Communication Costs in Parallel Machines

- Along with idling and contention, communication is a major overhead in parallel programs.

- The cost of communication is dependent on a variety of features including the programming model semantics, the network topology, data handling and routing, and associated software protocols.

Message Passing Costs in Parallel Computers

The total time to transfer a message over a network comprises of the following:

- *Startup time* (t_s): Time spent at sending and receiving nodes (executing the routing algorithm, programming routers, etc.).

- *Per-hop time* (t_h): This time is a function of number of hops and includes factors such as switch latencies, network delays, etc.

- *Per-word transfer time* (t_w): This time includes all overheads that are determined by the length of the message. This includes bandwidth of links, error checking and correction, etc.

Store-and-Forward Routing

- A message traversing multiple hops is completely received at an intermediate hop before being forwarded to the next hop.

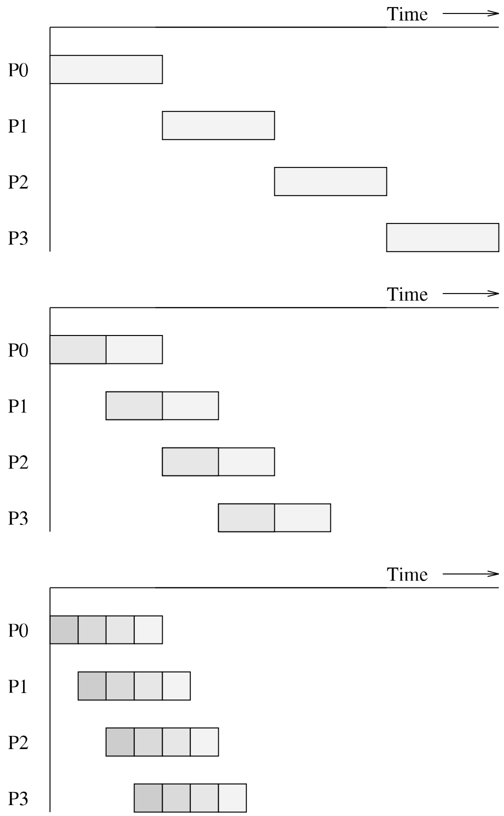
- The total communication cost for a message of size m words to traverse l communication links is

$$t_{comm} = t_s + (mt_w + t_h)l. \tag{1}$$

- In most platforms, t_h is small and the above expression can be approximated by

$$t_{comm} = t_s + mlt_w.$$

Routing Techniques



(a) A single message sent over a store-and-forward network

(b) The same message broken into two parts and sent over the network.

(c) The same message broken into four parts and sent over the network.

Passing a message from node P_0 to P_3 (a) through a store-and-forward communication network; (b) and (c) extending the concept to cut-through routing. The shaded regions represent the time that the message is in transit. The startup time associated with this message transfer is assumed to be zero.

Packet Routing

- Store-and-forward makes poor use of communication resources.
- Packet routing breaks messages into packets and pipelines them through the network.
- Since packets may take different paths, each packet must carry routing information, error checking, sequencing, and other related header information.
- The total communication time for packet routing is approximated by:

$$t_{comm} = t_s + t_h l + t_w m$$

- The factor t_w accounts for overheads in packet headers.

Cut-Through Routing

- Takes the concept of packet routing to an extreme by further dividing messages into basic units called flits.

- Since flits are typically small, the header information must be minimized.

- This is done by forcing all flits to take the same path, in sequence.

- A tracer message first programs all intermediate routers. All flits then take the same route.

- Error checks are performed on the entire message, as opposed to flits.

- No sequence numbers are needed.

Cut-Through Routing

- The total communication time for cut-through routing is approximated by:

$$t_{comm} = t_s + t_h l + t_w m.$$

- This is identical to packet routing, however, t_w is typically much smaller.

Simplified Cost Model for Communicating Messages

- The cost of communicating a message between two nodes l hops away using cut-through routing is given by

$$t_{comm} = t_s + l t_h + t_w m.$$

- In this expression, t_h is typically smaller than t_s and t_w. For this reason, the second term in the RHS does not show, particularly, when m is large.

- Furthermore, it is often not possible to control routing and placement of tasks.

- For these reasons, we can approximate the cost of message transfer by

$$t_{comm} = t_s + t_w m.$$

Simplified Cost Model for Communicating Messages

- It is important to note that the original expression for communication time is valid for only uncongested networks.

- If a link takes multiple messages, the corresponding t_w term must be scaled up by the number of messages.

- Different communication patterns congest different networks to varying extents.

- It is important to understand and account for this in the communication time accordingly.

Cost Models for Shared Address Space Machines

- While the basic messaging cost applies to these machines as well, a number of other factors make accurate cost modeling more difficult.

- Memory layout is typically determined by the system.

- Finite cache sizes can result in cache thrashing.

- Overheads associated with invalidate and update operations are difficult to quantify.

- Spatial locality is difficult to model.

- Prefetching can play a role in reducing the overhead associated with data access.

- False sharing and contention are difficult to model.

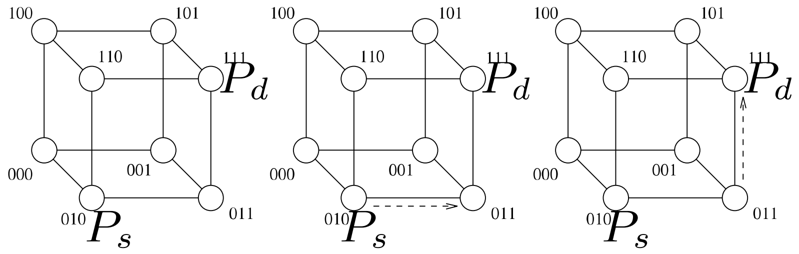
Routing Mechanisms for Interconnection Networks

How does one compute the route that a message takes from source to destination?

- Routing must prevent deadlocks – for this reason, we use dimension-ordered or e-cube routing.

- Routing must avoid hot-spots – for this reason, two-step routing is often used. In this case, a message from source s to destination d is first sent to a randomly chosen intermediate processor i and then forwarded to destination d.

Routing Mechanisms for Interconnection Networks



Step 1 (010 →110)

Step 2 (110 →111)

Routing a message from node P_s (010) to node P_d (111) in a three-dimensional hypercube using E-cube routing.