EE/CSCI 451
Fall 2020
Homework 5

Assigned: September 28, 2020
Due: October 4, 2020, AOE
Total Points: 100

1 [20 points]

Explain the following terms:

1. Hypercube network
2. Store and forward routing
3. Cut through routing
4. Minimal routing
5. Flow control digits (FLITS)
6. LogP model
7. Dilation in a graph embedding
8. Congestion in a graph embedding
9. Expansion in a graph embedding
10. Network model of parallel computers

2 [20 points]

Consider the routing of messages in a parallel computer that uses store-and-forward routing. In such a network, the cost of sending a single message of size m from P_{source} to $P_{destination}$ via a path of length d is $t_s + t_w \times d \times m$. An alternate way of sending a message of size m is as follows. The user breaks the message into k parts each of size $\frac{m}{k}$, and then sends these k distinct messages one by one from P_{source} to $P_{destination}$. For this new method, derive the expression for the time to transfer a message of size m to a node d hops away under the following two cases:

1. Assume that another message can be sent from P_{source} as soon as the previous message has reached the next node in the path.
2. Assume that another message can be sent from P_{source} only after the previous message has reached $P_{destination}$.

3 [15 points]

In the LogP model, assume L=8, o=2, g=2 and P=4.

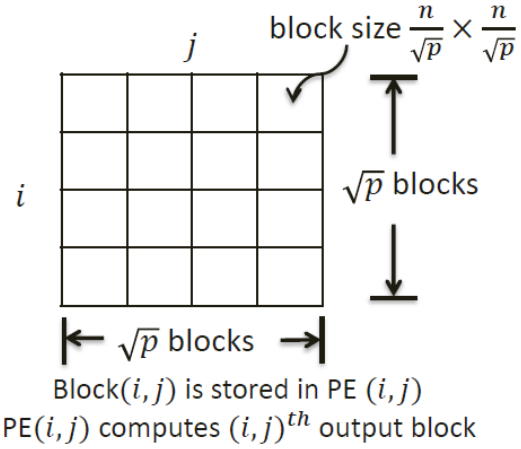
- What is the overall latency for sending a single message from one processor to another processor?
- What is the overall latency for sending n separate messages from one processor to another processor?
- Consider broadcasting a single message from P_0 to all the other processors. All processors that have received the message transmit it as quickly as possible, while ensuring that no processor receives more than one message. P_0 begins transmitting the message at time 0. Show the broadcast procedure (at what time each processor receives the message and starts to transmit the message). What is the overall latency for broadcasting the message to all the processors?

Broadcast procedure:

- At time 0, P_0 transmits to P_1.
- At time 2, the message sent by P_0 enters the network; P_0 transmits to P_2.
- At time 4, ...
- ...
- ...

4 [15 points]

In this problem we will analyze block matrix multiplication using the LogP model. Assume the matrix is of size $n \times n$ and the number of processors is p. The data layout is shown in Figure 1. Each processor locally stores a block of the input matrices and computes an output block. Each block is of size $\frac{n}{\sqrt{p}} \times \frac{n}{\sqrt{p}}$. Assume data communication and computation cannot be overlapped. To simplify analysis, assume a synchronous model with the same clock for all the processors and concurrent read is allowed.



Approach

In each processor

Repeat $\sqrt{p}$ times

Read an input block of size $(\frac{n}{\sqrt{p}} \times \frac{n}{\sqrt{p}})$

Read an input block of size $(\frac{n}{\sqrt{p}} \times \frac{n}{\sqrt{p}})$

Compute and update the output block

End

Figure 1: Block matrix multiplication using p processors

1. What is total volume of data communicated? You may use order notation.

2. In the LogP model, what is the total (parallel) time for communication? Explain.

3. Assuming each multiply-add operation takes 1 unit of time, what is the total execution (communication time and computation time) time? Explain.

5 [15 points]

Figure 2 depicts a process-processor mapping scheme. For example, process $(0, i)$, $0 \leq i < q$ are mapped to processor 0. Define the corresponding mapping function. Determine the congestion, dilation and expansion of the mapping in terms of p.

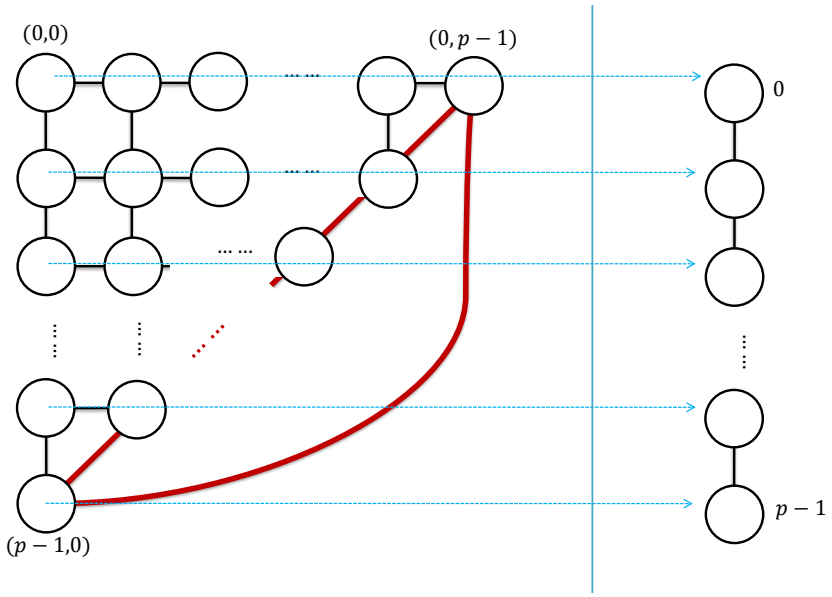


Figure 2: Process-processor mapping

Your mapping function:

f: $(i, j) \rightarrow$, $0 \leq i < p, 0 \leq j < p - i$

g: , (for the edges in red only)

6 [15 points]

1. Assume a 2-D mesh of size $n \times n$ is embedded into a 1-D mesh of size n by mapping each column to a processor (column-major mapping) as shown in Figure 3. Derive the exact expressions for dilation and congestion. Explain.

2. Assume the 2-D mesh is mapped to a 1-D mesh of same size by using row-major mapping. Derive the exact expressions for dilation and congestion. Explain.

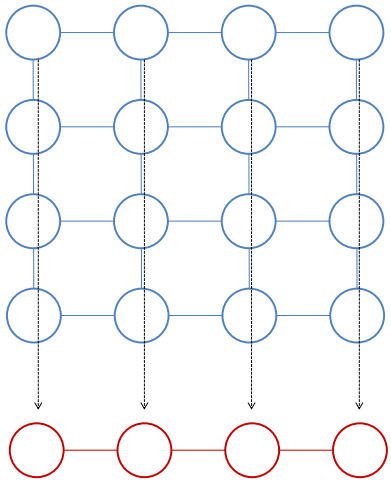


Figure 3: Embedded 2-D mesh