High Performance Computing: Concepts, Methods & Means

Enabling Technologies 2: Cluster Networks

Prof. Amy Apon
Department of Computer Science and Computer Engineering
University of Arkansas
March 15th, 2007
Outline

- Basics
- Networks in the Top 500
- Details on some cluster networks
  - Gigabit Ethernet
  - Myrinet
  - Infiniband
- Software stack
  - Low latency messaging
  - MPI implementation
A few basics

- $100\text{Mbps} = 100 \times 10^6 \text{bits/second}$
- $1000\text{Mbps} = 1\text{Gbps} = 10^9 \text{bits/second}$
- Latency (one-way): the time from the source sending a packet to the destination receiving it
- Throughput: data per unit time that is delivered
- Networks are built in layers
Network Layers

Application (e.g., FTP, HTTP, MPI)
Transport (e.g., TCP, UDP)
Network (e.g., IP)
Data Link (e.g., Ethernet)
Physical (e.g., cables, etc.)
When a message is sent

1. The application constructs a message

2. The message is packaged (encapsulated) with a header from the transport layer (e.g., TCP) and sent to the network layer. TCP provides end-to-end reliable transfer of messages.

3. The network layer adds a header. IP provides a packet format and address across multiple networks.

4. The data link layer adds a header, and the frame is sent out on the network. The data link layer provides point-to-point transfer of frames.
Interconnect Family, Top 500 Supercomputers since 2003

- Gigabit Ethernet
- Myrinet
- InfiniBand
- SP Switch
- Proprietary
- NUMAlink
- Quadrics
- Crossbar
- Cray Interconnect
- Mixed
- RapidArray
- NUMAflex
- Fast Ethernet
- Fireplane
- HIPPI
- Giganet
- Giganet
- SCI
- PARAMNet
- N/A
Cray XT3 Interconnect

2. Each processing node in the Cray XT3 consists of an Opteron processor, memory, and a SeaStar 3D interconnect chip. Each SeaStar is connected to its six nearest neighbors.

http://www.ed-china.com
Interconnect Family, Top 500 Supercomputers since 2003

(simplified view)

Top 500 List

Count

June, 2003
Nov, 2003
June, 2004
Nov, 2004
June, 2005
Nov, 2005
June, 2006
Nov, 2006

Gigabit Ethernet
Myrinet
InfiniBand
SP Switch
Proprietary
Others
Gigabit Ethernet


http://www.timbercon.com/Gigabit-Ethernet/Gigabit-Ethernet.jpg

http://image.compusa.com/prodimages/2/6ba9cda0-2414-47c1-ae29-3b9c14c7b3f8.gif
Five important features of Gigabit Ethernet

1. User data rate of 1000Mbps (1Gbps)
2. Free on the node; small switches are inexpensive, too. 😊
3. Is usually a switched network, however, the link from the switch to the node is a collision domain using CSMA-CD
3. Carrier Sense Multiple Access with Collision Detection (CSMA-CD)

Used in Gigabit Ethernet, Fast Ethernet, Ethernet, and wireless Ethernet

- To send: listen to wire; if not busy then send,
- If busy then stay on wire listening until the wire becomes not busy, then send immediately
- After send, listen long enough to be sure there was no collision,
- If not, assume it was successful
- If collision, wait a little while and then try to send same frame again, up to 10 tries, then report a failure
  - Exponential backoff – the wait time increases as a power of 2
CSMA-CD, continued

- Handling the carrier sense and collisions takes time
- Time and latency for messaging is bad in a cluster

Good news: On a switch, the CSMA-CD protocol is only used on the link from the node to the switch

Moral: always use switches in clusters - all Gigabit Ethernet products are switched
Using a Switched Network

Important to notice:
- One sender gets 100% of the network
- A collision can only happen when two or more senders send on a common link

![Diagram showing OK and Collision scenarios in a switched network]
4. Jumbo Frames

- Some Gigabit Ethernet products also have a "bulk mode" that bundles frames together before sending.
- This helps to reduce the number of collisions, hurts with latency on clusters.
- If this is an option, disable it for cluster computers if Gigabit Ethernet is used for application messaging.
5. Source Routing

- Ethernets also use a source routing algorithm to find a path when there is a loop in the topology
  - The source calculates all paths and stores them prior to any messaging

- Can't take advantage of multiple paths in switches to increase bandwidth, because only one path is used from one node to another
GigE Advantages and Disadvantages

● **Advantages:**
  - Inexpensive
    ● NICs are free!
    ● A 5-port switch is $35.99 after a $10 rebate, 24-port switch for $239.99 (www.newegg.com)
  - easily available, easy to install and use, uses well-understood protocols and drivers

● **Disadvantage:** performance is not as good as it could be
10Gbps Ethernet

- Optical fiber media only (no copper)
- Full duplex only – no CSMA-CD
- Up to 40km distance
- Direct attachment to SONET (a wide area network technology)
- New encoding scheme
- More expensive!
  - About $3K now just for the network interface card (NIC)

Considering these for connecting to the head node of an HPC cluster
Myrinet

http://www.myri.com
Myrinet

- Developed and used in the early 1990's by the **academic community** specifically for supporting message-passing supercomputers (Cal-Tech, Berkeley NOW project, …)
- Widely used in clusters, especially those on the Top 500
- Fast data rate, low error rate
- **Goal:** replicate the distinctive characteristics of the interconnect of high-performance message passing computers of the 1980’s:
Five important features of Myrinet

1. Regular topologies and multiple data paths
   - Makes the cluster easier to build if the topology is regular
   - Scales to 1000's
   - Can take advantage of multiple data paths using a **fat tree** to increase the bandwidth between nodes
Example Fat Tree (Clos Network)

http://www.top500.org/orsc/2006/myrinet.html
Another Example Fat Tree

2. Cut-through routing

3. Flow control on every communication link

- Point-to-point full duplex links connect hosts and switches
- Flow control through the use of STOP and GO bits on every link
  - A buffer exists on the receiving side to hold bits.
  - If the buffer gets too full, a STOP message is sent.
  - If the buffer gets too empty, a GO message is sent.
Flow control, continued

Operation of the Myrinet slack buffer
Flow control, continued

The size of the buffer depends on the transmission rate, the rate of the signal in the wire, and the length of the wire, and can be calculated. E.g., suppose the transmission rate is 1Gbps (it's a lot higher on the newest NICs). Suppose the cables is at most 25m long, and suppose the rate that the signal travels in the wire is 200m/usec.
Then, the number of bits "on the wire" is
Flow control, continued

\[ 1 \times 10^9 \text{ bits/sec} \times 25\text{m} \]

\[ \#\text{bits} = \frac{10^3 \times .125}{200\text{m/ } 10^{-6}\text{sec}} = 125 \]

This is the number of bits that can fill a buffer on the receiving side while a STOP message is being sent.

Times 2 for the bits that enter the wire while the STOP message is on the wire, = 250 bits = 250/8 bytes = 31.25 bytes (about 32 bytes)

This matches the figure in the Boden paper, plus some extra bytes for hysteresis.
More Myrinet Features

4. Variable length header that holds the route, stripped at each switch and routed

5. DMA engine and processor for implementing a low-latency protocol (and we will see one more disadvantage of Gigabit Ethernet – hold this thought!)
### Current Myrinet Products

<table>
<thead>
<tr>
<th>Product series</th>
<th>Myrinet-2000</th>
<th>Myri-10G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full-duplex data rate for the links, NIC ports, and switch ports</td>
<td>2+2 Gigabits/s</td>
<td>10+10 Gigabits/s</td>
</tr>
<tr>
<td>Link cables</td>
<td>LC-connectorized duplex multimode fiber to 200m</td>
<td>Selected 10-Gigabit Ethernet cables, copper and fiber</td>
</tr>
<tr>
<td>NICs</td>
<td>Single-port and dual-port PCI-X</td>
<td>Single-port PCI-Express, dual-protocol 10G Myrinet or 10G Ethernet</td>
</tr>
<tr>
<td>Switches</td>
<td>Based on 16-port and 32-port crossbar switches</td>
<td>Based on 16-port crossbar switches</td>
</tr>
<tr>
<td>Switch networks</td>
<td>Up to 256 host ports with a single &quot;Network in a Box&quot; component, and up to tens of thousands of hosts by combining these components</td>
<td>Up to 128 host ports with a single &quot;Network in a Box&quot; component, and up to tens of thousands of hosts by combining these components</td>
</tr>
<tr>
<td>Interoperability</td>
<td>Gigabit Ethernet</td>
<td>10-Gigabit Ethernet</td>
</tr>
<tr>
<td>Myrinet software support</td>
<td>Myrinet Express (MX-2G) or GM-2</td>
<td>Myrinet Express (MX-10G)</td>
</tr>
<tr>
<td>MX or MPI latency</td>
<td>2.6µs–3.2µs</td>
<td>2µs</td>
</tr>
<tr>
<td>MX unidirectional data rate</td>
<td>247 MBytes/s (one-port NICs) 495 MBytes/s (two-port NICs)</td>
<td>1.2 GBytes/s</td>
</tr>
<tr>
<td>TCP/IP (MX ethernet emulation) data rate</td>
<td>1.98 Gbits/s (one-port NICs) 3.95 Gbits/s (two-port NICs)</td>
<td>9.6 Gbits/s</td>
</tr>
</tbody>
</table>
Summary of Myrinet

Advantages of Myrinet
- very high performance
- high data rate
- low probability of blocking
- fault tolerant
- supports low-latency messaging
- scales to very large clusters
- 10G Myrinet connects to 10GigE

Disadvantages of Myrinet
- higher cost than GigE
  - $500 Myrinet 2000 NIC, $1000 Myri-10G NIC
  - $4K Myrinet 2000 8-port switch
  - $4K 16-port line card for 10G Myrinet
- single supplier

http://www.pcpcdirect.com/highperf.html
InfiniBand

- First, let’s look at the storage and file system structure on a typical supercomputer (e.g., Red Diamond)
InfiniBand Architecture (IBA)

- Developed by *industry* and is the result of the merging of two competing standards for I/O and storage
- Not yet widely adopted by storage industry, but is gaining presence in the Top 500
- Fast data rate, low error rate
- **Goal:** replace just about every datacenter I/O standard including PCI, Fibre Channel, SCSI, and various networks like Ethernet.
Five important features of InfiniBand

1. InfiniBand is both a hardware and software standard, see http://www.infinibandta.org/home

2. Components include:
   - Target Channel Adapter (to devices like RAID disk subsystems)
   - Host Channel Adapters (to nodes, can sit on PCI bus, or on a modified motherboard)
   - Point-to-point switches that create a switched fabric – redundant paths are supported
Example System Area Network based on InfiniBand

More InfiniBand Features

3. Configurable Virtual Lane routing that can support multiple logical channels on the same physical link.
   - A mechanism to avoid Head of Line blocking and to support QoS (as in ATM)

4. Credit-based flow control on every link
   - Receiver tells sender how many bytes to send
   - Sender sends that many, and then waits until it gets more “credits”
   - Credits can be sent in acks or separately

5. Remote Direct Memory Access (RDMA) and low-latency protocol support (hold this thought!)
InfiniBand Data Rates

Effective theoretical throughput in different configurations

<table>
<thead>
<tr>
<th></th>
<th>single</th>
<th>double</th>
<th>quad</th>
</tr>
</thead>
<tbody>
<tr>
<td>1X</td>
<td>2 Gbit/s</td>
<td>4 Gbit/s</td>
<td>8 Gbit/s</td>
</tr>
<tr>
<td>4X</td>
<td>8 Gbit/s</td>
<td>16 Gbit/s</td>
<td>32 Gbit/s</td>
</tr>
<tr>
<td>12X</td>
<td>24 Gbit/s</td>
<td>48 Gbit/s</td>
<td>96 Gbit/s</td>
</tr>
</tbody>
</table>
## InfiniBand Products

### Cisco SFS 7000 Series InfiniBand Server Switches

<table>
<thead>
<tr>
<th>Technology</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>InfiniBand (4X SDR = 10Gbps, 12X SDR = 30Gbps; 4X DDR = 20Gbps)</td>
<td>24 autosensing InfiniBand 4X dual-speed DDR/SDR interfaces</td>
</tr>
<tr>
<td><strong>Products</strong></td>
<td></td>
</tr>
<tr>
<td>Cisco SFS 7000D</td>
<td>24 autosensing InfiniBand 4X dual-speed DDR/SDR interfaces</td>
</tr>
<tr>
<td>Cisco SFS 7008P</td>
<td>96 4x SDR InfiniBand ports</td>
</tr>
<tr>
<td>Cisco SFS 7012D</td>
<td>Up to 144 autosensing InfiniBand 4X dual-speed DDR/SDR interfaces</td>
</tr>
<tr>
<td>Cisco SFS 7024D</td>
<td>Up to 288 autosensing InfiniBand 4X dual-speed DDR/SDR interfaces</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td></td>
</tr>
<tr>
<td>All switches are fully non-blocking with fully redundant components</td>
<td></td>
</tr>
<tr>
<td>All switches support optically powered ports</td>
<td></td>
</tr>
<tr>
<td>Cisco SFS 7000D and SFS 7008P support an embedded subnet manager</td>
<td></td>
</tr>
</tbody>
</table>
IB cabling at Sandia on Thunderbird (#6 in Nov. 2006)
Summary of InfiniBand

- Advantages of InfiniBand
  - very high performance
  - low probability of blocking
  - fault tolerant, supports low-latency messaging
  - scales to very large clusters
  - multiple vendors
  - can potentially connect directly to storage devices

- Disadvantages
  - higher cost than GigE (and Myrinet?)
    - HCA $895, 24 port SDR switch $11,495, 5m cable $200
Recall the Network Layers

- **Application (e.g., MPI)**
- **Transport (e.g., TCP)**
- **Network (e.g., IP)**

**Data Link**
- Flow control on links: CSMA-CD, stop and go, credits
- Routing through switches: source routing, stripped header, VL
- Point-to-point transfer of frames

**Physical**
- Encoding of bits, signaling, boxes and wires
An MPI message over Gigabit Ethernet

Application builds message in user memory

Message is copied to kernel, checksum and TCP/IP and Ethernet headers are added

Message is sent onto network to receiver

You only see the communication at THIS LEVEL

Message arrives to user memory and the application is notified

Message is copied to kernel and TCP/IP and Ethernet headers are stripped

Message arrives to user memory and the application is notified

All of this copying and kernel intervention takes time and increases latency!!
Six low-latency messaging concepts

1. Use buffers in pinned memory on the send and receive side
   - This is direct memory access by the DMA engine on the NIC
   - Pinned memory is not allowed to be paged out to disk
   - Also sometime called “kernel bypass”

2. Use scatter/gather lists in the specification of the data to be sent, so that data is copied from a list of locations to the NIC directly
Scatter/gather in Quadrics

3. Hardware support on the NIC for non-blocking sends and receives

Non-blocking send requires 2 steps:
1. 'post' a send, notify the NIC that data is available, the NIC gets the data,
2. some time later the program checks to see if the send is complete.

The user is responsible for not modifying the buffer before the send completes

Non-blocking receive also requires 2 steps:
1. post a receive to tell NIC where to put the data when the message arrives,
2. check later to see if the receive is complete

This all can be emulated in software but the work is similar to that of ordinary network protocols.
More low-latency messaging concepts

4. Polling by the CPU for receives is allowed, but is (almost) cheating

- Makes the receive appear to happen faster when a program is doing nothing but waiting for a message
- But wastes CPU cycles when other activity has to happen at the same time
More low-latency messaging concepts

5. If the receive is not posted when the message arrives the message will be dropped

Solution: pre-post a number of buffers for messages that arrive "unexpectedly"

- This works fine for small messages.
  - An "eager" protocol puts all small messages into prepared buffers
- Works badly for large messages
  - Use a "rendezvous" protocol instead – send a message that says a message is coming, receive an ack, then send the message

The eager/rendezvous threshold is configurable in Myrinet (and InfiniBand?) See http://www.myri.com/serve/cache/166.html
Example of the effect of the eager/rendezvous threshold

6. **Low-latency messaging is too low a level for user-level programming**

- It requires the ability to pin user memory. This can be dangerous in a multi-user system, and is an expensive operation in a single-user system.
- Thus, network companies have built their own low-latency messaging systems
  - Myrinet: GM, MX
  - InfiniBand: Sockets Direct Protocol (SDP), Virtual Interface Architecture (VIA)

Special MPI implementations are built over these (e.g., MVAPICH over InfiniBand)
Network Layers Without TCP/IP

Application (e.g., MPICH)

Low latency messaging: GM or MX, Sockets Direct Protocol, ...

Flow control on links: stop and go, credits

routing through switches: stripped header, VLanes

Point-to-point transfer of frames

encoding of bits, signaling, boxes and wires
MPICH2 Implementation Structure

Liu, et. al, “Design and Implementation of MPICH2 over InfiniBand with RDMA Support,” PDPS 2004
Implementation of MPI Collective Operations

Implementation issues include:

- Segmentation and message size
- Non-blocking versus blocking messages
- Virtual topologies

Message Segmentation
Barrier Topologies

Flat Tree

Double Ring

Recursive Doubling
Barrier Performance

![Graph showing Barrier Performance - TCP/IP]

- **Barrier Native (Double ring)**
- **Barrier Recursive Doubling**
- **Barrier Flat tree**

The graph illustrates the minimum test duration in seconds as a function of the communicator size for different barrier operations. The x-axis represents the communicator size, while the y-axis shows the minimum test duration.
Broadcast Topologies

Binomial Tree

Flat Tree

Binary Tree*
Broadcast Performance

Broadcast - 32 nodes

Minimum Test Duration [sec]

Message Size [byte]
Reduce Analysis

Successful algorithms:

- **Small messages** - Flat tree
- **Large messages** - K-Chain (for small K values – 1,2!)
- **Intermediate messages** – K-Chain but it depends on number of processors, operation, data type, etc.
Reduce Performance

![Graph showing the impact of different network structures on reduce performance]

- Linear
- Binary
- Binomial
- Chain 8
- Chain 4
- Chain 2
- Chain 1

The graph illustrates the minimum duration (in seconds) required for different total message sizes (in bytes) for various network structures, highlighting the reduction in performance as message size increases.
Useful Reading Material

1) CSMA-CD

2) Myrinet
   http://www.myricom.com/myrinet/overview/

3) InfiniBand
   http://www.infostor.com/articles/article_display.cfm?Section=ARCHI&C=Newst&ARTICLE_ID=248655
Review Questions

- Know these acronyms: FTP, HTTP, MPI, TCP, UDP, IP, SP Switch, HIPPI, SCI, CSMA-CD, Berkeley NOW project, NIC, DMA, GigE, IBA, RAID, PCI, TCA, HCA, RDMA, QoS, SDP, VIA, MVAPICH.
- What, exactly, is 10Gbps?
- Define latency, throughput
- Name, in order from bottom to top, the network layers that we talked about in class. Describe in a sentence the primary function of each layer.
- What are the top three most prominent networks on the Top 500 list?
- Describe five important features of Gigabit Ethernet.
- Give the steps of CSMA-CD. Why is CSMA-CD bad for cluster architectures? In what cases will it not be a hindrance?
- Give four ways that 10Gbps Ethernet is different from Gigabit Ethernet.
- Compare and contrast Myrinet and InfiniBand on at least three criteria.
Review Questions, continued

- Describe five important features of Myrinet.
- Suppose that the transmission rate of my network is 2Gbps, that my cables are 50m long, and that the signal travels in the wire at the rate of 200m/usec. How large should the buffers be to support a STOP and GO protocol correctly? Justify your answer.
- Describe five important features of InfiniBand.
- Place each of these into the correct network layer, according to the description we used in class: TCP, IP, CSMA-CD, credit-based flow control in InfiniBand, source routing in Ethernet, Virtual Lanes, fiber cable, GM, SDP, MPI
- Describe six important concepts in low-latency messaging.
- What is pinned memory? Why don’t we want users to do it?
- Explain how non-blocking sends and receives work.
- Describe two differences between and eager protocol and a rendezvous protocol in a low-latency messaging application?
- Suppose you are given $50,000 to purchase a computer cluster. What network do you choose and why? You just justify your answer for credit. What if you are given $500,000?