Developments in High-Performance Interconnection Networks

Timothy Mark Pinkston

*SMART* Interconnects Group
Electrical Engineering – Systems Department
University of Southern California, Los Angeles
http://www.usc.edu/dept/ceng/pinkston/SMART.html

*Superior Multiprocessor ARchiTecture*
Introduction

The need for high-performance networking exists at many system levels:

- Metropolitan and Local Area Networks (MANs/LANs)
- Networks of Workstations and System Area Networks (NOWs/SANs)
- Storage Area Networks (STANs) and Server-I/O

The switch/router is critical to the types of services and performance provided by the network at each of these levels.
High-Performance Networking
How to increase system bandwidth capacity?

- high-speed dense I/O

\[
T_{\text{latency}} = T_c \left( \text{hop\_distance} \times \frac{\text{packet\_size}}{\text{channel\_width}} \right) + T_{\text{contention}}
\]
How to utilize system resources efficiently?
- cut-through switching
- adaptive routing
- virtual channel flow

How to increase system bandwidth capacity?
- high-speed dense I/O

\[ T_{\text{latency}} = T_c (hop_{\text{distance}} + \frac{\text{packet\_size}}{\text{channel\_width}}) + T_{\text{contention}} \]
Networking at the MAN/LAN Level
# Ethernet & Link Technologies

<table>
<thead>
<tr>
<th></th>
<th>Ethernet 10 Base T</th>
<th>Fast Ethernet 100 Base T</th>
<th>Gigabit Ethernet 1000 Base X</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Rate</strong></td>
<td>10 Mbps</td>
<td>100 Mbps</td>
<td>1 Gbps</td>
</tr>
<tr>
<td><strong>Category 5 UTP</strong></td>
<td>100 m</td>
<td>100 m</td>
<td>10’s m*</td>
</tr>
<tr>
<td><strong>STP/Coax-Cable</strong></td>
<td>500 m</td>
<td>100 m</td>
<td>25 m*</td>
</tr>
<tr>
<td><strong>Multimode Fiber</strong></td>
<td>2 km</td>
<td>2 km</td>
<td>550 m</td>
</tr>
<tr>
<td><strong>Single-mode Fiber</strong></td>
<td>25 km</td>
<td>20 km</td>
<td>5 km</td>
</tr>
</tbody>
</table>

Note: technical decisions made by the IEEE 802.3 standards group

UTP = unshielded twisted pair, STP = shielded twisted pair

* = expected to be finalized by IEEE 802.3ab
CoreBuilder 9000 (3Com)

- Main components: backplane, central switching fabric, and switching module
- Throughput: Up to 48Gbps aggregate bandwidth
- Scalability: Provides up to 16 switching module slots:
  - Gigabit Ethernet switching module: 9 1000BASE-SX (1Gbps) ports
  - Fast Ethernet switching module: up to 36 10/100 Mbps ports

- Services: Support wire-speed interfaces for Gigabit/Fast Ethernet, ATM, FDDI, etc.
- High availability: full redundancy for critical components, e.g., switching fabric: primary and backup, and distributed switching modules
- Reach: Up to 2km with Multimode fiber, up to 40km with single-mode fiber and long reach interface
Cisco 12016 Gigabit Switch/Router (GSR)

- **Scalability and throughput:** a gigabit router processor (GRP), a switch fabric, and up to 16 line cards having 4 1.25Gbps serial lines (80 Gbps switching capacity)
- **Flexibility:** GRP determines network topology, calculates best routing paths and updates routing tables of line cards (which perform packet forwarding)
- Routing table size in a line card: 64MB (up to 1 million routing entries)
- **Services:** Optimized for IP routing (delivers Layer 3 routing at gigabit speeds)
- Supports CoS/QoS: weighted fair queuing (WFQ), random early detection (RED) and weighted RED (wRED)
Cajun™ P550 Gigabit Switch (Lucent)

- **Scalable:** up to 24 Gigabit, 120 10/100 Mbps, or up to 60 fiber-optic Fast Ethernet ports
- **Throughput:** 22.88 Gbps switching capacity
- **Reach:** Up to 10Km using single-mode fiber

- 24,000 address forwarding table entries and 240,000 maximum flows
- **Services:** Supports fault-tolerance with full redundancy & hot-swapping
- Signal-based QoS using RSVP and queue engine (traffic priorities)
- **Reach:** Up to 10Km using single-mode fiber
BlackDiamond 6800 and Summit 24

BlackDiamond 6800

- **Scalable:** BlackDiamond 6800: 48 gigabit or 256 10/100 Mbps Ethernet ports
  Summit 24: 24 10/100 Mbps and one gigabit Ethernet ports
- **Throughput:** Layer 3 switching (route 48 million packets per second)
- **Services:** Non-blocking, wire-speed IP routing on all ports
- **Policy-based QoS:** using port-based VLANs (virtual LANs)
- **Hot swappable chassis** components plus dual management and switch fabric maximize network up-time
Terabit Switch/Router TSR (Avici)

- **Scalability and Throughput**: up to 560 OC-192 (10Gbps) ports: 5.6 Tbps aggregate BW
- **TSR chassis** supports 40 line cards:
  - Each line card can handle 7 million packets per second at either layer 2 or layer 3 switching
- **Services**: Supports CoS/QoS: using WFQ, RED and DiffServ (Differentiated Services) for end-to-end QoS
- **Remote network configuration and management** via telnet, Simple Network Management Protocol (SNMP) or Nortel Network’s INM management software.
Pluris 20000 Series Terabit Network Router (TNR)

- **Scalable**: From 1 to 128 shelves per system and 1 to 15 line cards per shelf (maximum of 1920 per system)
- Line cards support either 16 OC-12c (622Mbps) ports, 4 OC-48 (2.4Gbps) ports, or a single OC-192 (10Gbps) port
- **Throughput**: up to 150 Gbps per shelf, up to 19.2 Tbps per system using n-dimensional fiber optic interconnect
- Line card capability: 33 million packets per second

- **Services**: High availability (**99.999% up-time**): shelf level redundancy (switch fabric, control cards, line cards, power, fans)
- **IP DiffServ** for **QoS** Traffic management (CBR, VBR, WFQ, RED)
Networking at the NOW/SAN Level
Example of a NOW/SAN
Examples of Regular Topologies for Direct Networks

- Mesh (i.e., 3-ary 3-cube)
- Hypercube (i.e., 2-ary 4-cube)
- Torus (i.e., 4-ary 2-cube)
Switch-based Irregular Networks in LAN settings:

Switch (8-ports)

PEs Connected to Switches

Irregular Network Topology
## Examples of Multiprocessor which use Direct Networks

<table>
<thead>
<tr>
<th></th>
<th>year</th>
<th>topology</th>
<th>virtual channels</th>
<th>cut-through switching</th>
<th>adaptive routing</th>
</tr>
</thead>
<tbody>
<tr>
<td>nCube2, nCube3</td>
<td>1988</td>
<td>Hypercube</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intel Paragon</td>
<td>1992</td>
<td>2D Mesh</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MIT J-Machine</td>
<td>1992</td>
<td>3D Mesh</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cray T3D</td>
<td>1993</td>
<td>3D Torus</td>
<td>✓ (4)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cray T3E</td>
<td>1996</td>
<td>3D Torus</td>
<td>✓ (5)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SGI Origin 2000</td>
<td>1997</td>
<td>“Fat Bristled” Hypercube</td>
<td>✓ (4)</td>
<td>✓</td>
<td>✓ (partial)</td>
</tr>
</tbody>
</table>
### Examples of Switches Available for Use in NOWs

<table>
<thead>
<tr>
<th>Switch</th>
<th>Year</th>
<th>Topologies</th>
<th>Virtual Channels</th>
<th>Cut-through Routing</th>
<th>Adaptive Routing</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEC Autonet</td>
<td>1990</td>
<td>arbitrary</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Myrinet</td>
<td>1994</td>
<td>arbitrary</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ServerNet</td>
<td>1995</td>
<td>arbitrary</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>SGI Spider</td>
<td>1996</td>
<td>arbitrary</td>
<td>✓ (4)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Intel Cavallino</td>
<td>1996</td>
<td>arbitrary</td>
<td>✓ (4)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ServerNet II</td>
<td>1997</td>
<td>arbitrary</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
Sun Gigabit Switch

- **Scalability and Throughput:** 2 1Gbps and 8 10/100Mbps Ethernet ports
- **Gigabit Layer 3** Ethernet switch
- **Services:** Gigabit Ethernet PCI adapter: data forwarding hardware, dual DMA channels, Ethernet MAC (compatibility)
- **Buffer memory:** 256KB on 10/100 Mbps ports, 1MB on each 1Gbps port
- **Reach:** Up to 500m with 50 micron multimode optic fiber
Tandem ServerNet II (Compaq)

- System area network (SAN) switch
- Scalability: 6 1.25+1.25 Gbps bi-directional ports
- Latency: 7.37 \(\text{us}\) for best kernel-to-kernel latency
  
  Wormhole switching: less than 300 ns per switch
- Services: Fault-tolerance with redundancy: two parallel network layers
- Reach: Up to 5km single mode fiber, up to 25m 1000BaseCX copper
Myrinet Switch

- Scalable: 4 to 16 1.28 Gbps LAN(L)/SAN(S) ports
- Cut-through: L-L (300ns), L-S (200ns), S-S (100ns)
- Arbitrary network, source routing, phys. chan.
- Power consumption: 6 to 32 Watts
- Price range: $1200 (4 ports) to $6000 (16 ports)

- Bandwidth: SAN cable: unshielded, 1.28Gbps up to 10 feet
- LAN cable: shielded, 0.64Gbps up to 60 feet, 1.28Gb/s up to 35 feet
- LAN/SAN optical converter/fiber: 1Gb/s up to 10 km (single-mode), up to 550m (multi-mode)
Networking at the STAN/Server-I/O Level

Components:
- CPU
- Memory
- SCSI controller
- VME interface
- ATM controller
- LAN controller
- PCI interface

Interconnections:
- TNet system area network
- Disks
- VME controller
- ATM
- Ethernet
- PCI controller

Terms:
- VME: Virtual Memory Extended (bus)
- ATM: Asynchronous Transfer Mode

Superior Multiprocessor Architecture Interconnects
Why Storage Area Networking (STAN)?

Limitations of existing network server connectivity (e.g., SCSI) to facilitate the rapid growth in data intensive, mission-critical applications

<table>
<thead>
<tr>
<th>Attribute</th>
<th>UltraSCSI Limit</th>
<th>FC-AL SAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Transmission</td>
<td>Half-Duplex</td>
<td>Full-Duplex</td>
</tr>
<tr>
<td>Effective Data Bandwidth</td>
<td>40 MB/sec</td>
<td>200 MB/sec</td>
</tr>
<tr>
<td>Protocol Support</td>
<td>SCSI</td>
<td>SCSI, IP, others</td>
</tr>
<tr>
<td>Connection Scalability</td>
<td>15 drives/bus</td>
<td>126 nodes/loop</td>
</tr>
<tr>
<td>Connection Distance</td>
<td>25m</td>
<td>10km</td>
</tr>
<tr>
<td>Relative Storage Capacity</td>
<td>136 Gbytes</td>
<td>9,172 Gbytes</td>
</tr>
</tbody>
</table>

- High Bandwidth
- Modular Scalability
- High Availability
- Manageability
- Ease of Integration
A Fibre Channel Arbitration-Loop STAN
Next-Generation Server-I/O (*InfiniBand*)

Problems of traditional server-I/O architecture, i.e., shared bus (PCI):
- **Low scalability**: $1/n$ effective bandwidth per port, low clock speed
- Suffer from arbitration, sharing and **single point of failure**

### Traditional I/O Arch

```
CPU  CPU
Mem Ctrl
PCI, PCI-X
SCSI  Enet  FC
```

### InfiniBand

```
CPU  Mem Ctrl  Sys Mem
Host interconnect
HCA  Link  Switch  Link  Target
xCA  Router  Network  xCA  Link
```

A next-generation server-I/O architecture, *InfiniBand*
- **Scalable bandwidth** (switch)
- Improved latency and fewer interrupts compared to current architectures
- **Unified interconnect** for IPC, network I/O, storage I/O
- Scalable to thousands of endpoints
- **High reliability** and **availability**
Conclusion

- The need for high-performance networking exists at many system levels, including MANs, LANs, NOWs, SANs, STANs, and Server-I/O.

- There is a trend toward increasing the scalability, types of services, and performance provided.

- The switch/router is critical to the types of services and performance provided by the network at all of these levels.
Sources:

- CoreBuilder 9000 (3Com)
  http://www.3com.com/solutions/key_net/gigabit/gigabit.html
- GSR 12016 (Cisco)
  http://www.cisco.com/univercd/cc/td/doc/pcat/12000.htm#CIHDCCHIA
- Cajun (Lucent)
- BlackDiamond 6800 (Extreme Networks, subsidiary of Hewlett-Packard)
  http://www.extremenetworks.com/technology/whitepapers/design.asp
- Terabit Switch/Router (TSR) (Avici)
- 20000 Series Terabit Network Router (TNR) (Pluris)
- Gigabit Switch (SUN)
  http://www.sun.com/products-n-solutions/hw/networking/
- ServerNet II (Tandem, subsidiary of Compaq)
  http://www.servernet.com/flat/public/prod_des/snsw1pd/snsw1pd.pdf
- Myrinet (Myricom)
  http://www.myri.com/myrinet/product_list.html