SFWRENG 4F03
Parallel Computing

Winter 2020

05 Communication and Latency

Dr Asghar Bokhari

Faculty of Engineering, McMaster University

February 4, 2020

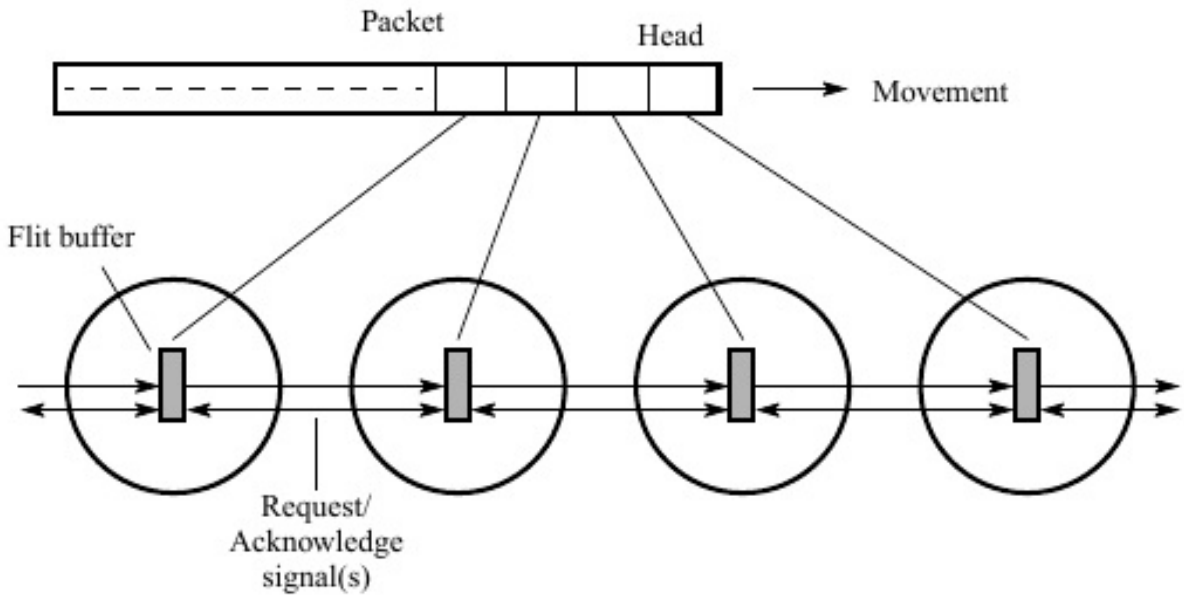
Communication Methods

- Circuit Switching
  - establish path and maintain all links until entire message has passed through (like old telephone systems).
- Packet switching
  - Message broken up into "packets" of data.
  - Each packet has a source and destination address.
  - High latency, large buffer space
  - Sometime called store and forward packet switching

Communication Methods

- Virtual Cut-through
  - If outgoing link available, message is immediately transmitted and not stored in buffer.
- Wormhole Routing
  - Alternative to normal store-and-forward routing
  - Goal to reduce latency and size of buffers.
  - Message divided into smaller units called *flits*. Size is usually 1-2 bytes
  - Only head of message (first flit) is transmitted from source node to next node when link available.
  - When head flit moves to next node, the next flit can move ahead.
  - Requires a request/acknowledge system

Wormhole Routing



1

[1] B. Wilkinson, and M. Allen, *Parallel Programming. Techniques and Applications Using Networked Workstations and Parallel Computers*, Prentice-Hall, 1999.

Communication Methods

- Wormhole Routing Continued
  - Necessary to reserve entire path for message as the flits are linked.
  - Other packets cannot be interleaved with *flits* along the same links.
  - Approach requires less storage and has latency independent of path length.
  - Flits (flow control digits)
  - Request/acknowledge signaling

Communication Cost

- Message Passing Time
  - Startup time, t_s
  - Per-hop time, t_h
  - Per-word time, t_w

  A message consists of a header and the actual data

Startup Time

- Startup time, t_s includes time
  - to prepare a message (adding header, trailer, error correction information)
  - to execute a routing algorithm
  - to establish interface between the local node and the router
- Startup time occurs **once** per message

Per-hop Time

- This is the time for a header to travel between two directly connected nodes (one link)
- Also called *node latency*

Per-word Transfer Time

- This is the time for a word to traverse a link
- Assume a channel bandwidth is r words/sec
  Then
  $t_w = \frac{1}{r}$

Implications

- $t_{comm} = t_s + lt_h + t_w m$
- t_s, t_w, and t_h are determined by hardware, software layers,and messaging semantics
- Programmers do not have control over them
- t_s is much larger thant_h and t_w
- Send larger messages versus many small ones
- Minimize the volume of data: better algorithms
- Minimize l: not much control by the user
- With MPI: little control over mapping processes to processors

Simplified Model

- t_h is usually dominated by t_s for small messages or t_wm for large messages
- Maximum number of hops l is usually small in parallel machines
- We can simplify
  $t_{comm} = t_s + lt_h + t_w m$
  to
  $t_{comm} = t_s + t_w m$

Network Latency Comparison

- Comparison of latencies of different networks
- Let
  - B stand for bandwidth (bits/s)
  - l be the number of links the message passes through, and
  - L the message size in bits.
- For sending L bits over a single link, the transmission time is: L/B

Network Latency Comparison

- Circuit Switching Networks
- Let
  - L_c be the length (in bits) of control packet sent to establish path.
  - $$Latency = (\frac{L_c}{B})I + (\frac{L}{B})$$
  - Latency reduces to L/B if $L_c \ll L$

Network Latency Comparison

- Store and Forward
  - $Latency = (\frac{L}{B})I$
- Virtual Cut Through
- Let
  - L_h be the length (in bits) of header field of first packet of message.

$$Latency = (\frac{L_h}{B})I + (\frac{L}{B})$$

  - Latency reduces to L/B if $L_h \ll L$

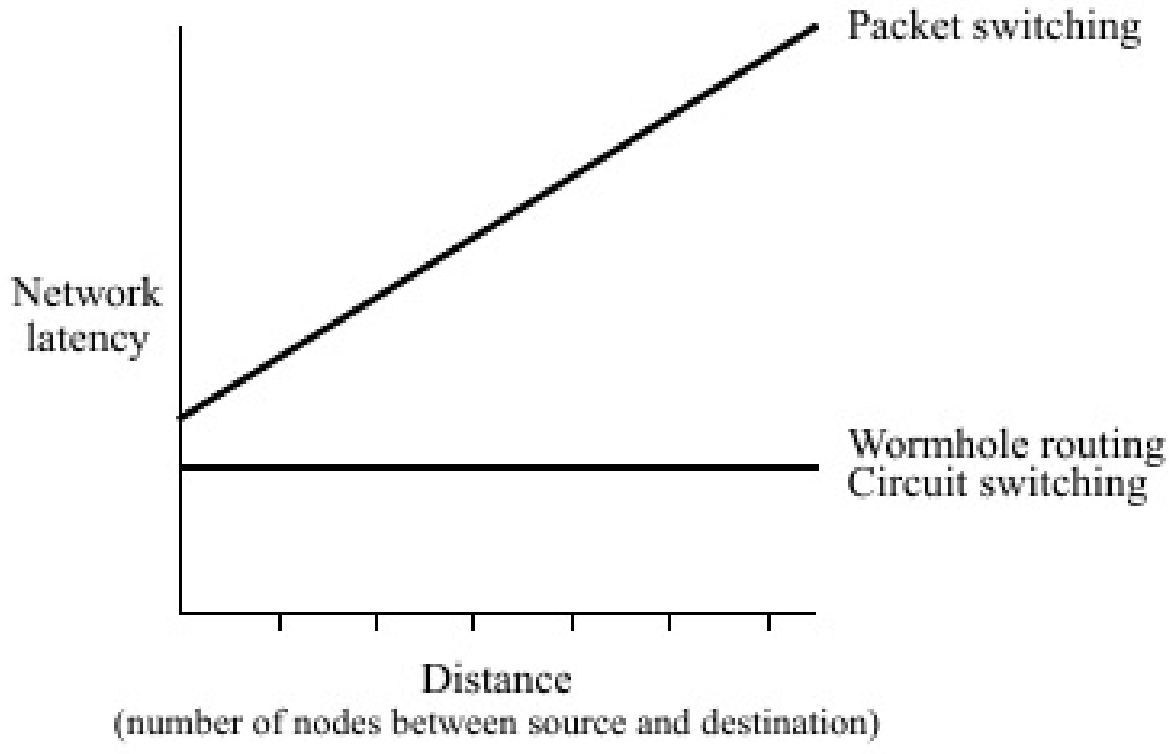
Network Latency Comparison

- Wormhole Routing
- Let
  - L_f be the length (in bits) of each flit.

$$Latency = (\frac{L_f}{B})l + (\frac{L}{B})$$

  - Latency reduces to L/B if $L_f \ll L$

Latency Comparison

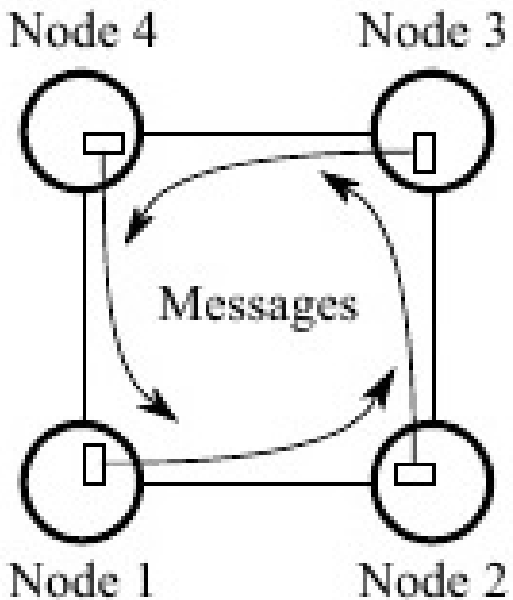


2

[2]B. Wilkinson, and M. Allen, *Parallel Programming. Techniques and Applications Using Networked Workstations and Parallel Computers*, Prentice-Hall, 1999.

Dead Lock and Live Lock

- Interconnection networks use routing algorithms to determine path.
- Can be adaptive; They can choose alternative paths depending on criteria - ie. local traffic
- Algorithm may *deadlock* or *livelock.*



Cluster Computing

- Use existing workstations and PCs connected together to form powerful computing platforms
- High performance computers available at low cost
- Latest processors can be incorporated into the system that can be incrementally expanded
- Such groups of computers were called **COWs**(Cluster of workstations) or **NOWs**(Networks of workstations)
- Usually called clusters these days
- Tools were developed to use these sets of computers collectively
- PVM (parallel virtual machine)
- MPI message passing interface
- More details later

Cluster Configurations

- Early ways to form a cluster:
  - Using existing network of computers
  - Grouping existing networked computer into a dedicated cluster for HPC
- **Beowulf Clusters:**
  - Standard off-the-shelf microprocessors
  - Open source OS (Linux)
  - Low cost connection (Ethernet)
  - Best cost/performance ratio
  - Another name - commodity computers

Cluster Configurations

- **Beyond Beoulf:**

  Interconnects: Gigabit Ethernet, specialized high performance networks- Myrinet(2.4 Gb/s)

  Clusters with Multiple Interconnects: Multiple Ethernet cards, channel bonding

  SMP clusters: message passing between SMPs, threads and other shared memory techniques within SMPs.

  Web Clusters: - grid computing

Title

-
-
- Template!