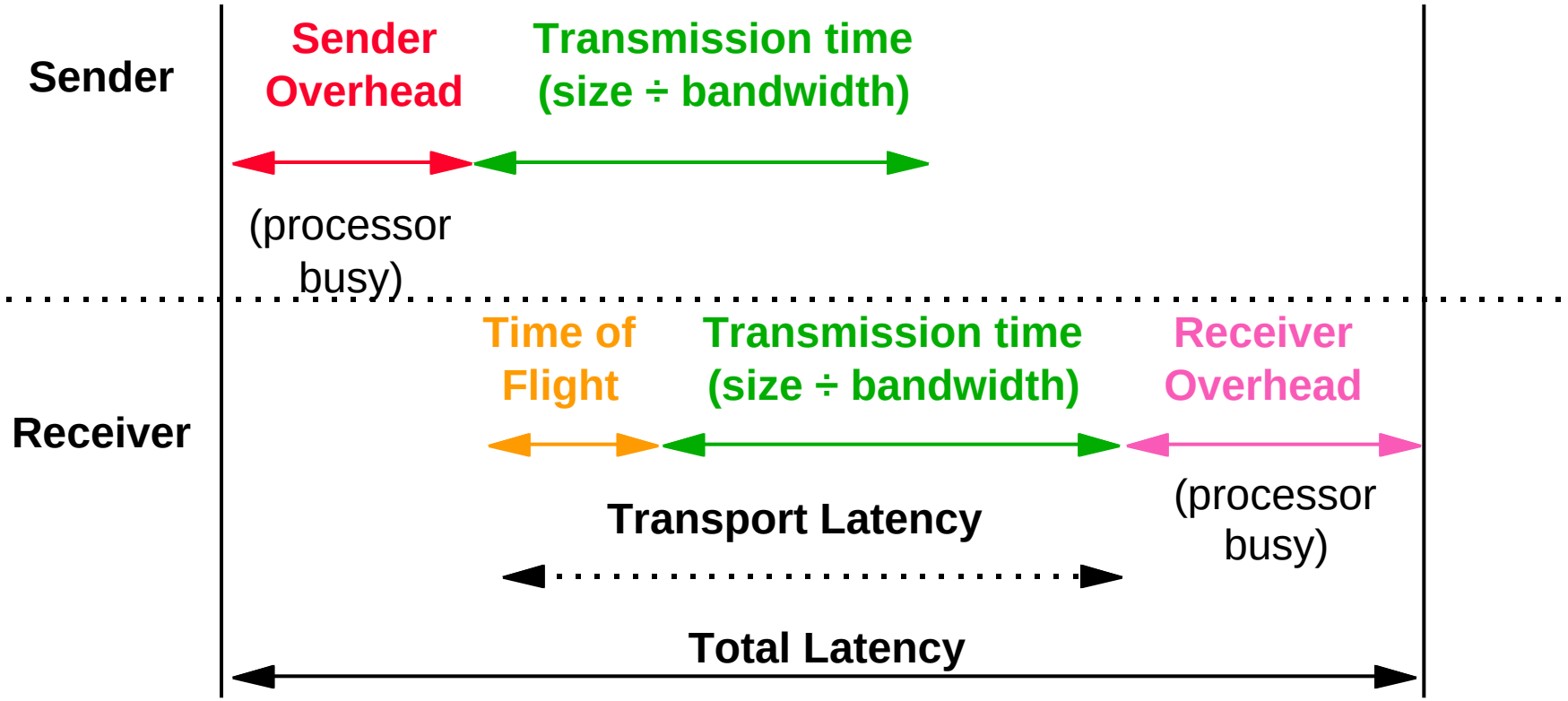
Lecture 16: Networks & Interconnect (Routing, Examples, Protocols) + Intro to Parallel Processing

Professor David A. Patterson

Computer Science 252

Spring 1998

Review: Performance Metrics



Total Latency = Sender Overhead + Time of Flight + Message Size ÷ BW + Receiver Overhead

Includes header/trailer in BW calculation?

Review: Interconnections

- Communication between computers
- Packets for standards, protocols to cover normal and abnormal events
- Performance issues: HW & SW overhead, interconnect latency, bisection BW
- Media sets cost, distance
- Shared vs. Switched Media determines BW
- HW and SW Interface to computer affects overhead, latency, bandwidth
- Topologies: many to chose from, but (SW) overheads make them look alike; cost issues in topologies, should not be programming issue

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

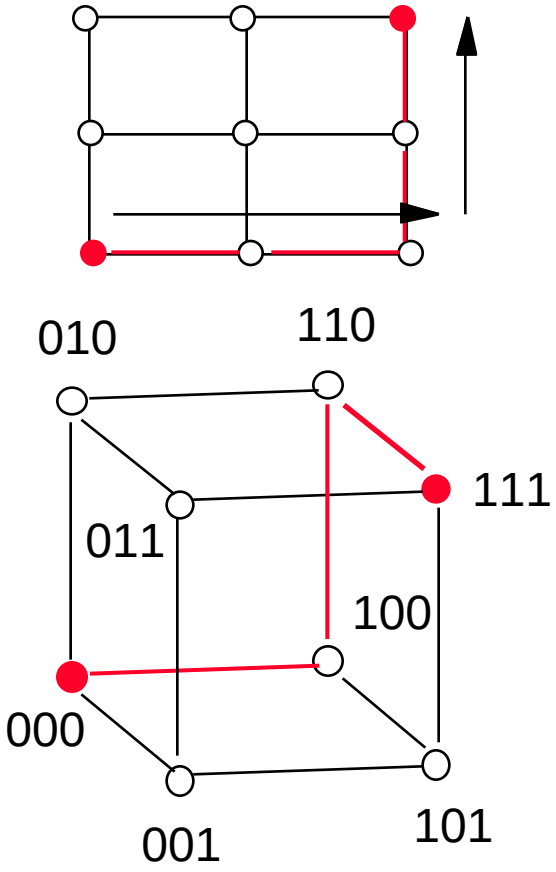
- **Connectionless: every package of information must have an address => packets**
  - **Each package is routed to its destination by looking at its address**
  - **Analogy, the postal system (sending a letter)**
  - **also called "Statistical multiplexing"**
  - **Note: "Split phase buses" are sending packets**

Routing Messages

- **Shared Media**
  - **Broadcast to everyone**
- **Switched Media needs real routing. Options:**
  - **Source-based routing: message specifies path to the destination (changes of direction)**
  - **Virtual Circuit: circuit established from source to destination, message picks the circuit to follow**
  - **Destination-based routing: message specifies destination, switch must pick the path**
    - » **deterministic: always follow same path**
    - » **adaptive: pick different paths to avoid congestion, failures**
    - » **Randomized routing: pick between several good paths to balance network load**

Deterministic Routing Examples

- **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**
  - **$X = x_o x_1 x_2 \ldots x_n$ -> $Y = y_o y_1 y_2 \ldots y_n$**
  - **R = X xor Y**
  - **Traverse dimensions of differing address in order**
- **tree: common ancestor**
- **Deadlock free?**



Store and Forward vs. Cut-Through

- **Store-and-forward policy: each switch waits for the full packet to arrive in switch before sending to the next switch (good for WAN)**
- **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 head of message is blocked, 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 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

Congestion Control

- **Packet switched networks do not reserve bandwidth; this leads to *contention* (connection based limits input)**
- **Solution: prevent packets from entering until contention is reduced (e.g., freeway on-ramp metering lights)**
- **Options:**
  - **Packet discarding: If packet arrives at switch and no room in buffer, packet is discarded (e.g., UDP)**
  - **Flow control: between pairs of receivers and senders; use feedback to tell sender when allowed to send next packet**
    - » **Back-pressure: separate wires to tell to stop**
    - » **Window: give original sender right to send N packets before getting permission to send more; overlapslatency of interconnection with overhead to send & receive packet (e.g., TCP), adjustable window**
  - **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 % (e.g., ATM)**

Practical Issues for Inteconnection Networks

- **Standardization advantages:**
  - **low cost (components used repeatedly)**
  - **stability (many suppliers to chose from)**
- **Standardization disadvantages:**
  - **Time for committees to agree**
  - **When to standardize?**
    - » **Before anything built? => Committee does design?**
    - » **Too early suppresses innovation**
- **Perfect interconnect vs. Fault Tolerant?**
  - **Will SW crash on single node prevent communication? (MPP typically assume perfect)**
- **Reliability (vs. availability) of interconnect**

Practical Issues

Interconnection	MPP	LAN	WAN
Example	CM-5	Ethernet	ATM
Standard	No	Yes	Yes
Fault Tolerance?	No	Yes	Yes
Hot Insert?	No	Yes	Yes

- 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!

Cross-Cutting Issues for Networking

- **Efficient Interface to Memory Hierarchy vs. to Network**
  - **SPEC ratings => fast to memory hierarchy**
  - **Writes go via write buffer, reads via L1 and L2 caches**
- **Example: 40 MHz SPARCStation(SS)-2 vs 50 MHz SS-20, no L2$ vs 50 MHz SS-20 with L2$ I/O bus latency; different generations**
- **SS-2: combined memory, I/O bus => 200 ns**
- **SS-20, no L2$: 2 busses +300ns => 500ns**
- **SS-20, w L2$: cache miss+500ns => 1000ns**

CS 252 Administrivia

- **Upcoming events in CS 252**

23-Mar to 27-Mar Spring Break

Wed 8-Apr Multiprocessors

Fri 10-Apr Multiprocessors

Wed 15-Apr Project Reviews: all day (no lecture)

Fri 17-Apr Searching the Computer Science Literature: Techniques & Tips by Camille Wanat

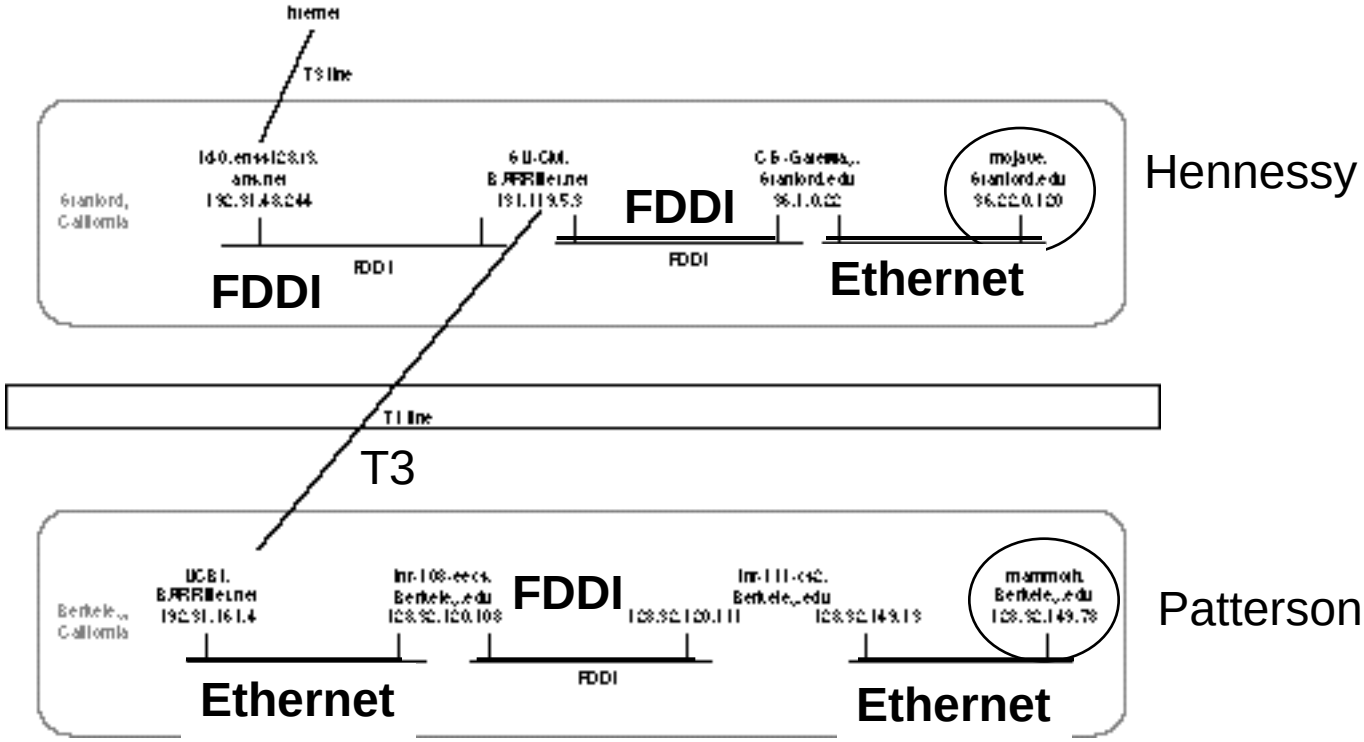
Wed 22-Apr Quiz # 2 5:30-8:30 (no lecture)

- **Next reading is Chapter 8 of CA:AQA 2/e and Sections 1.1-1.4, Chapter 1 of upcoming book by Culler, Singh, Gupta called "Parallel Computer Architecture-A Hardware/Software Approach"**
- **`www.cs.berkeley.edu/~culler/`**

Protocols: HW/SW Interface

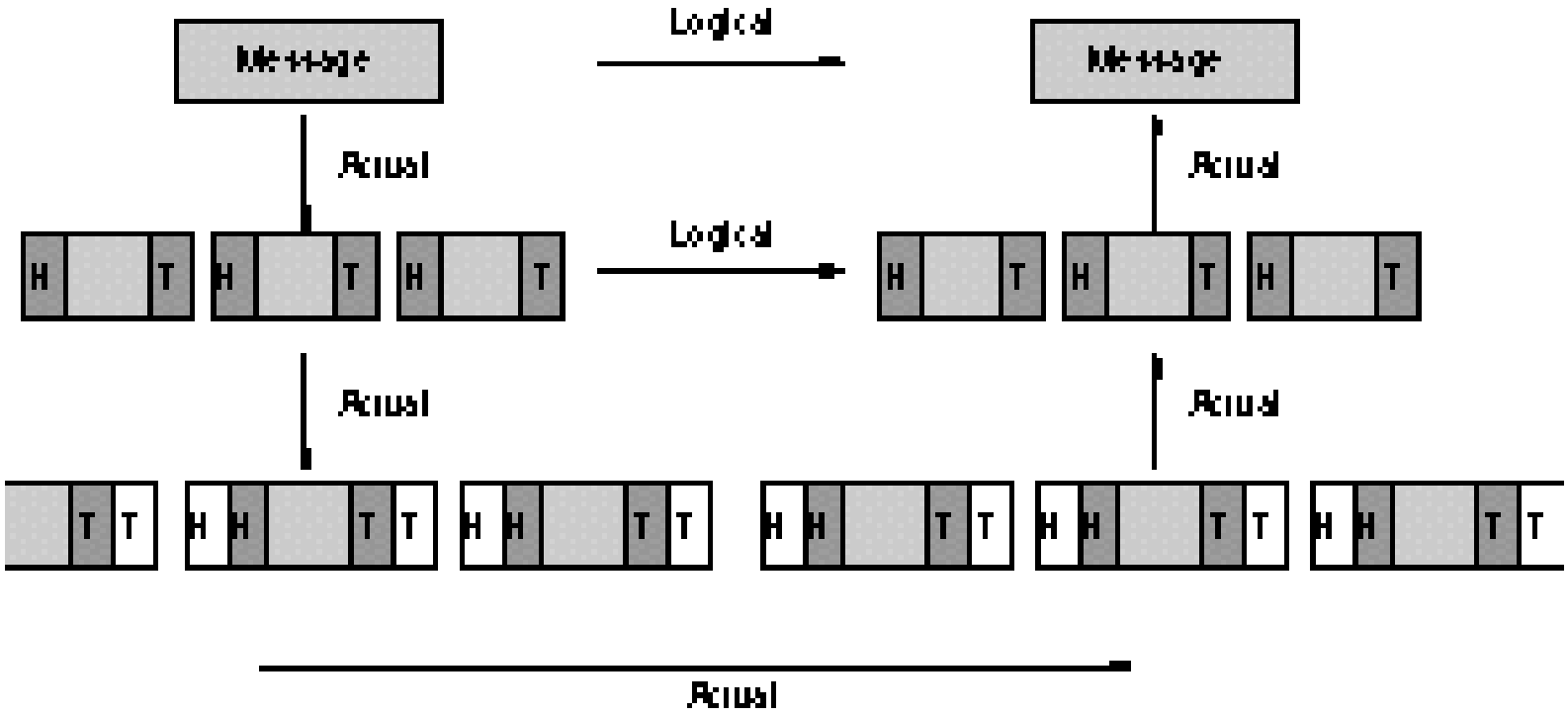
- Internetworking: allows computers on independent and incompatible networks to communicate reliably and efficiently;
  - Enabling technologies: SW standards that allow reliable communications without reliable networks
  - Hierarchy of SW layers, giving each layer responsibility for portion of overall communications task, called protocol families or protocol suites
- Transmission Control Protocol/Internet Protocol (TCP/IP)
  - This protocol family is the basis of the Internet
  - IP makes best effort to deliver; TCP guarantees delivery
  - TCP/IP used even when communicating locally: NFS uses IP even though communicating across homogeneous LAN

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**

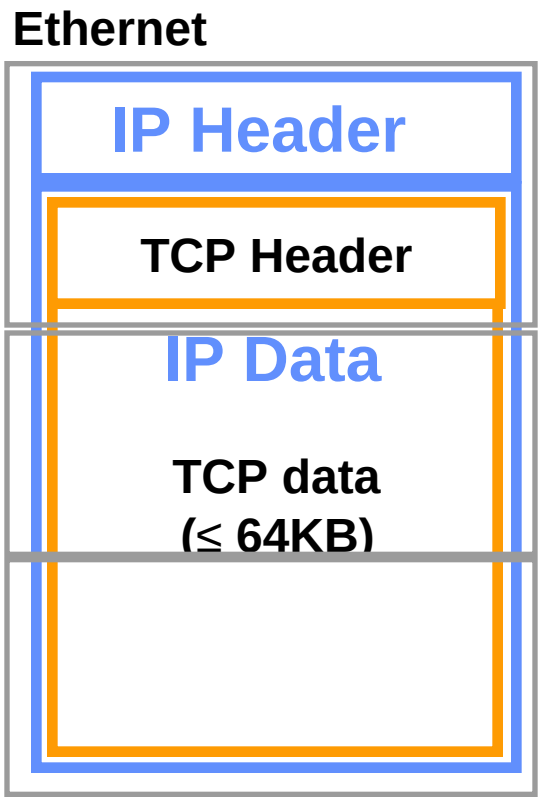
Protocol



- 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 if implemented as hierarchy (e.g., multiple check sums)

TCP/IP packet

- Application sends message
- TCP breaks into 64KB segements, adds 20B header
- IP adds 20B header, sends to network
- If Ethernet, broken into 1500B packets with headers, trailers
- Header, trailers have length field, destination, window number, version, ...



Example Networks

- **Ethernet: shared media 10 Mbit/s proposed in 1978, carrier sensing with expotential backoff on collision detection**
- **15 years with no improvement; higher BW?**
- **Multiple Ethernets with devices to allow Ehternets to operate in parallel!**
- **10 Mbit Ethernet successors?**
  - **FDDI: shared media (too late)**
  - **ATM (too late?)**
  - **Switched Ethernet**
  - **100 Mbit Ethernet (Fast Ethernet)**
  - **Gigabit Ethernet**

Connecting Networks

- **Bridges: connect LANs together, passing traffic from one side to another depending on the addresses in the packet.**
  - **operate at the Ethernet protocol level**
  - **usually simpler and cheaper than routers**
- **Routers or Gateways: these devices connect LANs to WANs or WANs to WANs and resolve incompatible addressing.**
  - **Generally slower than bridges, they operate at the internetworking protocol (IP) level**
  - **Routers divide the interconnect into separate smaller subnets, which simplifies manageability and improves security**
- **Cisco is major supplier; basically special purpose computers**

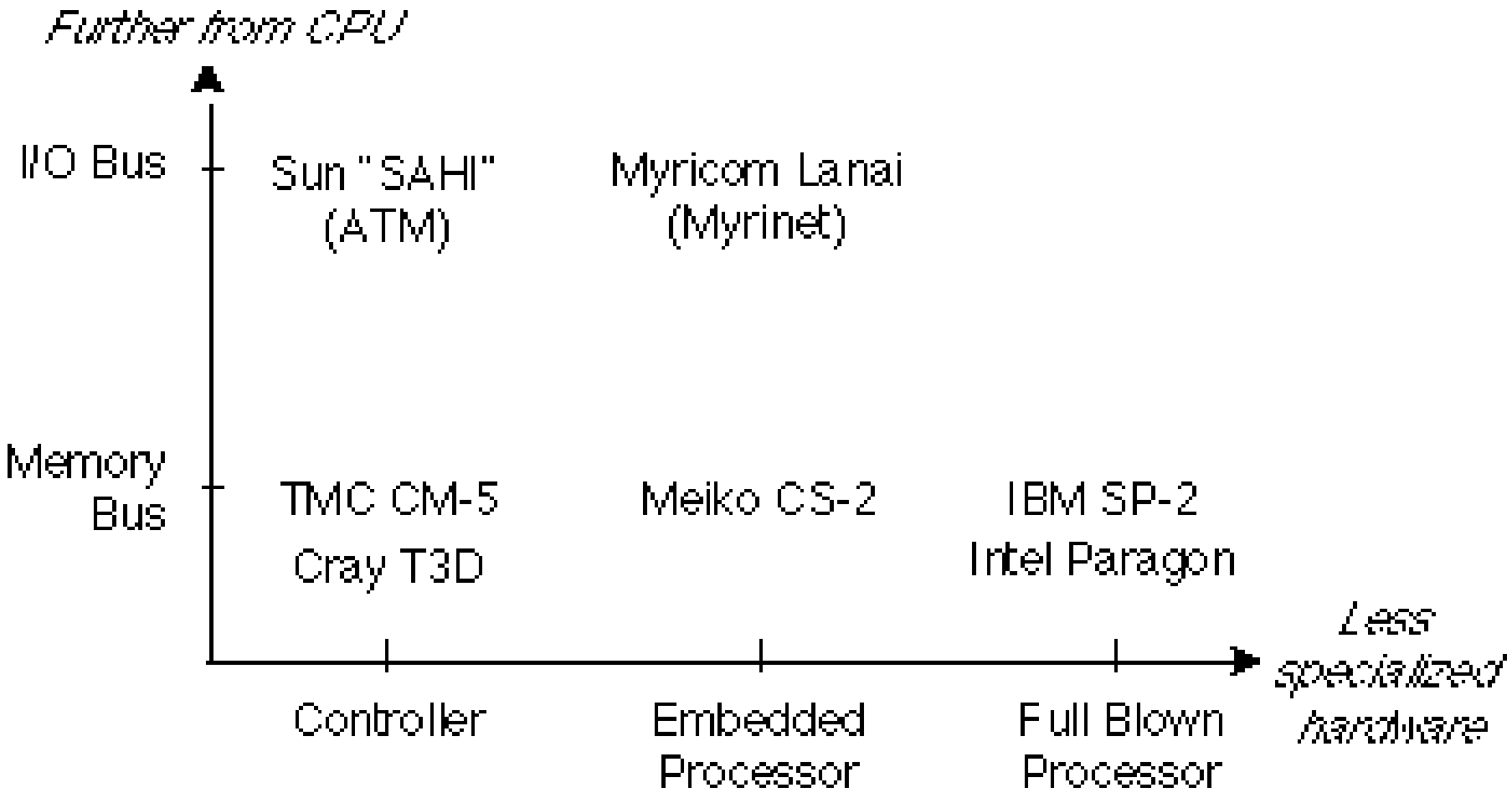
Example Networks

	MPP	LAN	WAN
	IBM SP-2	100 Mb Ethernet	ATM
Length (meters)	10	200	100/1000
Number data lines	8	1	1
Clock Rate	40 MHz	100 MHz	155/622…
Switch?	Yes	No	Yes
Nodes (N)	≤512	≤254	≈10000
Material	copper	copper	copper/fiber
Bisection BW (Mbit/s)	320xNodes	100	155xNodes
Peak Link BW (Mbits/s)	320	100	155
Measured Link BW	284	--	80

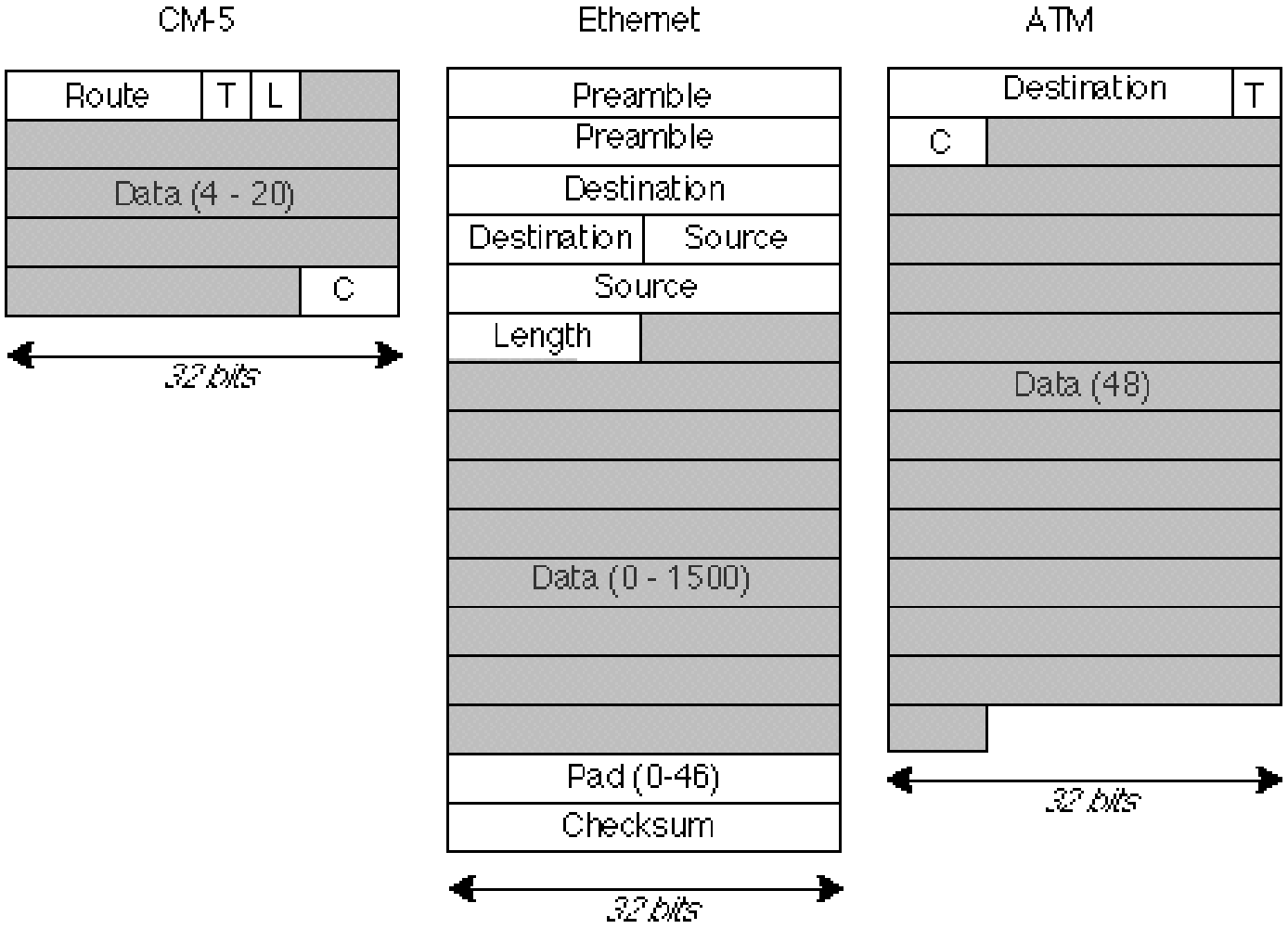
Example Networks (cont'd)

	MPP	LAN	WAN
	IBM SP-2	100 Mb Ethernet	ATM
Latency (µsecs)	1	1.5	≈50
Send+Receive Ovhd (µsecs)	39	440	630
Topology	Fat tree	Line	Star
Connectionless?	Yes	Yes	No
Store & Forward?	No	No	Yes
Congestion Control	Back-pressure	Carrier Sense	Choke packets
Standard	No	Yes	Yes
Fault Tolerance	Yes	Yes	Yes

Examples: Interface to Processor



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**

Example Switched LAN Performance

Network Interface	*Switch*	*Link BW*
AMD Lance Ethernet	Baynetworks EtherCell 28115	10 Mb/s
Fore SBA-200 ATM	Fore ASX-200	155 Mb/s
Myricom Myrinet	Myricom Myrinet	640 Mb/s

- On SPARCstation-20 running Solaris 2.4 OS
- Myrinet is example of "System Area Network": networks for a single room or floor: 25m limit
  - shorter => wider faster, less need for optical
  - short distance => source-based routing => simpler switches
  - Compaq-Tandem/Microsoft also sponsoring SAN, called "ServerNet"

Example Switched LAN Performance (1995)

Switch	*Switch Latency*
Baynetworks EtherCell 28115	52.0 µsecs
Fore ASX-200 ATM	13.0 µsecs
Myricom Myrinet	0.5 µsecs

- Measurements taken from “LogP Quantyified: The Case for Low-Overhead Local Area Networks”, K. Keeton, T. Anderson, D. Patterson, Hot Interconnects III, Stanford California, August 1995.

UDP/IP performance

Network	*UDP/IP roundtrip, N=8B*	*Formula*
Bay. EtherCell	1009 µsecs	+2.18*N
Fore ASX-200 ATM	1285 µsecs	+0.32*N
Myricom Myrinet	1443 µsecs	+0.36*N

- Formula from simple linear regression for tests from N = 8B to N = 8192B
- Software overhead not tuned for Fore, Myrinet; EtherCell using standard driver for Ethernet

NFS performance

Network	*Avg. NFS response*	*LinkBW/Ether*	*UDP/E.*
Bay. EtherCell	**14.5 ms**	**1**	**1.00**
Fore ASX-200 ATM	**11.8 ms**	**15**	**1.36**
Myricom Myrinet	**13.3 ms**	**64**	**1.43**

- **Last 2 columns show ratios of link bandwidth and UDP roundtrip times for 8B message to Ethernet**

Estimated Database performance (1995)

Network	*Avg. TPS*	*LinkBW/E.*	*TCP/E.*
Bay. EtherCell	77 tps	1	1.00
Fore ASX-200 ATM	67 tps	15	1.47
Myricom Myrinet	66 tps	64	1.46

- Number of Transactions per Second (TPS) for DebitCredit Benchmark; front end to server with entire database in main memory (256 MB)
  - Each transaction => 4 messages via TCP/IP
  - DebitCredit Message sizes < 200 bytes
- Last 2 columns show ratios of link bandwidth and TCP/IP roundtrip times for 8B message to Ethernet

Summary: Networking

- **Protocols allow hetereogeneous networking**
  - **Protocols allow operation in the presense of failures**
  - **Internetworking protocols used as LAN protocols => large overhead for LAN**
- **Integrated circuit revolutionizing networks as well as processors**
  - **Switch is a specialized computer**
  - **Faster networks and slow overheads violate of Amdahl's Law**

Parallel Computers

- **Definition: “A parallel computer is a collection of processiong elements that cooperate and communicate to solve large problems fast.”**

  Almasi and Gottlieb, *Highly Parallel Computing* ,1989

- **Questions about parallel computers:**
  - **How large a collection?**
  - **How powerful are processing elements?**
  - **How do they cooperate and communicate?**
  - **How are data transmitted?**
  - **What type of interconnection?**
  - **What are HW and SW primitives for programmer?**
  - **Does it translate into performance?**

Parallel Processors “Religion”

- **The dream of computer architects since 1960: replicate processors to add performance vs. design a faster processor**
- **Led to innovative organization tied to particular programming models since “uniprocessors can’t keep going”**
  - **e.g., uniprocessors must stop getting faster due to limit of speed of light: 1972, … , 1989**
  - **Borders religious fervor: you must believe!**
  - **Fervor damped some when 1990s companies went out of business: Thinking Machines, Kendall Square, ...**
- **Argument instead is the “pull” of opportunity of scalable performance, not the “push” of uniprocessor performance plateau**

Opportunities: Scientific Computing

- **Nearly Unlimited Demand (Grand Challenge):**

App	*Perf (GFLOPS)*	*Memory (GB)*
48 hour weather	0.1	0.1
72 hour weather	3	1
Pharmaceutical design	100	10
Global Change, Genome	1000	1000

(Figure 1-2, page 25, of Culler, Sighn, Gupta [CSG97])

- **Successes in some real industries:**
  - **Petrolium: reservoir modeling**
  - **Automotive: crash simulation, drag analysis, engine**
  - **Aeronautics: airflow analysis, engine, structural mechanics**
  - **Pharmaceuticals: molecular modeling**
  - **Entertainment: full length movies ("Toy Story")**

Example: Scientific Computing

- **Molecular Dynamics on Intel Paragon with 128 processors (1994)**
  - **(see Chapter 1, Figure 1-3, page 27 of Culler, Sighn, Gupta [CSG97])**
  - **Classic MPP slide: processors v. speedup**
- **Improve over time: load balancing, other**
- **128 processor Intel Paragon = 406 MFLOPS**
- **C90 vector = 145 MFLOPS (or ≈ 45 Intel processors)**

Opportunities: Commercial Computing

- **Transaction processing & TPC-C bencmark**
  - **(see Chapter 1, Figure 1-4, page 28 of [CSG97])**
  - **small scale parallel processors to large scale**
- **Througput (Transactions per minute) vs. Time (1996)**
- **Speedup:**

	1	4	8	16	32	64	112
IBM RS6000	**735**	**1438**	**3119**				
	1.00	***1.96***	***4.24***				
Tandem Himilaya				**3043**	**6067**	**12021**	**20918**
				1.00	***1.99***	***3.95***	***6.87***

  - **IBM performance hit 1=>4, good 4=>8**
  - **Tandem scales: 112/16 = 7.0**
- **Others: File servers, eletronic CAD simulation (multiple processes), WWW search engines**

What level Parallelism?

- **Bit level parallelism: 1970 to ≈1985**
  - **4 bits, 8 bit, 16 bit, 32 bit microprocessors**
- **Instruction level parallelism (ILP): 1985 through today**
  - **Pipelining**
  - **Superscalar**
  - **VLIW**
  - **Out-of-Order execution**
  - **Limits to benefits of ILP?**
- **Process Level or Thread level parallelism; mainstream for general purpose computing?**
  - **Servers are parallel (see Fig. 1-8, p. 37 of [CSG97])**
  - **Highend Desktop dual processor PC soon?? (or just the sell the socket?)**

Whither Supercomputing?

- **Linpack (dense linear algebra) for Vector Supercomputers vs. Microprocessors**
- **"Attack of the Killer Micros"**
  - **(see Chapter 1, Figure 1-10, page 39 of [CSG97])**
  - **100 x 100 vs. 1000 x 1000**
- **MPPs vs. Supercomputers when rewrite linpack to get peak performance**
  - **(see Chapter 1, Figure 1-11, page 40 of [CSG97])**
- **500 fastest machines in the world: parallel vector processors (PVP), bus-based shared memory (SMP), and MPPs**
  - **(see Chapter 1, Figure 1-12, page 41 of [CSG97])**

Parallel Architecture

- **Parallel Architecture extends traditional computer architecture with a communication architecture**
  - **abstractions (HW/SW interface)**
  - **organizational structure to realize abstraction efficiently**

Parallel Framework

- **Layers:**
  - **(see Chapter 1, Figure 1-13, page 42 of [CSG97])**
  - **Programming Model:**
    - » **Multiprogramming : lots of jobs, no communication**
    - » **Shared address space: communicate via memory**
    - » **Message passing: send and recieve messages**
    - » **Data Parallel: several agents operate on several data sets simultaneously and then exchange information globally and simultaneously (shared or message passing)**
  - **Communication Abstraction:**
    - » **Shared address space: e.g., load, store, atomic swap**
    - » **Message passing: e.g., send, recieve library calls**
    - » **Debate over this topic (ease of programming, scaling) => many hardware designs 1:1 programming model**

Shared Address Model Summary

- Each processor can name every physical location in the machine
- Each process can name all data it shares with other processes
- Data transfer via load and store
- Data size: byte, word, ... or cache blocks
- Uses virtual memory to map virtual to local or remote physical
- Memory hierarchy model applies: now communication moves data to local processor cache (as load moves data from memory to cache)
  - Latency, BW, scalability when communicate?

Networking Summary

- Protocols allow hetereogeneous networking
- Protocols allow operation in the presense of failures
- Routing issues: store and forward vs. cut through, congestion, ...
- Standardization key for LAN, WAN
- Internetworking protocols used as LAN protocols => large overhead for LAN
- Integrated circuit revolutionizing networks as well as processors
- Switch is a specialized computer
- High bandwidth networks with high overheads violate of Amdahl's Law