Lecture 28: Networks & Interconnect—
Architectural Issues

Professor Randy H. Katz
Computer Science 252
Spring 1996
Review: ABCs of Networks

• **Starting Point**: Send bits between 2 computers

  ![Diagram of two computers connected by a line with arrows indicating data flow]

• Queue on each end

• Can send both ways (“Full Duplex”)

• Rules for communication? “protocol”
  – Inside a computer?
  – Loads/Stores: Request(Address) & Response (Data)
  – Need Request & Response
  – Name for standard group of bits sent: **Packet**
Review: Questions about Simple Network

- What if more than 2 computers want to communicate?
  - Need computer address field in packet
- What if packet is garbled in transit?
  - Add error detection field in packet
- What if packet is lost?
  - More elaborate protocols to detect loss
- What if multiple processes/machine?
  - Queue per process
- Questions such as these lead to more complex protocols and packet formats
## Review: Implementation Issues

<table>
<thead>
<tr>
<th>Interconnect</th>
<th><strong>MPP</strong></th>
<th><strong>LAN</strong></th>
<th><strong>WAN</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>CM-5</td>
<td>Ethernet</td>
<td>ATM</td>
</tr>
<tr>
<td>Maximum length between nodes</td>
<td>25 m</td>
<td>500 m;</td>
<td>copper: 100 m</td>
</tr>
<tr>
<td></td>
<td>≤5 repeaters</td>
<td></td>
<td>optical(multimode): 2 km</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>optical(single mode): 25 km</td>
</tr>
<tr>
<td>Number data lines</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Clock Rate</td>
<td>40 MHz</td>
<td>10 MHz</td>
<td>≥ 155.5 MHz</td>
</tr>
<tr>
<td>Shared vs. Switch</td>
<td>Switch</td>
<td>Shared</td>
<td>Switch</td>
</tr>
<tr>
<td>Maximum number of nodes</td>
<td>2048</td>
<td>254</td>
<td>&gt; 10,000</td>
</tr>
<tr>
<td>Media Material</td>
<td>Copper</td>
<td>Twisted pair</td>
<td>Twisted pair copper wire or optical fiber</td>
</tr>
<tr>
<td></td>
<td></td>
<td>copper wire or Coaxial cable</td>
<td></td>
</tr>
</tbody>
</table>
• Overhead: latency of interface vs. Latency: network
**Review: Impact of Overhead on Delivered BW**

- **BW model:** Time = overhead + msg size/peak BW
- > 50% data transferred in packets = 8KB

<table>
<thead>
<tr>
<th>Peak BW (MB/sec)</th>
<th>MinTime (one-way μsecs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>1</td>
</tr>
<tr>
<td>1.00</td>
<td>10</td>
</tr>
<tr>
<td>10.00</td>
<td>100</td>
</tr>
<tr>
<td>100.00</td>
<td>1000</td>
</tr>
<tr>
<td>1000.00</td>
<td></td>
</tr>
</tbody>
</table>

**Diagram:**
- The graph illustrates the relationship between peak bandwidth and delivered bandwidth over different min times.
- The model shows how overhead affects the delivered bandwidth compared to peak bandwidth.

**Legend:**
- Red square: 1
- Green square: 10
- Blue square: 100
- Yellow square: 1000

---

RHK.S96 6
Review: Interconnect Issues

- Implementation Issues
- Performance Measures
- Architectural Issues
- Practical Issues
# Example Architecture Measures

<table>
<thead>
<tr>
<th>Interconnect</th>
<th>MPP</th>
<th>LAN</th>
<th>WAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example Topology</td>
<td>CM-5</td>
<td>Ethernet</td>
<td>ATM</td>
</tr>
<tr>
<td></td>
<td>“Fat” tree</td>
<td>Line, Bus</td>
<td>Variable, constructed from multistage switches</td>
</tr>
</tbody>
</table>

| Connection based? | No        | No          | Yes                      |
| Data Transfer Size| Variable: 4 to 20B | Variable: 0 to 1500B | Fixed: 48B |
| Store & Forward?  | No        | n.a.        | Yes                      |
| Congestion control| At source: Flow control via back pressure | At source: Listen for E-net idle | Rate based via choke packets |
Topology

• Structure of the interconnect

• Determines
  – **Degree**: number of links from a node
  – **Diameter**: max number of links crossed between nodes
  – **Average distance**: number of hops to random destination
  – **Bisection**: minimum number of links that separate the network into two halves

• **Warning**: these three-dimensional drawings must be mapped onto chips and boards which are essentially two-dimensional media
  – Elegant when sketched on the blackboard may look awkward when constructed from chips, cables, boards, and boxes
## Important Topologies

<table>
<thead>
<tr>
<th>Type</th>
<th>Degree</th>
<th>Diameter</th>
<th>Ave Dist</th>
<th>Bisection</th>
<th>Diam</th>
<th>Ave D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D mesh</td>
<td>≤ 2</td>
<td>N-1</td>
<td>N/3</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2D mesh</td>
<td>≤ 4</td>
<td>2(N^{1/2} - 1)</td>
<td>2N^{1/2} / 3</td>
<td>N^{1/2}</td>
<td>63</td>
<td>21</td>
</tr>
<tr>
<td>3D mesh</td>
<td>≤ 6</td>
<td>3(N^{1/3} - 1)</td>
<td>3N^{1/3} / 3</td>
<td>N^{2/3}</td>
<td>~30</td>
<td>~10</td>
</tr>
<tr>
<td>nD mesh</td>
<td>≤ 2n</td>
<td>n(N^{1/n} - 1)</td>
<td>nN^{1/n} / 3</td>
<td>N^{(n-1) / n}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ring</td>
<td>2</td>
<td>N / 2</td>
<td>N/4</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2D torus</td>
<td>4</td>
<td>N^{1/2}</td>
<td>N^{1/2} / 2</td>
<td>2N^{1/2}</td>
<td>32</td>
<td>16</td>
</tr>
<tr>
<td>k-ary n-cube</td>
<td>2n</td>
<td>n(N^{1/n})</td>
<td>nN^{1/n/2}</td>
<td>15</td>
<td>8</td>
<td>(3D)</td>
</tr>
<tr>
<td>(N = k^n)</td>
<td></td>
<td>nk/2</td>
<td>nk/4</td>
<td>2k^{n-1}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypercube</td>
<td>n</td>
<td>n = LogN</td>
<td>n/2</td>
<td>N/2</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Cube-Connected Cycles</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Topologies (cont)

<table>
<thead>
<tr>
<th>Type</th>
<th>Degree</th>
<th>Diameter</th>
<th>Ave Dist</th>
<th>Bisection</th>
<th>Diam</th>
<th>Ave D</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D Tree</td>
<td>3</td>
<td>$2 \log_2 N$</td>
<td>$\sim 2 \log_2 N$</td>
<td>1</td>
<td>20</td>
<td>$\sim 20$</td>
</tr>
<tr>
<td>4D Tree</td>
<td>5</td>
<td>$2 \log_4 N$</td>
<td>$2 \log_4 N - 2/3$</td>
<td>1</td>
<td>10</td>
<td>9.33</td>
</tr>
<tr>
<td>$kD$</td>
<td>$k+1$</td>
<td>$\log_k N$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2D fat tree</td>
<td>4</td>
<td>$\log_2 N$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2D butterfly</td>
<td>4</td>
<td>$\log_2 N$</td>
<td>$N/2$</td>
<td></td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

N = 1024

![CM-5 Thinned Fat Tree](image)
**Butterfly**

**Multistage: nodes at ends, switches in middle**

- All paths equal length
- Unique path from any input to any output
- Conflicts

**Benes Network**

- Routes all permutations w/o conflict
- Notice similarity to Fat Tree
- Randomization is major breakthrough
Multistage Fat Tree

- Randomly assign packets to different paths on the way up to spread the load
## Example Networks

<table>
<thead>
<tr>
<th>Name</th>
<th>Number</th>
<th>Topology</th>
<th>Bits</th>
<th>Clock</th>
<th>Link</th>
<th>Bisect.</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>nCube/ten</td>
<td>1-1024</td>
<td>10-cube</td>
<td>1</td>
<td>10 MHz</td>
<td>1.2</td>
<td>640</td>
<td>1987</td>
</tr>
<tr>
<td>iPSC/2</td>
<td>16-128</td>
<td>7-cube</td>
<td>1</td>
<td>16 MHz</td>
<td>2</td>
<td>345</td>
<td>1988</td>
</tr>
<tr>
<td>MP-1216</td>
<td>32-512</td>
<td>2D grid</td>
<td>1</td>
<td>25 MHz</td>
<td>3</td>
<td>1,300</td>
<td>1989</td>
</tr>
<tr>
<td>Delta</td>
<td>540</td>
<td>2D grid</td>
<td>16</td>
<td>40 MHz</td>
<td>40</td>
<td>640</td>
<td>1991</td>
</tr>
<tr>
<td>CM-5</td>
<td>32-2048</td>
<td>fat tree</td>
<td>4</td>
<td>40 MHz</td>
<td>20</td>
<td>10,240</td>
<td>1991</td>
</tr>
<tr>
<td>CS-2</td>
<td>32-1024</td>
<td>fat tree</td>
<td>8</td>
<td>70 MHz</td>
<td>50</td>
<td>50,000</td>
<td>1992</td>
</tr>
<tr>
<td>Paragon</td>
<td>4-1024</td>
<td>2D grid</td>
<td>16</td>
<td>100 MHz</td>
<td>200</td>
<td>6,400</td>
<td>1992</td>
</tr>
<tr>
<td>T3D</td>
<td>16-1024</td>
<td>3D Torus</td>
<td>16</td>
<td>150 MHz</td>
<td>300</td>
<td>19,200</td>
<td>1993</td>
</tr>
</tbody>
</table>

No standard topology!
Connection-Based vs. Connectionless

• Telephone: operator sets up connection between the caller and the receiver
  – Once the connection is established, conversation can continue for hours

• Share transmission lines over long distances by using switches to multiplex several conversations on the same lines
  – “Time division multiplexing” divide B/W transmission line into a fixed number of slots, with each slot assigned to a conversation

• Problem: lines busy based on number of conversations, not amount of information sent

• Advantage: reserved bandwidth
Connection-Based vs. Connectionless

- **Connectionless**: every package of information must have an address => packets
  - Each package is routed to its destination by looking at its address, e.g., the postal system
  - Split phase buses are send packets
  - Statistical multiplexing
Packet Formats

- Fields: Destination, Checksum(C), Length(L), Type(T)
- Data/Header Sizes in bytes: (4 to 20)/4, (0 to 1500)/26, 48/5
Congestion Control

- Packet switched networks do not reserve bandwidth; this leads to *contention*.
- Solution: prevent packets from entering until contention is reduced (e.g., metering lights).
- Options:
  - Packet discarding: If a packet arrives at a switch and there is no room in the buffer, the packet is discarded.
  - Flow control: between pairs of receivers and senders; use feedback to tell the sender when it is allowed to send the next packet.
    - Back-pressure: separate wires to tell to stop.
    - Window: give the original sender the right to send N packets before getting permission to send more (overlap the latency of the interconnection with the overhead to send and receive a packet).
  - Choke packets: aka “rate-based”; Each packet received by busy switch in warning state sent back to the source via choke packet. Source reduces traffic to that destination by a fixed % (ATM Forum).
Store and Forward vs. Cut-Through

• **Store-and-forward policy**: each switch waits for the full packet to arrive in the switch before it is sent on to the next switch.

• **Cut-through routing** or **worm hole routing**: switch examines the header, decides where to send the message, and then starts forwarding it immediately.

  – In worm hole routing, when the head of the message is blocked the message stays strung out over the network, potentially blocking other messages (needs only buffer the piece of the packet that is sent between switches). CM-5 uses it, with each switch buffer being 4 bits per port.

  – Cut through routing lets the tail continue when the head is blocked, accordioning the whole message into a single switch. (Requires a buffer large enough to hold the largest packet).
Store and Forward vs. Cut-Through

• Advantage
  – Latency reduces from function of:
    number of intermediate switches \( X \) by the size of the packet
to
  time for 1st part of the packet to negotiate the switches + the packet size ÷ interconnect BW
Switching

w = wire width, m = message size
b = m/w, H = number of hops
R = delay per router

- **Circuit switching**
  - Establish end-to-end route, then transmit data
    \[ L = HR_c + b \]

- **Packet switching**
  - Route data as it moves forward
  - Store-and-forward: \[ L = H(b + R_{sf}) \]
  - Cut-through: \[ L = HR_{ct} + b \]
  - Typically, \( R_{ct} = \frac{32}{w} \)
  - Most modern networks are about 1 small packet deep
Examples

<table>
<thead>
<tr>
<th>Machine</th>
<th>Network</th>
<th>W</th>
<th>R</th>
<th>rate MHz</th>
<th>MB/s per link</th>
</tr>
</thead>
<tbody>
<tr>
<td>nCUBE/2</td>
<td>13-D hypercube</td>
<td>1</td>
<td>40</td>
<td>40</td>
<td>2.2</td>
</tr>
<tr>
<td>CM-5</td>
<td>4-D Thinned Fat Tree</td>
<td>4</td>
<td>8</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td>Delta</td>
<td>2-D Mesh</td>
<td>8</td>
<td>2</td>
<td>~40</td>
<td>40</td>
</tr>
<tr>
<td>T3D</td>
<td>3-D Torus</td>
<td>16</td>
<td>?</td>
<td>150</td>
<td>300</td>
</tr>
</tbody>
</table>

- What about performance? Degree vs. Distance
  - Generally, topologies with larger distance have better physical layout, so they have shorter, wider links and smaller routing delays. Hence, latencies are similar: $L = HR_{ct} + \frac{m}{w}$ (H increases, b decreases)
    - Low degree networks have few links per processor
    - If each message travels a larger number of links => few outstanding messages per processor
    - With skinny wires, message occupies path for a long time
Routing

• Deterministic—follows a pre-specified route
  – mesh: dimension-order routing
    » $(x_1, y_1) -> (x_2, y_2)$
    » first $\Delta x = x_2 - x_1$,
    » then $\Delta y = y_2 - y_1$,
  – hypercube: edge-cube routing
    » $X = x_0x_1x_2...x_n -> Y = y_0y_1y_2...y_n$
    » $R = X \text{xor} Y$
    » Traverse dimensions of differing address in order
  – tree: common ancestor

• Adaptive—route based on network state (e.g., contention)

• Deadlock free
## Practical Issues

<table>
<thead>
<tr>
<th>Interconnection</th>
<th>MPP</th>
<th>LAN</th>
<th>WAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
<td>CM-5</td>
<td>Ethernet</td>
<td>ATM</td>
</tr>
<tr>
<td>Standard</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Fault Tolerance?</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Hot Insert?</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- **Standards**: required for WAN, LAN!
- **Fault Tolerance**: Can nodes fail and still deliver messages to other nodes? required for WAN, LAN!
- **Hot Insert**: If the interconnection can survive a failure, can it also continue operation while a new node is added to the interconnection? required for WAN, LAN!
Protocols: HW/SW Interface

• Internetworking: allows computers on independent and incompatible networks to communicate reliably and efficiently;
  – Enabling technologies: SW standards that allow reliable communications without reliable networks
  – Hierarchy of layers, giving each layer responsibility for portion of overall communications task, called protocol families or protocol suites

• Transmission Control Protocol/Internet Protocol (TCP/IP)
  – This protocol family is the basis of the Internet
Protocol

- Key to protocol families is that communication occurs logically at the same level of the protocol, called peer-to-peer, but is implemented via services at the lower level.
- Danger is each level increases latency.
FTP From Stanford to Berkeley

- BARRNet is WAN for Bay Area
- T1 is 1.5 mbps leased line; T3 is 45 mbps; FDDI is 100 mbps LAN
- IP sets up connection, TCP sends file
Summary: Interconnections

- Communication between computers
- Packets for standards, protocols to cover normal and abnormal events
- Implementation issues: length, width, media
- Performance issues: overhead, latency, bisection BW
- Topologies: many to chose from, but (SW) overheads make them look alike; cost issues in topologies