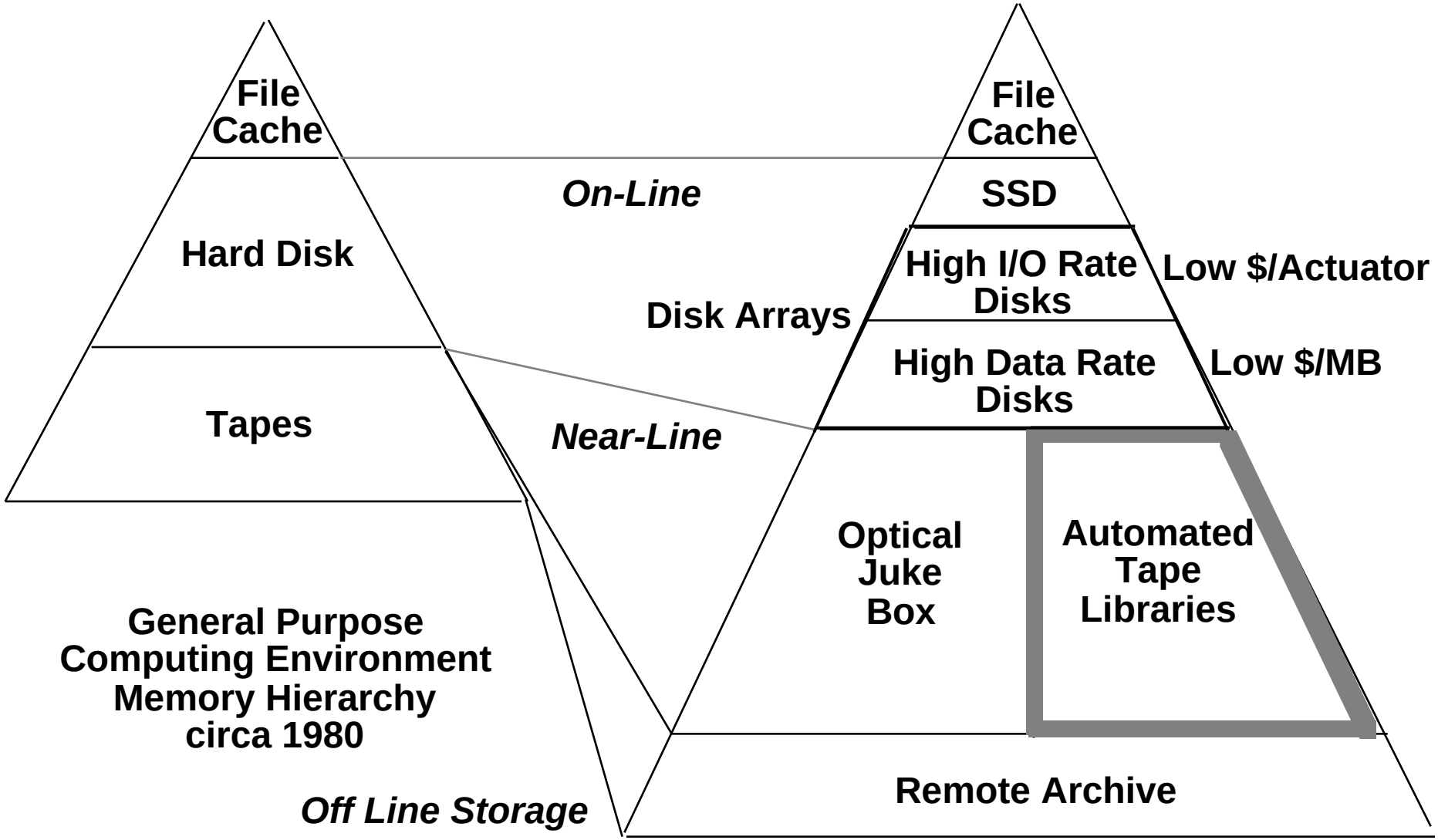
Lecture 15: Networks & Interconnect—Interface, Switches, Routing, Examples

Professor David A. Patterson

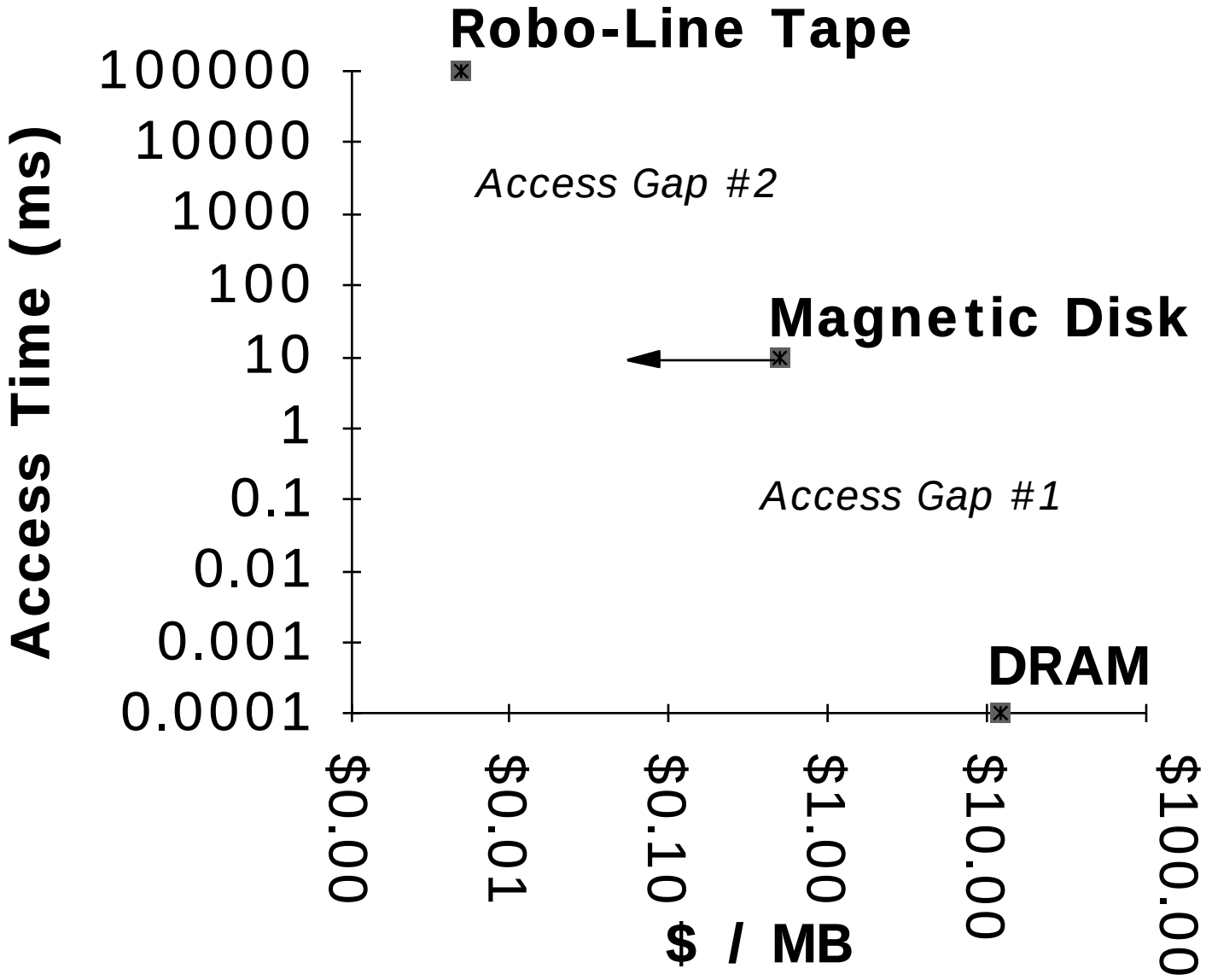
Computer Science 252

Fall 1996

Review: Memory Hierarchies



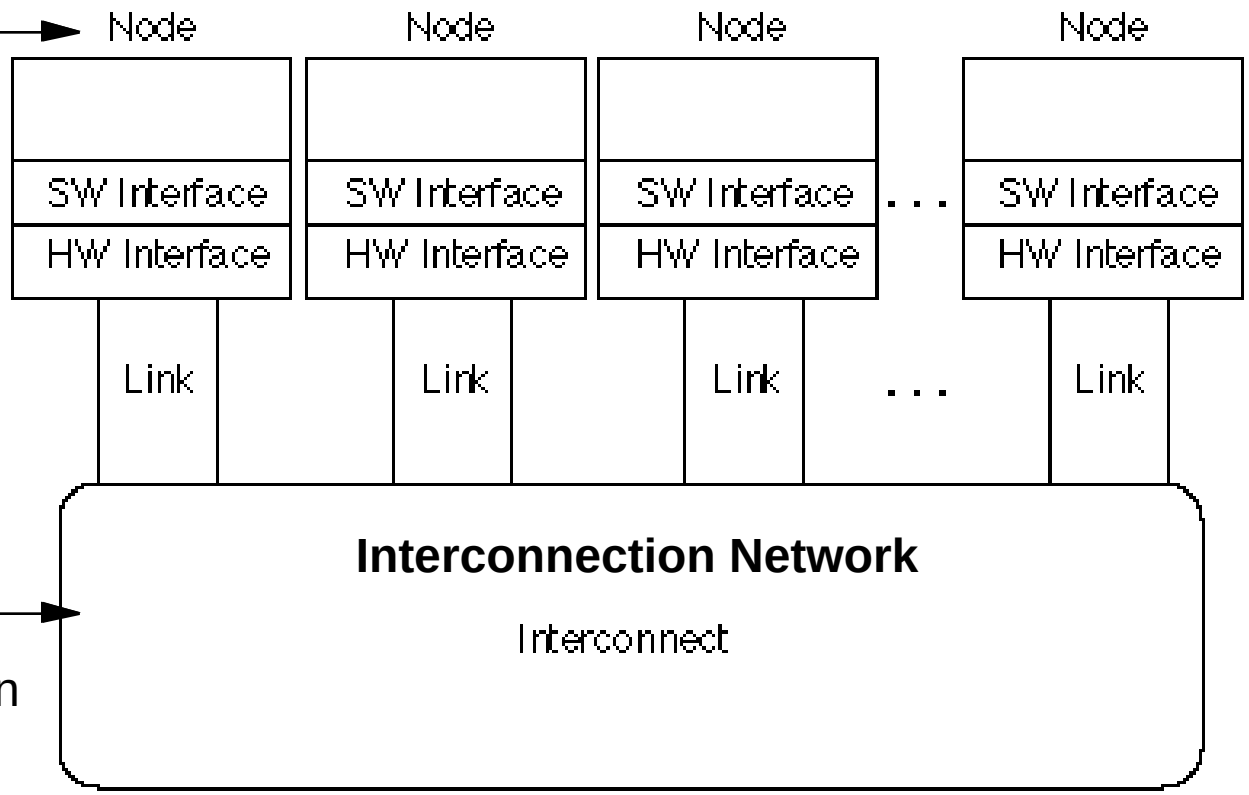
MSS: Review


Robo-Line Tape
Access Gap #2
Magnetic Disk
Access Gap #1
DRAM
Access Time (ms)
100000
10000
1000
100
10
1
0.1
0.01
0.001
0.0001
$0.00
$0.01
$0.10
$1.00
$10.00
$100.00
$ / MB

Review: Interconnections (Networks)

- **Examples:**
  - **MPP networks (SP2): 1000s nodes; ≤ 25 meters per link**
  - **Local Area Networks (Ethernet): 100s nodes; ≤ 1000 meters**
  - **Wide Area Network (ATM): 1000s nodes; ≤ 5,000,000 meters**

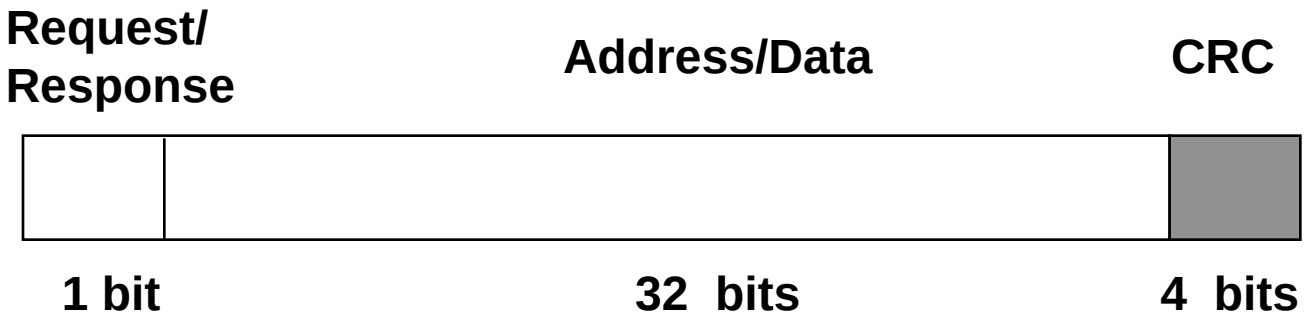
a.k.a.
end systems,
hosts



a.k.a.
network,
communication
subnet

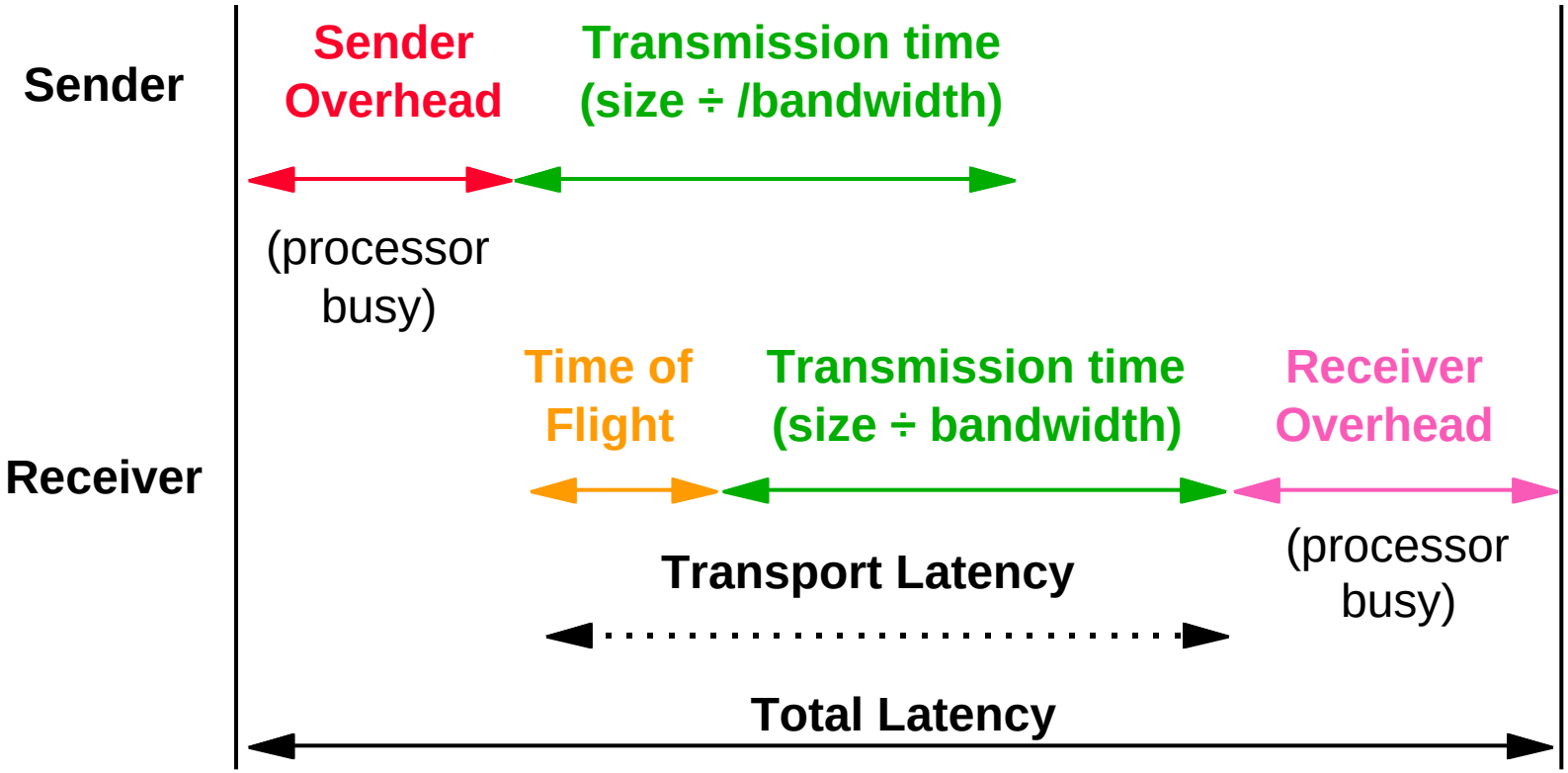
Review: A Simple Example Revisted

- **What is the format of packet?**
  - **Fixed? Number bytes?**



00: Request—Please send data from Address
01: Reply—Packet contains data corresponding to request
10: Acknowledge request
11: Acknowledge reply

Review: Performance Metrics



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

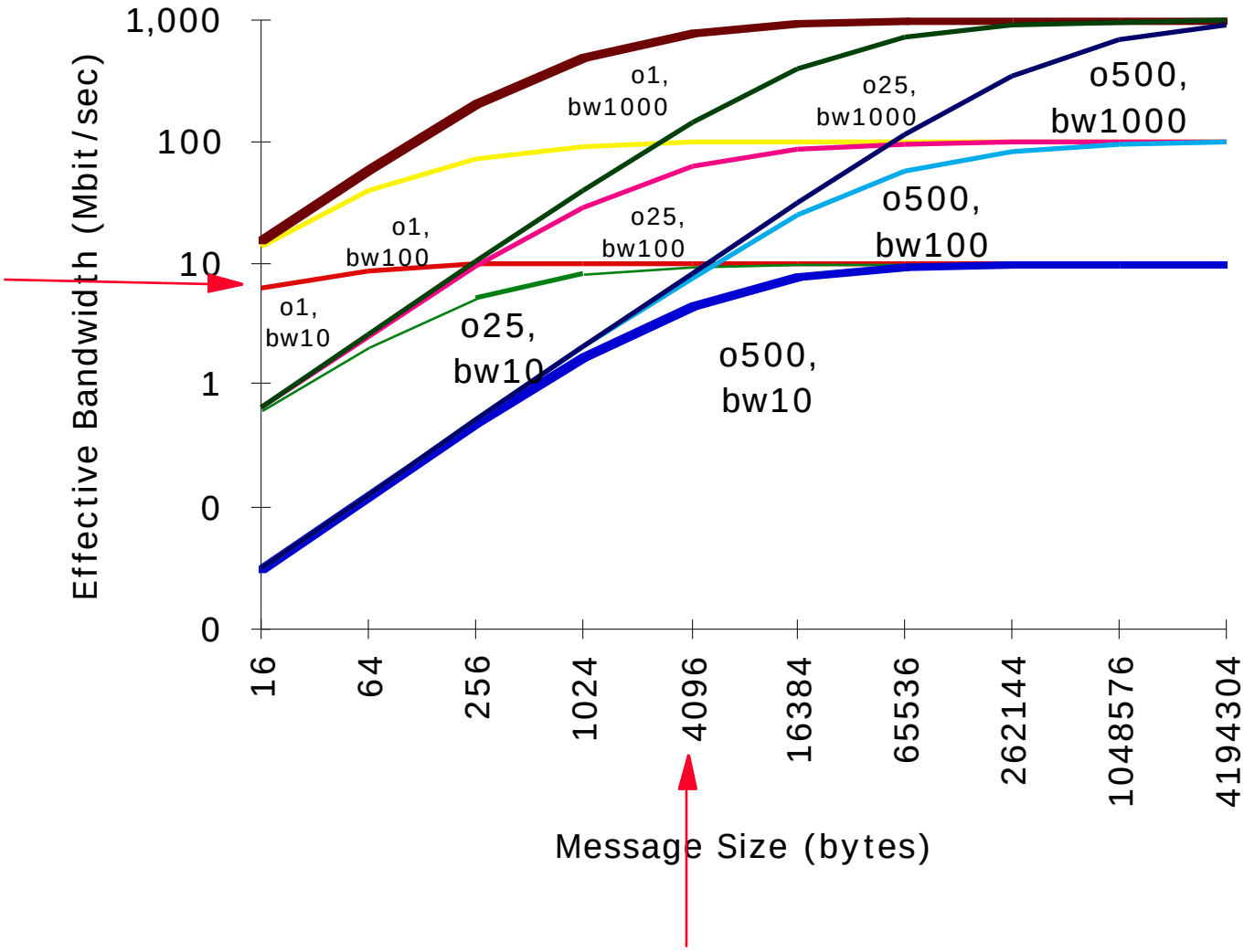
Summary: Interconnections

- **Communication between computers**
- **Packets for standards, protocols to cover normal and abnormal events**
- **Performance issues: HW & SW overhead, interconnect latency, bisection BW**

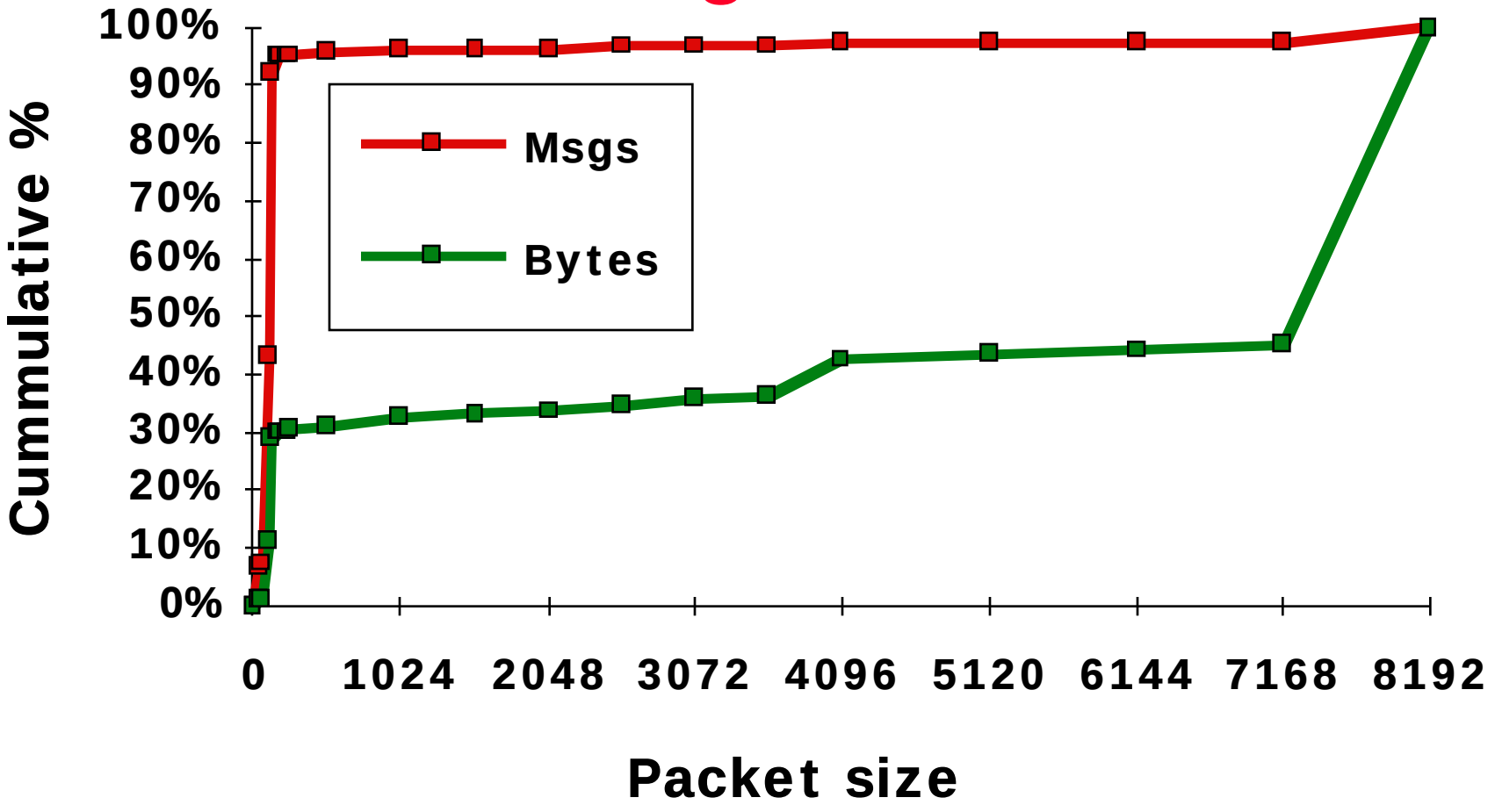
Simplified Latency Model

- **Total Latency ≈ Overhead + Message Size ÷ BW**
- **Overhead = Sender Overhead + Time of Flight + Receiver Overhead**
- **Example: show what happens as vary**
  - **Overhead: 1, 25, 500 µsec**
  - **BW:10,100, 1000 Mbit/sec**
  - **Message Size: 16 Bytes to 4 MB**
- **If overhead 500 µsec, how big a message > 10 Mb/s?**
- **How big are messages?**

Overhead, BW, Size

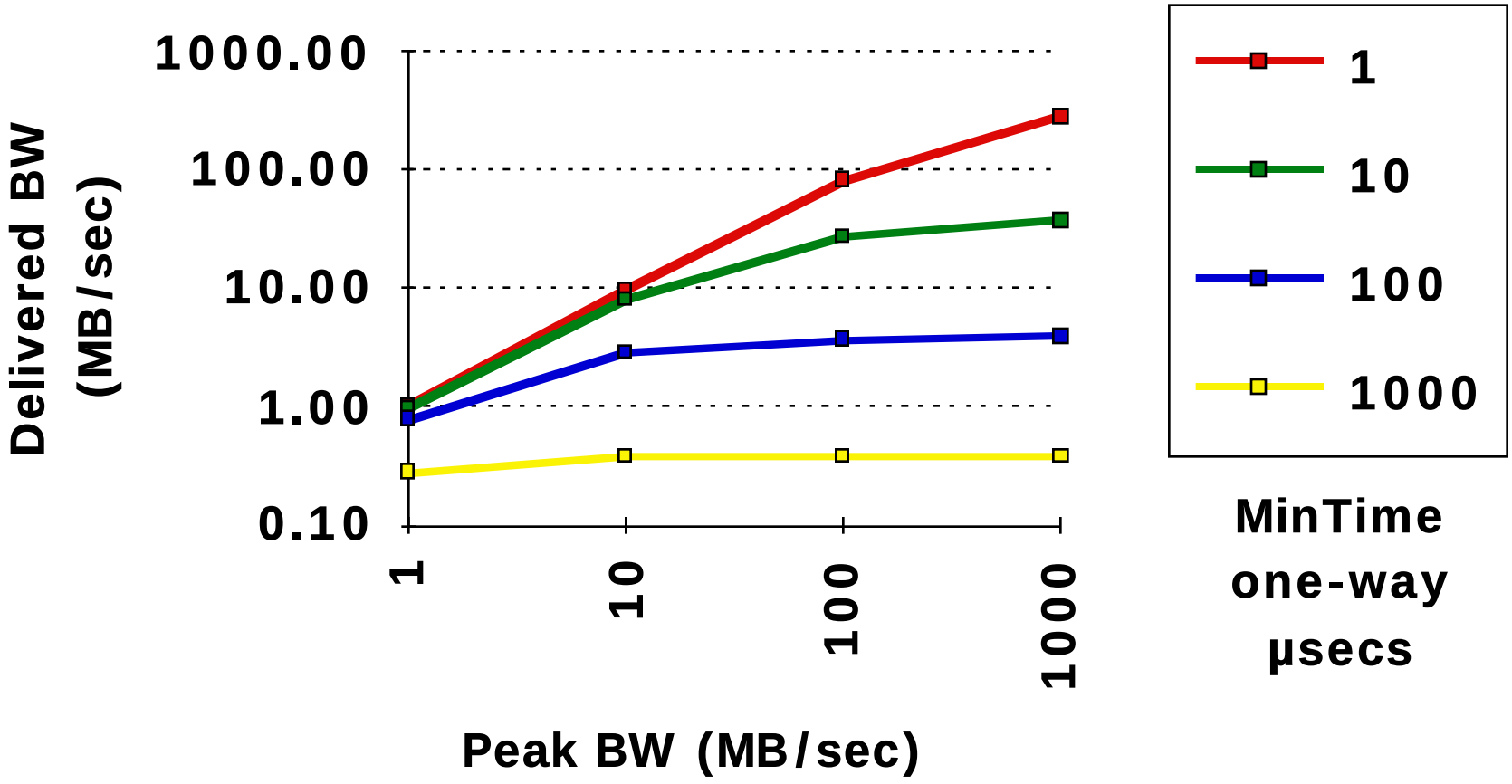


Measurement: Sizes of Message for NFS



- 95% Msgs, 30% bytes for packets ≤ 200 bytes
- > 50% data transfered in packets = 8KB

Impact of Overhead on Delivered BW



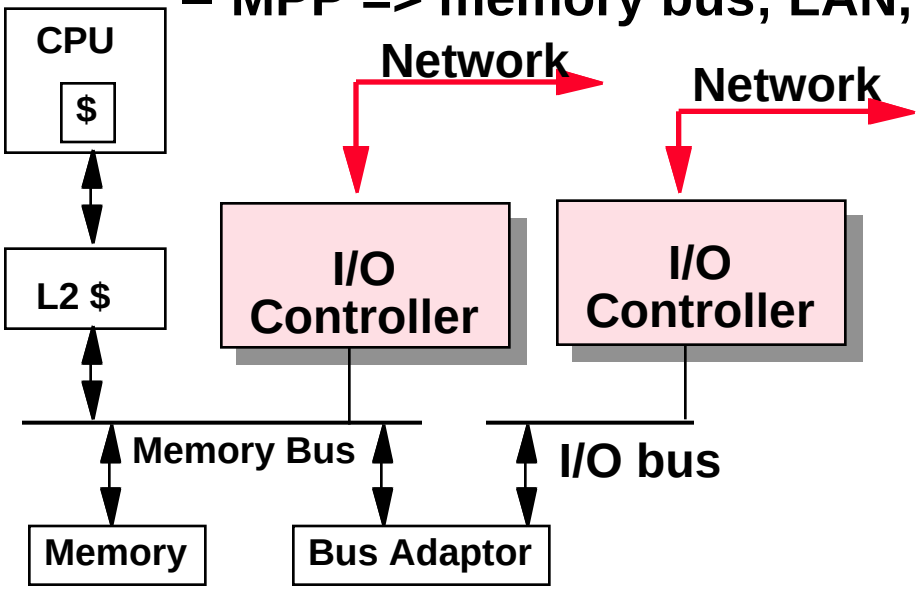
- BW model: Time = overhead + msg size/peak BW
- > 50% data transfered in packets = 8KB

Interconnect Issues

- **Performance Measures**
- **Interface Issues**

Interface Issues

- **Where to connect network to computer?**
  - **Cache consistent to avoid flushes? (memory bus)**
  - **Standard interface card? (I/O bus)**
  - **Latency and bandwidth? (memory bus)**
  - **MPP => memory bus; LAN, WAN => I/O bus**



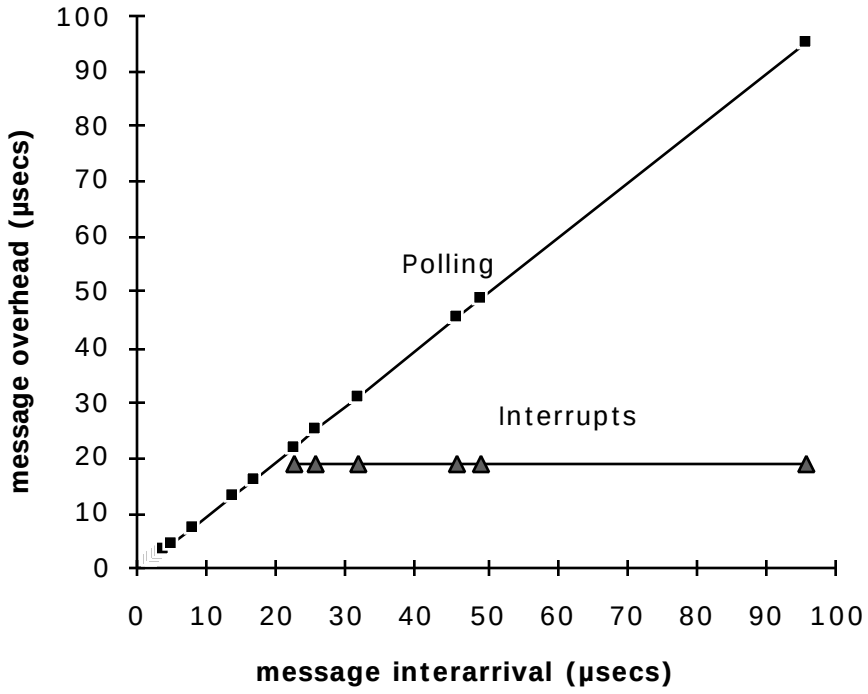
ideal: high bandwidth, low latency, standard interface

Interface Issues

- **How to connect network to software?**
  - **Programmed I/O?(low latency)**
  - **DMA? (best for large messages)**
  - **Receiver interrupted or received polls?**
- **Things to avoid**
  - **Invoking operating system in common case**
  - **Operating at uncache memory speed (e.g., check status of network interface)**

CM-5 Software Interface

- **CM-5 example (MPP)**
  - **Time per poll 1.6 µsecs; time per interrupt 19 µsecs**
  - **Minimum time to handle message: 0.5 µsecs**
  - **Enable/disable 4.9/3.8 µsecs**
- **As rate of messages arrving changes, use polling or interrupt?**
  - **Solution: Always enable interrupts, have interrupt routine poll until until no messages pending**
  - **Low rate => ≈ interrupt**
  - **High rate => ≈ polling**

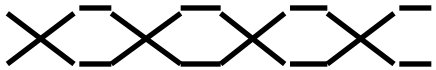


Interconnect Issues

- **Performance Measures**
- **Interface Issues**
- **Network Media**

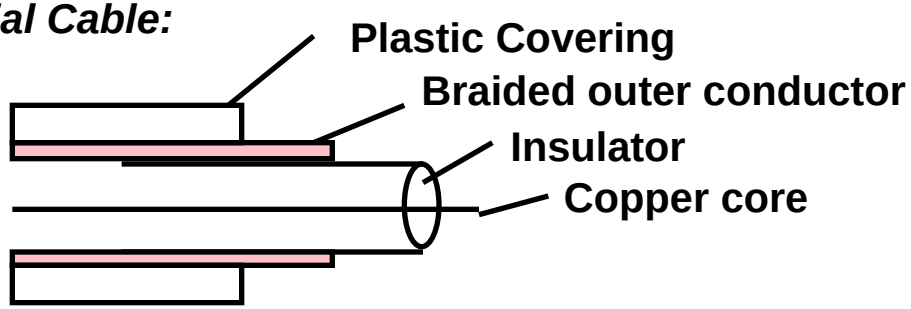
Network Media

Twisted Pair:



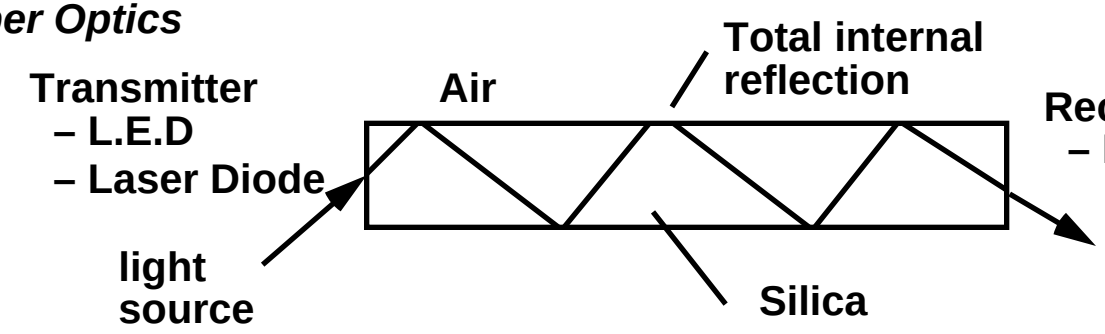
Copper, 1mm think, twisted to avoid attenna effect (telephone)

Coaxial Cable:



Used by cable companies: high BW, good noise immunity

Fiber Optics



Light: 3 parts are cable, light source, light detector. Multimode light disperse (LED), Single mode sinle wave (laser)

Network Media

Media	Bandwidth	Distance	Cost/meter	Cost/interface
twisted pair copper wire	1 Mb/s (20 Mb/s)	2 km (0.1 km)	$0.23	$2
coaxial cable	10 Mb/s	1 km	$1.64	$5
multimode optical fiber	600 Mb/s	2 km	$1.03	$1000
single mode optical fiber	2000 Mb/s	100 km	$1.64	$1000

CS 252 Administrivia

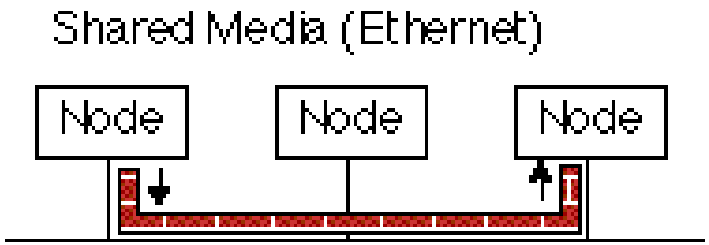
- **Project Survey #2 due today**
- **Homework on Chapter 7 due Monday 11/4 at 5 PM in 252 box, done in pairs:**
  - **Exercises 7.1, 7.3, 7.10**

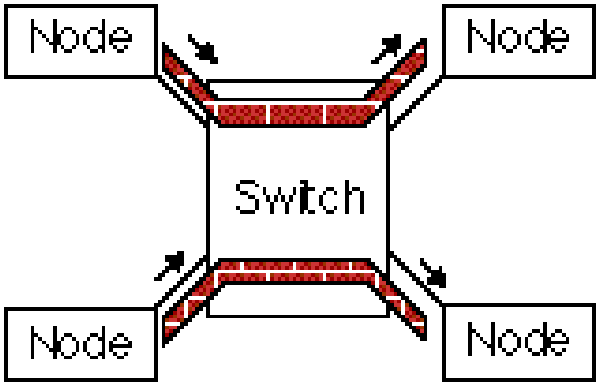
Interconnect Issues

- Performance Measures
- Interface Issues
- Network Media
- Connecting Multiple Computers

Connecting Multiple Computers

- **Shared Media vs. Switched: pairs communicate at same time: "point-to-point" connections**
- **Aggregate BW in switched network is many times shared**
  - **point-to-point faster since no arbitration and simpler interface**
- **Arbitration in Shared network?**
  - **Central arbiter for LAN?**
  - **Listen to check if being used ("Carrier Sensing")**
  - **Listen to check if collision ("Collision Detectino")**
  - **Random resend to avoid repeated collisions**





(A. K. A. data switching interchanges, multistage interconnection networks, interface message processors)

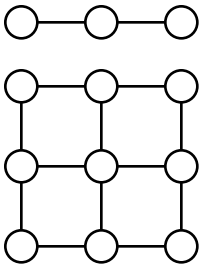
Example Interconnects

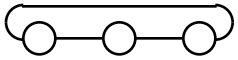
Interconnect	*MPP*	*LAN*	*WAN*
Example	CM-5	Ethernet	ATM
Maximum length between nodes	25 m	500 m; ≤5 repeaters	copper: 100 m optical: 2 km—25 km
Number data lines	4	1	1
Clock Rate	40 MHz	10 MHz	≥ 155.5 MHz
Shared vs. Switch	Switch	Shared	Switch
Maximum number of nodes	2048	254	> 10,000
Media Material	Copper	Twisted pair copper wire or Coaxial cable	Twisted pair copper wire or optical fiber

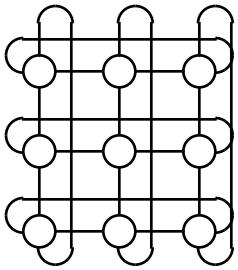
Switch Topology

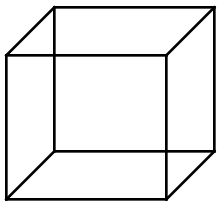
- Structure of the interconnect
- Determines
  - Degree: number of links from a node
  - Diameter: max number of links crossed between nodes
  - Average distance: number of hops to random destination
  - Bisection: minimum number of links that separate the network into two halves
- Warning: these three-dimensional 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





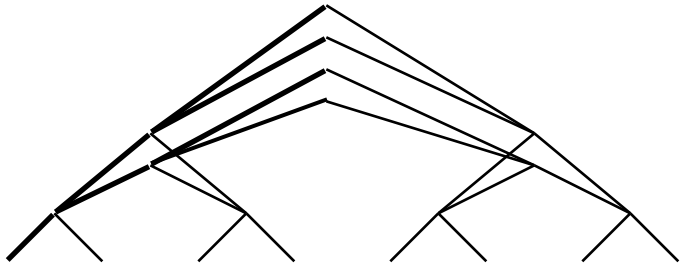


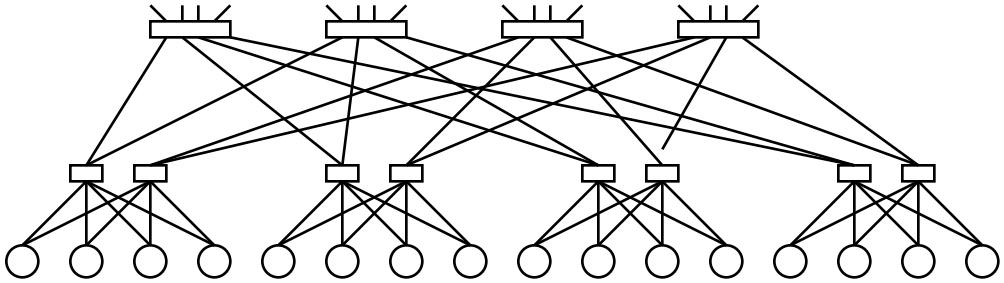


Type	Degree	Diameter	Ave Dist	Bisection	N = 1024 Diam	Ave D
1D mesh	≤ 2	N-1	N/3	1		
2D mesh	≤ 4	$2(N^{1/2} - 1)$	$2N^{1/2} / 3$	$N^{1/2}$	63	21
3D mesh	≤ 6	$3(N^{1/3} - 1)$	$3N^{1/3} / 3$	$N^{2/3}$	~30	~10
nD mesh ($N = k^n$)	≤ 2n	$n(N^{1/n} - 1)$	$nN^{1/n} / 3$	$N^{(n-1) / n}$		
Ring	2	N / 2	N/4	2		
2D torus	4	$N^{1/2}$	$N^{1/2} / 2$	$2N^{1/2}$	32	16
k-ary n-cube ($N = k^n$)	2n	$n(N^{1/n})$ nk/2	$nN^{1/n}/2$ nk/4	$2k^{n-1}$	15	8 (3D)
Hypercube	n	n = LogN	n/2	N/2	10	5
Cube-Connected Cycles						

Topologies (cont)

					N = 1024	
Type	Degree	Diameter	Ave Dist	Bisection	Diam	Ave D
2D Tree	3	$2\text{Log}_2 N$	$\sim 2\text{Log}_2 N$	1	20	~20
4D Tree	5	$2\text{Log}_4 N$	$2\text{Log}_4 N - 2/3$	1	10	9.33
kD	k+1	$\text{Log}_k N$				
2D fat tree	4	$\text{Log}_2 N$		N		
2D butterfly	4	$\text{Log}_2 N$		N/2	20	20

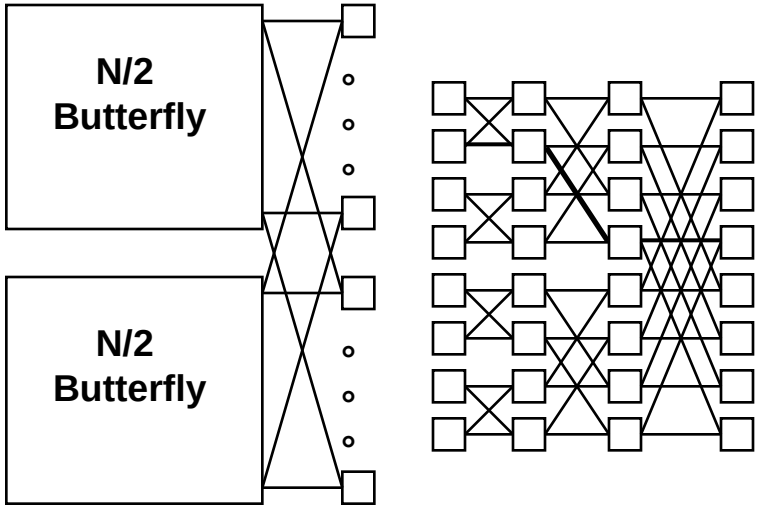




CM-5 Thinned Fat Tree

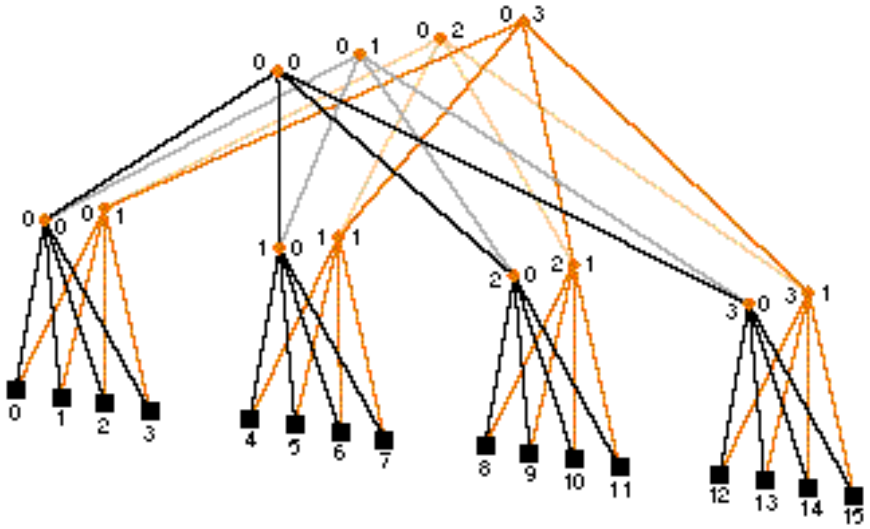
Butterfly

Multistage: nodes at ends, switches in middle



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

Multistage Fat Tree



- **Randomly assign packets to different paths on way up to spread the load**

Example Networks

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

MBytes/second

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

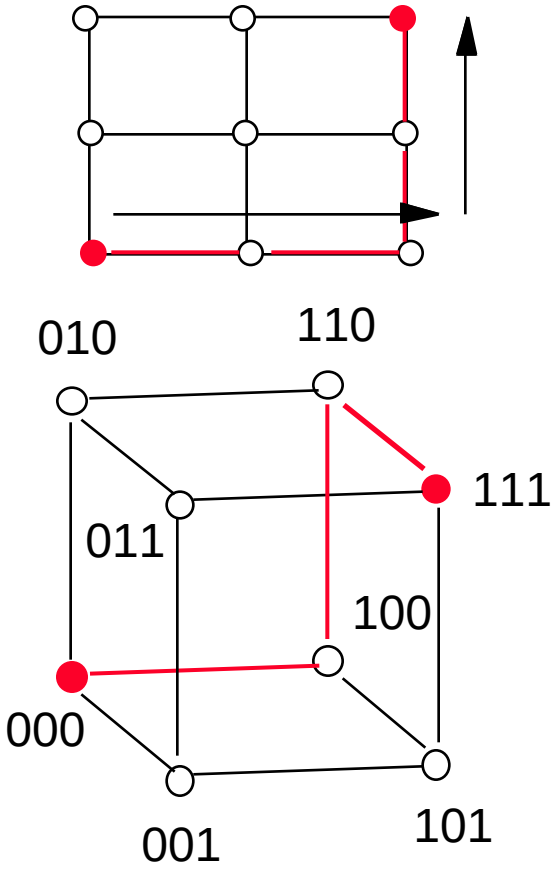
- **Connectionless: every package of information must have an address => packets**
  - **Each package is routed to its destination by looking at its address, e.g., the postal system**
  - **also called “Statistical multiplexing”**
  - **“Split phase buses” are sending packets**

Routing Messages

- **Shared Media**
  - **Broadcast to everyone**
- **Switched Media needs real routing. Options:**
  - **Source-based routing: message specifies path to the destination**
  - **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 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

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., metering lights)**
- **Options:**
  - **Packet discarding: If a packet arrives at a switch and there is no room in the buffer, the packet is discarded (e.g., UDP)**
  - **Flow control: between pairs of receivers and senders; use feedback to tell the sender when 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 interconnection with overhead to send & receive a packet (e.g., TCP)**
  - **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

- **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?**
- **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!

Example Networks

- **Ethernet: shared media 10 MB/s proposed in 1978, carrier sensing with expotential backoff on collision detect**
- **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**
  - **100 Mbit Ethernet (Fast Ethernet)**
  - **Switched Ethernet**
  - **ATM**

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 level.**
  - **Routers divide the interconnect into separate smaller subnets, which simplifies manageability and improves security.**

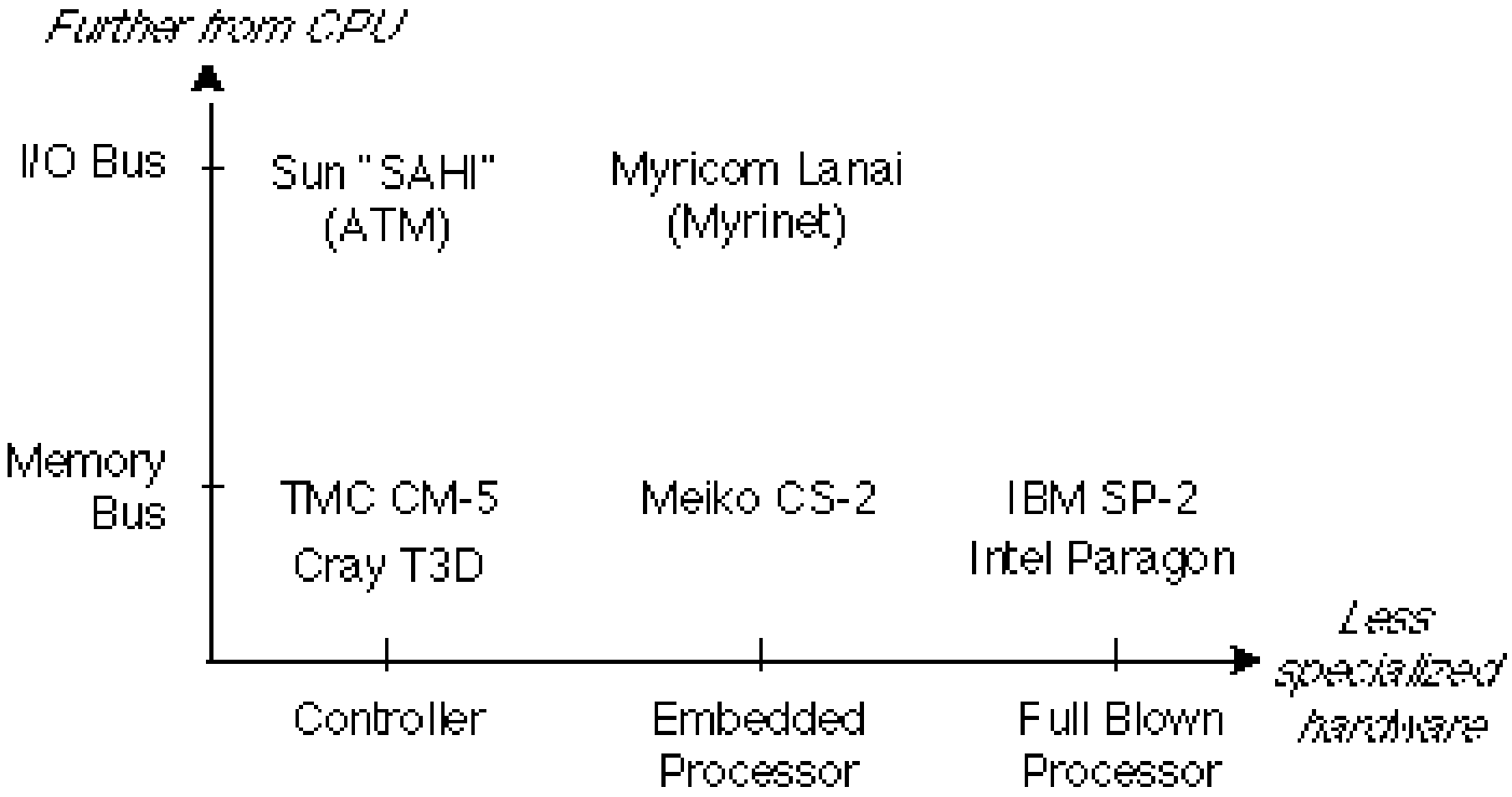
Example Networks

	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**

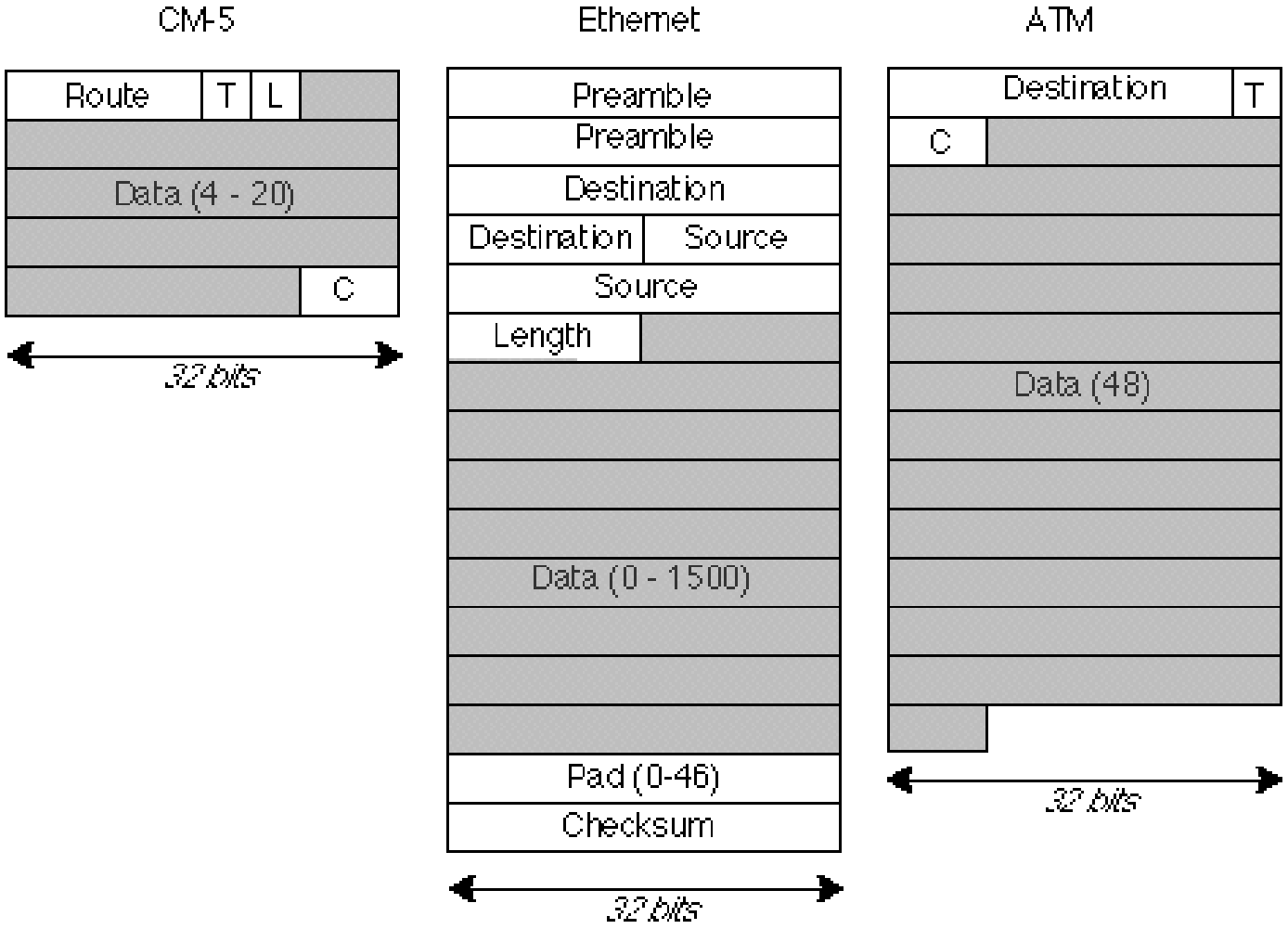
Example Networks (cont'd)

	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**

Summary: Interconnections

- **Media sets cost, distance**
- **Shared vs. Swicthed 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**
- **Routing issues: store and forward vs. cut through, congestion, ...**
- **Standardization key for LAN, WAN**