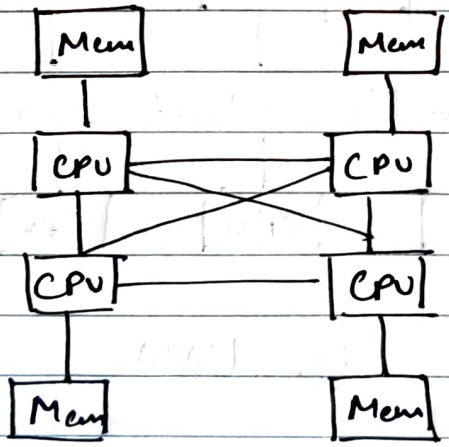
* High Performance Computing *
Shiva Saran
BE B - 20

* Unit Test 1

Question 2(a)

* Non Uniform Memory Access

- It allows memory access to every processor without any restrictions.
- A block of is attached to the processor and all blocks memory can be accessed through the path provided by the use interconnected network.



- A process has direct divided path to the block of memory attached to it.
  Eg. If accessing Memory block Mem1 from CPU1 will be much faster than accessing block Mem 2 from CPU1.
- It has significant implication ie. If we map address carefully it may be possible to keep most of the info required by a processor in the block attached to it.

- Therefore, the CPU can access that memory directly and reducing the contention for the common bus. Since the time to access a memory location depends on whether it is attached to invoking CPU or not.

This model is called NUMA.

Question 2 (b)

* Uniform Memory Access
- In this, each processor gets equal priority to access the main memory of the microprocessor is called as UMA.
- In fact, memory access of each processor is identical in UMA platform. as shown below.

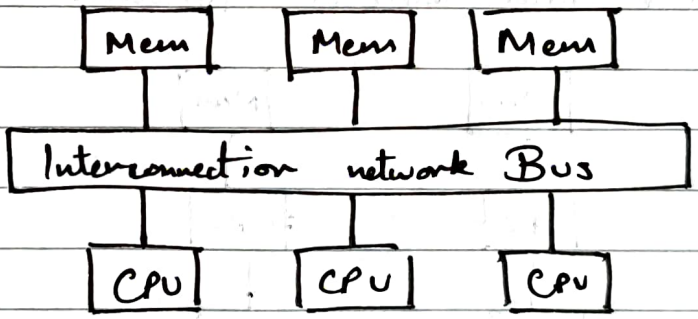


Fig. UMA

- The programming is much ~~easire~~ easier in such platform because of availability of global memory space in the system.
- The coding for read only instruction among the no. of programs running in different ~~processes~~ processors is not at all seen by the programmers.
- This happens because coding for such platforms is

similar to coding usually done in serial program for single microprocessor.
- The interaction among the read/write operation are handled requires the use of mutual exclusion or some other tools for synchorization.

Question 3 (a)

* Principle of Parallel ~~Agt~~ Algorithm Design:
- An algorithm provides step by step solution of the given problem. An algorithm accepts from the user and based upon input. after performing the defined computations, it provides the output.
- The basic steps in the design of the algorithm are:
~~Pat~~ Partitioning of overall ~~of~~ computation into smaller computation & assignments of these smaller computation into different processors.

Question 3 (b)

* Decomposition, tasks & Dependancy graphs:
° Decomposition:
- The overall computation can be partioned into number of small size computation so that these computation execute in parallel.
- The decomposition deals with the approaches of partioning the overall computation into sub-problems. When a computation is divided into many small tasks it is referred as fine-graived composition.

• Task
- In a program, the basic unit of computation is refered as a task & is controlled by OS.
- In contrast of the program, the task have unit of computation based upon that the overall computation is decomposed.

• Task-Dependancy Graph
- It is directed acyclic graph, typically a graph is a collection of nodes & edges, the task-dependancy graph also contains nodes & edges.
- The node in this graph is a task whereas edges between any two nodes represent dependancy between them.

Eg. There is edge that exists between 2 nodes T_1 & T_2. If T_2 is to be executed, it must be after T_1.

Question 5 (a)

* One to All Broadcast Communication

(i) One to all broadcast is the operation in which a single process sends identical data to all other processes.

(ii) Parallel algorithms always needs this operation.

(iii) Consider that data of size M is to be sent to all the processes.
- Initially alway source process has the data.
- After termination of algorithm there will be copy of ~~not~~ initial data with each process.
- ~~Processor~~ copies of data will be generated whereas ~~processor~~ P is the number of processor as shown in fig.

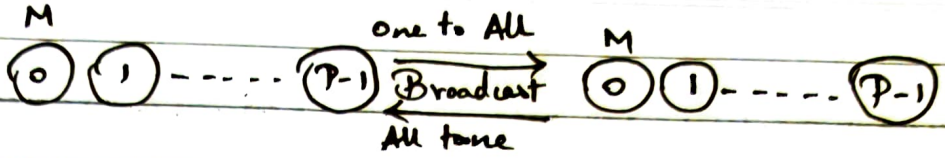


Fig. One to All Broadcast & All to one ~~broadcast~~ Reduction.

Question 5(b)

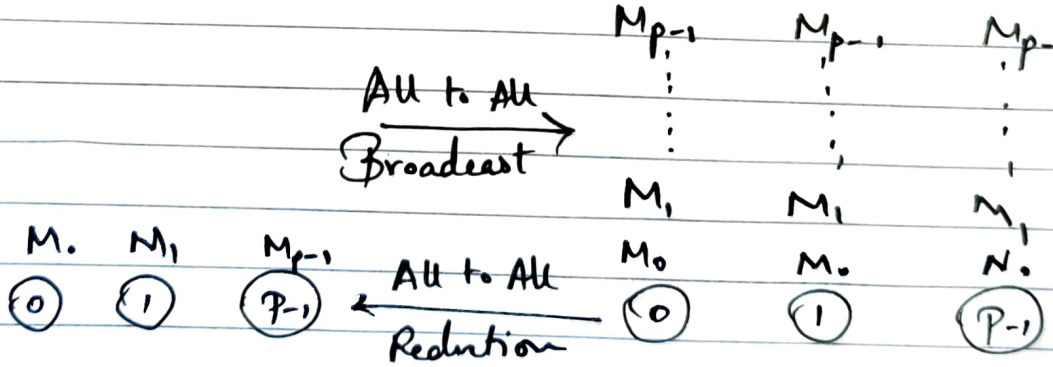
* All to All Broadcast & Reduction.

(i) All to All broadcast is a generation of one to all broadcast in which all P nodes simultaneously initiate a broadcast.

(ii) A process sends M-word message to every other process but different processes may broadcast different message.

(iii) It is used ~~to~~ in matrix operation including matrix multiplication & matrix-vector multiplication.

(iv) The dual of all to all reduction, in which every node is the destination of an all-to-one reduction.



- One approach to perform all to all broadcast is performing P one to all broadcast.
- This approach may take P times to complete communication.
- Communication link can be used more efficiently by simultaneously performing all to P one to all broadcast.
- By this there will be contention of all messages traversing the same path at the same time into a single message.