Tutorial 4
Message-passing Programming
Interconnection Networks

CS3210 – 2020/21 Semester 1

Learning Outcomes

1. Identify and solve synchronization problems in communication.
2. Apply different parallel programming patterns to solve problems.
3. Apply metrics to estimate throughput of different topologies in interconnection networks.

1. **[MPI]** Consider the following incomplete piece of an MPI program:

```
int rank, p, size=8;
int left, right;
char send_buffer1[8], recv_buffer1[8];
char send_buffer2[8], recv_buffer2[8];
...

MPI_Comm_rank(MPI_COMM_WORLD, &rank);
MPI_Comm_size(MPI_COMM_WORLD, &p);
left = (rank-1+p) %p;
right = (rank+1) %p;
...

MPI_Send(send_buffer1, size, MPI_CHAR, left, ...);
MPI_Recv(recv_buffer1, size, MPI_CHAR, right, ...);

MPI_Send(send_buffer2, size, MPI_CHAR, right, ...);
MPI_Recv(recv_buffer2, size, MPI_CHAR, left, ...);
...
```

   (a) In the program, the processors are arranged in a logical ring and each processor should exchange its name with its neighbor to the left and its neighbor to the right. Assign a unique name to each MPI process and fill out the missing pieces of the program such that each process prints its own name as well as it neighbors' names.

   (b) In the given program, MPI_Send() and MPI_Recv() operations are arranged such that depending on the implementation a deadlock can occur. Describe how a deadlock may occur.

   (c) Change the program such that no deadlock is possible by arranging the order of the MPI_Send() and MPI_Recv() operations appropriately.

   (d) Change the program such that MPI_Sendrecv() is used to avoid deadlocks.

   (e) Change the program such that MPI_ISend() and MPI_IRecv are used.

2. **[Parallel Programming Patterns and MPI]** Assume you need to compute prefix sums for multiple arrays for image convolution operation in MPI. You are not allowed to use MPI_Scan unless mentioned otherwise.

   (a) Explain how would you design a MPI program to compute the prefix sums in a pipeline programming model (you may use pseudo-code).

   (b) How many operations do you need for each stage of the pipeline? What are these operations? Enumerate and justify your assumptions about array size and processing units.

   (c) What is the maximum achievable speedup for arrays of size N?

(d) Briefly explain how would your answers to the previous questions change if you could use MPI_Scan operation?

3. **[Topology]** Bob is trying to setup a 64 nodes NUMA, but he could not decide which direct interconnection network to use. One of the main concerns for Bob is cost, especially the cost of the link between nodes. He found this network company NotInfiniBand (NIB) selling the links with the following price range:

Each Uni-direction Link (Base cost with a base data rate of 1Mbps)	SGD100
Enhanced data rate (for every additional 1Mbps)	SGD50

Table 1: Cost list for NotInfiniBand (NIB) links

Example – Two links with 5Mbps data rate will cost $(2 \times (\text{SGD } 100 + 4 \times \text{SGD } 50)) = \text{SGD } 600$

The computing nodes Bob acquired can process (send & receive) messages at 2Mbps. Help Bob out by estimating of the link cost for the following topologies:

(a) 3D Torus

(b) Hypercube

(c) Cycle-Connected-Cube

4. **[XY Routing]** Intel Xeon Phi processor uses a mesh network with "YX" message routing among its cores. In YX routing, messages move vertically in the mesh until they reach the destination row, and then move horizontally through the network until they reach the destination node.
Consider the mesh shown in Figure 1. Each link in the network is capable of transmitting one byte of data per clock.
Two messages are sent on the mesh. Both messages are sent at the same time. Message 1 is sent from node 0 to node 14. Message 2 is sent from node 11 to node 13. Both messages contain two packets of information and each packet is 4 bytes in size. Your friend looks at message workload and says: "It looks like we're going to need a more complicated routing scheme to avoid contention." Do you agree or disagree? Explain your answer.

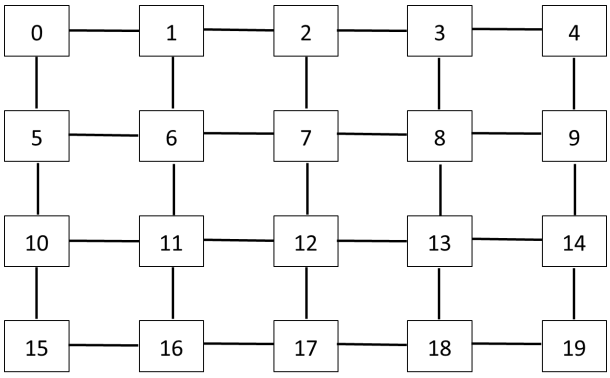

Figure 1: Mesh network

5. **[Indirect Interconnects]**

(a) Assume you are using an Omega network to connect 8 processing units with 8 memory units. How many stages do you need to use? How many switches are needed at each stage? Sketch this Omega network.

(b) In this Omega network (from point (a)) you want to send a packet from source 101 to 100. Explain how the packet is routed at each stage using XOR-Tag Routing.