Networking: Hardware and Protocols

Jasjeet Singh Bagla

Harish-Chandra Research Institute

Allahabad, India

jasjeet@mri.ernet.in

Networks

- There are, broadly speaking, two types of networks. Switched and Direct connection.

- Typically, these networks use the PCI bus for interfacing with the computer. Connecting directly to the North Bridge has been proposed for Infiniband.

- Several protocols are in use: TCP/IP, VIA, AM, FM, etc.

Ethernet Network

- This is the most commonly used network. It is typically used as a switched network, though un-switched configuration can be implemented for a small number of hosts.
- Latency is typically around 10^2 μs.
- Commonly used implementation operates at 100 Mbps, though Gigabit Ethernet is becoming increasingly popular and 10 Gigabit products are being released.
- Typically this is used with the TCP/IP protocol.
- Using other protocols like VIA or Active Messages can improve performance significantly.

TCP/IP

- This protocol is used in local area networks.
    - Has a layered structure.
    - Message is copied across these layers.
    - Message is divided into small packets before transmission.
    - Path of data is decided at every hop.
    - This introduces delays and latency is very poor.

Virtual Interface Architecture

- Transmission of data is negotiated with the destination. Route of data packets is decided in advance.
    - Header has information about all the hops along the way.
    - Can be implemented in hardware or emulated in software.
    - Is used in Myrinet and Param-net.
    - Software emulation for Ethernet is available.

Active Messages

- Direct access to hardware level buffers is provided in user space.
  - One sided communications.
  - Open source implementation for Linux/Ethernet (GAMMA).
  - Can co-exist with TCP/IP.
  - Very low latencies can be achieved, e.g., 20μs on Ethernet.
  - MPICH has been adapted for GAMMA.

Switched Networks

- Switched networks offer good bandwidth for multiple concurrent communications. Of course, back-plane of switches plays a crucial role here.

  - Ethernet. Bandwidth $\sim 1Gbps$, Latency $\simeq 100\ \mu$s.
  - Myrinet. Bandwidth $\sim 2Gbps$, Latency $\simeq 5\ \mu$s.
  - Param-net. Bandwidth $\sim 2Gbps$, Latency $\simeq 10\ \mu$s.
  - Infiniband. Bandwidth $\sim 6Gbps$, Latency $\simeq 10\ \mu$s.

Switches

- Backplane of the switch is very important, unless it is a X-bar switch.
- If a hierarchy of switches is used then switch-switch connections should have a higher bandwidth. Otherwise nodes connected to the same switch have a better connectivity.
- A network is said to have full bisection if any two halves of the network have the same connectivity irrespective of how we divide the network into these parts. Such a network is called Clos network.

Switches

- Switched networks, if backplane is not a bottleneck, offer a good all to all connectivity.
- Number of switches required increases in proportion with (`n log(n)`).
- Cost of switched networks is often dominated by the cost of switches, though the cost of network interface cards for high performance networks is also very high.

Unswitched Networks

- Rings. (Co-axial cables, SCI/Wulfkit)
    - Each machine connects to two neighbors.
    - Maximum number of hops for a message is $n/2$.
    - Failure of one node can lead to a breakdown of the network.

Unswitched Networks

- Torus/Mesh. (SCI/Wulfkit)
  - Each machine connects to four neighbors.
  - Maximum number of hops for a message is $2\sqrt{n}$.
  - Failure of one node does not disrupt the network.
  - Failure of more than one node disrupts messages.

Unswitched Networks

- 3d Torus. (SCI/Wulfkit)
  - Each machine connects to six neighbors.
  - Maximum number of hops for a message is $3n^{1/3}$.
  - Failure of one or two nodes does not disrupt the network.
  - Failure of more than two nodes disrupts messages.

Unswitched Networks

- Hyper-Cube.
  - Machines are connected along a k dimensional cube, $n = 2^k$.
  - There are at most k hops between nodes.
  - A total of $k * n$ cables are needed.

Unswitched Networks

- Fully connected.
    - All machines are connected to each other.
    - A total of $n^2/2$ cables are needed.
    - For $n = 128$, we require 8192 cables.

Unswitched Networks

- Topologies mentioned here can also be realized using switched networks.

  - Routing in multiply connected networks is a complex problem.
  - As long as the number of hops is small (~ 10), switch-less technologies like SCI/Wulfkit offer good performance. These have a latency of 5μs and a point to point bandwidth of more than 2Gbps.
  - Cost of these networks grows in proportion with the number of nodes.

Networks

- There are other solutions like Quadrics, Cray interconnect, etc. that have better performance than the solutions discussed here.

- Most networking solutions used here interface through the PCI bus. This will restrict bandwidth to less than 8 Gbps in near future.

- Infiniband can potentially interface through the memory bus directly. This will reduce latency by a significant amount and also remove the bottleneck on bandwidth.

Networks

- In last 30 years CPU speed has improved by a factor of $3,000$.

◦ In the same period, memory latency has improved by a factor of 30. This lead to the development of cache in order to hide the widening gap in performance of memory and CPUs.

◦ In the same period, latency has improved by a factor of 10 for Ethernet. For networks in general, it has improved by a factor of 50. How will technology be used to address this gap further?