Common word lengths are 8, 16, and 32 bits
- Number of words (M)
  - If A represents the number of bits of an address
  - Then number M of addressable units i.e. total number of memory locations are M = 2^A

Unit of Transfer

- Internal
  - Usually governed by data bus width
  - equal to the number of electrical lines into and out of the memory module
- External
  - Usually a block which is much larger than a word

Access Methods (1)

- Sequential
  - Start at the beginning and read through in order
  - Access time depends on location of data and previous location
  - e.g. Magnetic tape
- Direct
  - Individual blocks have unique address
  - Access is by jumping to the address plus sequential search
  - Access time depends on location and previous location
  - e.g. disk

Access Methods (2)

- Random
  - Each addressable location in memory has a unique, physically wired-in addressing mechanism
  - Access time is independent of location or previous access
  - e.g. RAM
- Associative
  - Data is located by a comparison with contents of a portion of the store.
  - Thus, a word is retrieved based on a portion of its contents rather than its address
  - Access time is independent of location or previous access. e.g. cache

Performance

- Access time
  - Time between presenting the address and getting the valid data
- Memory Cycle time
  - Time may be required for the memory to “recover” before next access
  - Cycle time is access + recovery
- Transfer Rate
  - Rate at which data can be moved

Physical Types

- Semiconductor
  - RAM
- Magnetic
  - Disk & Tape
- Optical
  - CD & DVD
- Others
  - Bubble
  - Hologram

Physical Characteristics

- Volatile (information decays naturally or is lost when electrical power is switched off)
- Non Volatile (information once recorded remains until deliberately changed)

- Erasable e.g. RAM
- Non Erasable e.g. ROM

Memory Hierarchy

- How much? How fast? How expensive?
- If the capacity is there, applications will likely be developed to use it.
- The question of how fast is, To achieve greatest performance, the memory must be able to keep up with the processor.
- As the processor is executing instructions, we would not want it to have to pause waiting for instructions or operands.
- The final question i.e. the cost of memory must be reasonable.

Trade Off – Capacity, Speed, Cost

- A variety of technologies are used to implement memory systems holds following relationships:
  - Faster access time, greater cost per bit
  - Greater capacity, smaller cost per bit
  - Greater capacity, slower access time

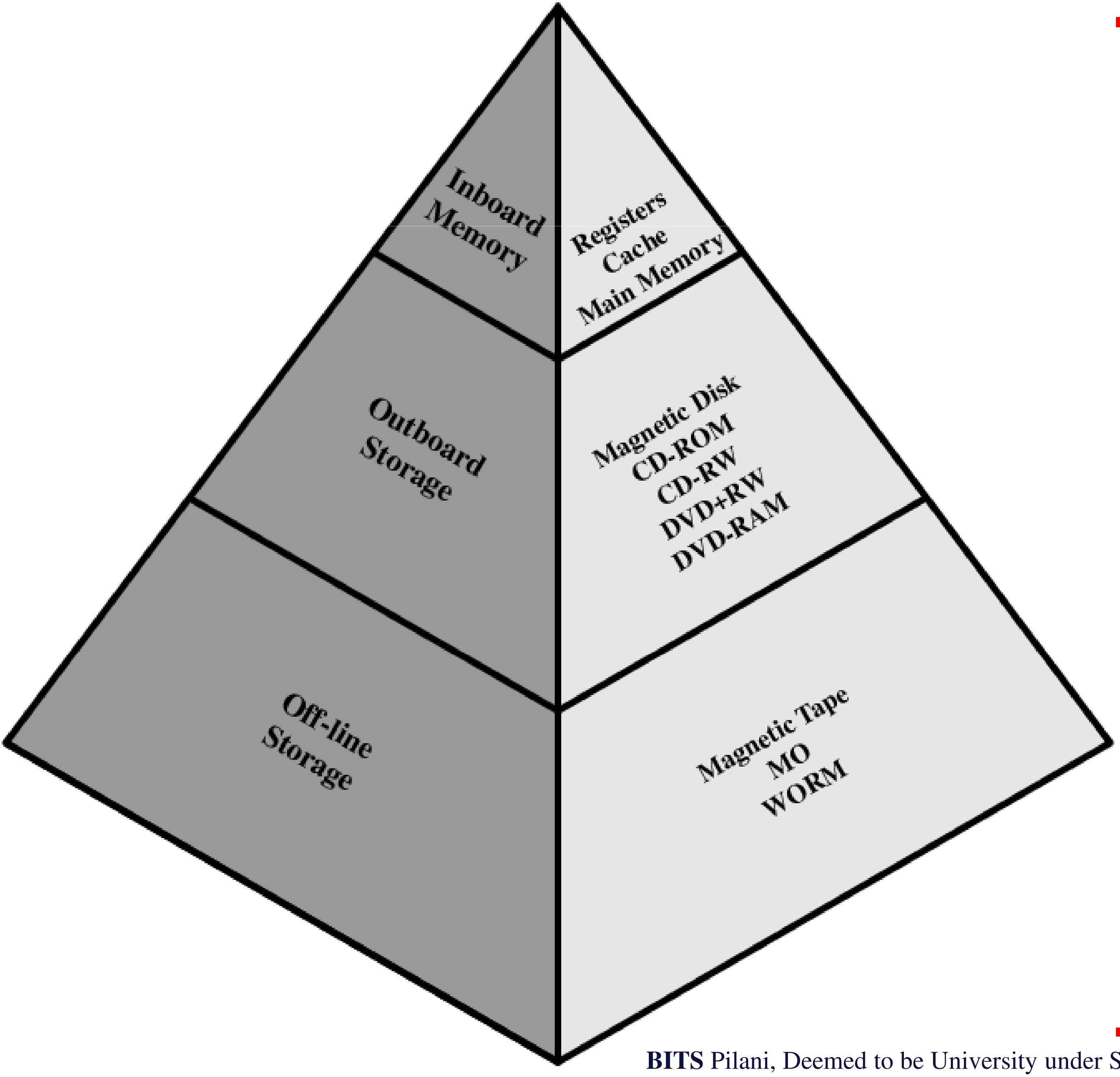
- *The designer would like to use memory technologies that provide for large-capacity memory, both because the capacity is needed and because the cost per bit is low. However, to meet performance requirements, the designer needs to use expensive, relatively lower-capacity memories with short access times.*

Memory Hierarchy

- The way out of this dilemma is not to rely on a single memory component or technology, but to employ a memory hierarchy .

- As one goes down the hierarchy, the following occur:
  - a. Decreasing cost per bit
  - b. Increasing capacity
  - c. Increasing access time
  - d. Decreasing frequency of access of the memory by the processor

Memory Hierarchy - Diagram


Inboard Memory
Registers
Cache
Main Memory
Outboard Storage
Magnetic Disk
CD-ROM
CD-RW
DVD+RW
DVD-RAM
Off-line Storage
Magnetic Tape
MO
WORM