Chapter – 9

Introduction to Parallel Processing

Parallel Processing:

Parallel processing is a method used to improve performance in a computer system. When a system processes two different instructions simultaneously then it is performing parallel processing.

a. Parallelism in Uniprocessor System:

Uniprocessor system can perform two or more task simultaneously. The tasks are not related to each other so a system that processes two different instructions simultaneously could be considered to perform parallel processing.

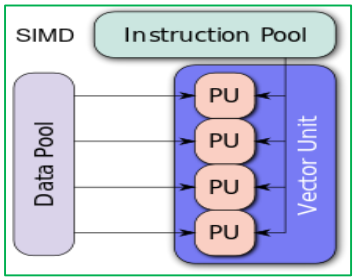
b. Parallelism in Multi-processor System:

Parallel processing system achieve parallelism by having more than one processor performing tasks simultaneously. Since, multi-processor systems are more complicated than uniprocessor system, there are many different ways to organize processors and memory. So, researcher Flynn purposed a classification based on the flow of instructions and data with in a computer called Flynn's classification.

Flynn's Classification:

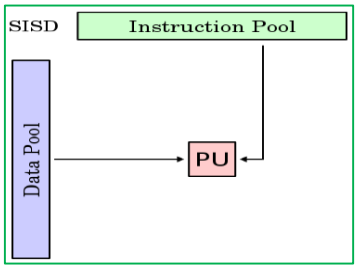
It is based on instructions and data processing:

1. SIMD (Single Instruction Multiple Data):



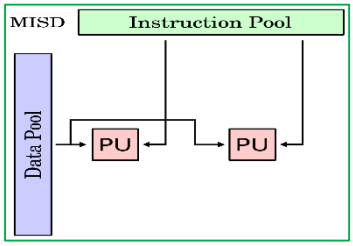
SIMD machine executes a single instruction on multiple data values simultaneously using multi-processors. Since, there is only one instruction each processor does not have to fetch and decode each instruction.

2. SISD (Single Instruction Single Data):



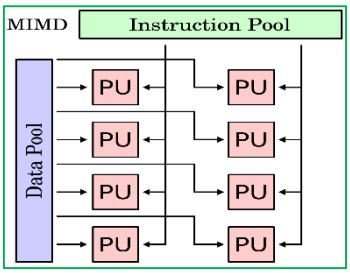
SISD machine executes a single instruction on individual data values using single processor, even if the processor incorporates internal parallelism such as an instruction pipelines, the computer would still be classified as SISD.

3. MISD (Multiple Instruction Single Data):



This classification is not practical to implement. So, number of significant MISD computer are not implemented.

4. MIMD (Multiple Instruction Multiple Data):



System referred to as multi-processor or multi-computers are usually MIMD. It may execute multiple instructions simultaneously so, each processor must include its own control unit. MIMD machines are well-suited for general purpose use.

The multiprocessors are further classified into two groups depending on the way their memory is organized. **The processors with shared memory are called tightly coupled or shared memory processors.** The information in these processors is shared through the common memory. Each of the processors can also have their local memory too. **The other class of multiprocessors is loosely coupled or distributed memory multi-processors.** In this, each processor have their own private memory, and they share information with each other through interconnection switching scheme or message passing.

The principal characteristic of a multiprocessor is its ability to share a set of main memory and some I/O devices. This sharing is possible through some physical connections between them called the interconnection structures.

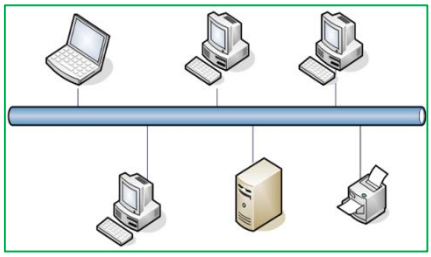
Interconnection Structures in Multiprocessors:

Although topology differs greatly, standard metrics or parameters such as diameter and bandwidth are used to quantify them.

- **Diameter:** The maximum distance between two processors in the computer system.
- **Bandwidth:** The capacity of a communication link.

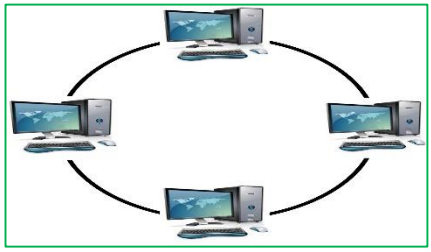
Types:

1. Shared Bus Topology:



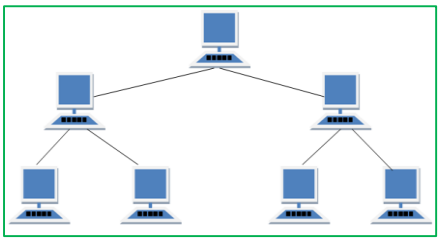
In this topology, processors communicate with each other exclusively through this bus. However, the bus can only handle one data transmission at a time. In most shared buses, processors directly communicate with their own local memory.

2. Ring Topology:



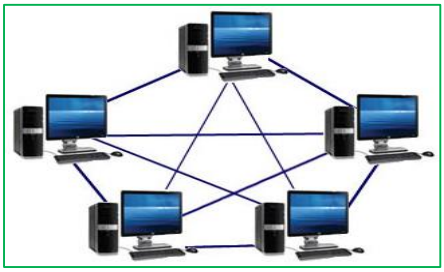
The ring topology uses direct communication between processors instead of a shared bus. This allows all communication links to be active simultaneously. Data may have to travel through several processors to reach destination.

3. Tree Topology:



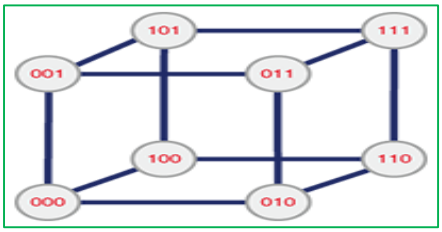
Like the ring, it uses direct communication between processors each having two or more connections. There is only one unique path between each pair of processors.

4. Mesh Topology:



In mesh topology, every processors are connected with one another, allowing for most transmissions to be distributed, even if one of the connections go down. A mesh topology can be a **full mesh topology** or a **partially connected mesh topology**.

5. Hyper-cube Topology:



The hyper-cube is a multi-dimensional mesh topology. Here, each processor connect to all other processors whose binary values differ by one bit.

CACHE COHERENCE:

When multiple processors with separate caches share a common memory, it is necessary to keep the caches in a state of coherence by ensuring that any shared operand that is changed in any cache is changed throughout the entire system to maintain data consistency. This is done in either of two ways:

1. DIRECTORY BASED PROTOCOL:

In a directory based system the data being shared is placed in a common directory that maintains the coherence between caches. The directory acts as a filter through which the processor must ask permission to load an entry from the primary memory to its cache. When an entry is changed the directory either updates or invalidates the other caches with that entry.

2. SNOOP PROTOCOL:

In a snooping system, all caches on the bus monitor (or snoop) the bus to determine, if they have a copy of the block of data that is requested on the bus. Every cache has a copy of the sharing status of every block of physical memory it has.

Cache misses and memory traffic due to shared data blocks limit the performance of parallel computing in multiprocessor computers or systems. Cache coherence aims to solve the problem associated with sharing data.

INTRODUCTION TO VECTOR PROCESSING AND ARRAY PROCESSORS:

1. VECTOR PROCESSING:

There is a class of computational problems that are beyond the capabilities of the conventional computer. These are characterized by the fact that they require vast number of computation and it take a conventional computers days or even weeks to complete. Computers with vector processing are able to handle such instruction and they have application in following fields:

- Long Range Weather Forecasting
- Petroleum Exploration
- Seismic Data Analysis
- Medical Diagnosis
- Aerodynamics And Space Simulation
- Artificial Intelligence And Expert System
- Mapping The Human Genome
- Image Processing

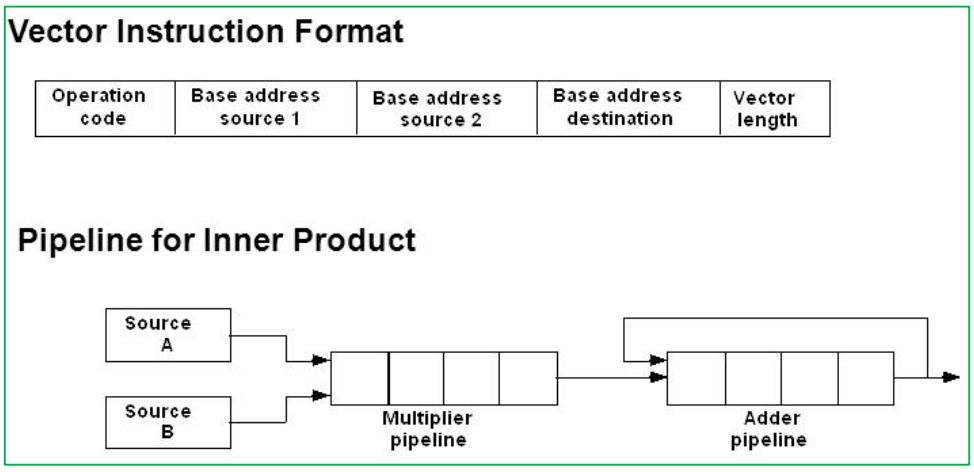
Vector Operation:

A vector 'V' of length 'n' is represented as row vector by V = [V_1, V_2, V_3, … V_n]. The element Vi of vector 'V' is written as V (I) and the index 'I' refers to a memory address or register where the number is stored.

Let us consider the program in assembly language that two vectors A and B of length 100 and put the result in vector C.

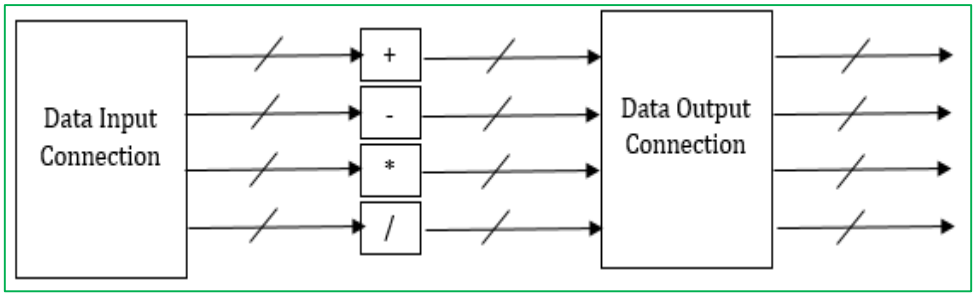
```
        Initialize I = 0
20      Read A (I)
        Read B (I)
        Store C (I) = A (I) + B (I)
        Increment I = I + 1
        If I <= 100 go to 20
        Continue
```

A computer capable of vector processing eliminates the overhead associated with the time it takes to fetch and execute the instructions in the program loop. It allows operations to be specified with a single vector instruction of the form: C (1:100) = A (1:100) + B (1:100)



Vector Arithmetic Unit:

Vector arithmetic unit is used to perform different arithmetic operations in parallel. A vector arithmetic unit contains multiple functional units. Some performs addition, other subtraction and in similar manner other performs different operations.



To add two numbers the control unit routes these value, to an adder unit.

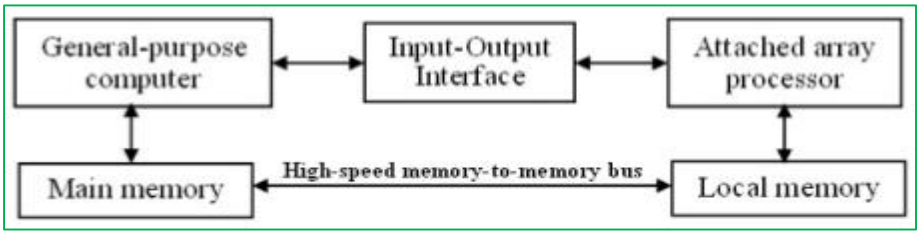
For the operations A ←B + C, D ← E – F, the CPU operations route B and C to adder and E and F to subtractor. This allows CPU to execute both instructions simultaneously.

2. Array Processor:

An array processor is a processor that performs the computations on large arrays of data. There are two different types of array processor:

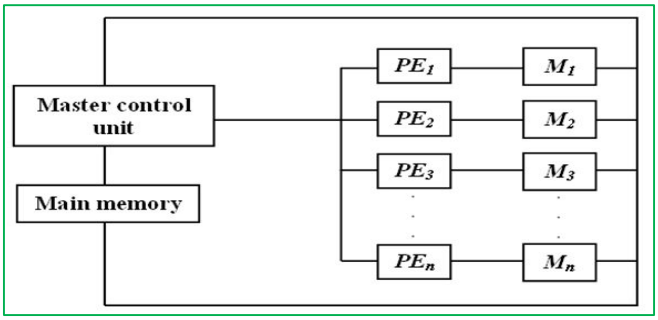
a. Attached Array Processor:

It is designed as a peripheral for a conventional host computer. Its purpose is to enhance the performance of the computer by providing vector processing. It achieves high performance by means of parallel processing with multiple functional units.



b. SIMD Array Processor:

It is processor which consists of multiple processing unit operating in parallel. The processing units are synchronized to perform the same task under control of common control unit. Each processor elements (PE) includes an ALU, a floating point arithmetic unit and working register.



INTRODUCTION TO MULTITHREADED ARCHITECTURE:

Multithreading is the ability of the central processing unit or a single core in a multi-core processor to execute multiple processes or threads concurrently and appropriately supported by operating system. This approach differs from multi-processing as threads have to share the resources of a single or multiple cores.

Multi-threading aims to increase utilization of a single core by using thread level as well as instruction level parallelism.

Merits:

- If a thread gets a lots of cache misses then other threads can continue taking the advantage of unused computing structure.
- If several threads work on the same set of data then they can actually share their cache leading to better cache uses.

Demerits:

- Multiple threads can interfere with each other while sharing hardware resources.