Lecture 15: Networks & Interconnect—Introduction

Professor David A. Patterson

Computer Science 252

Spring 1998

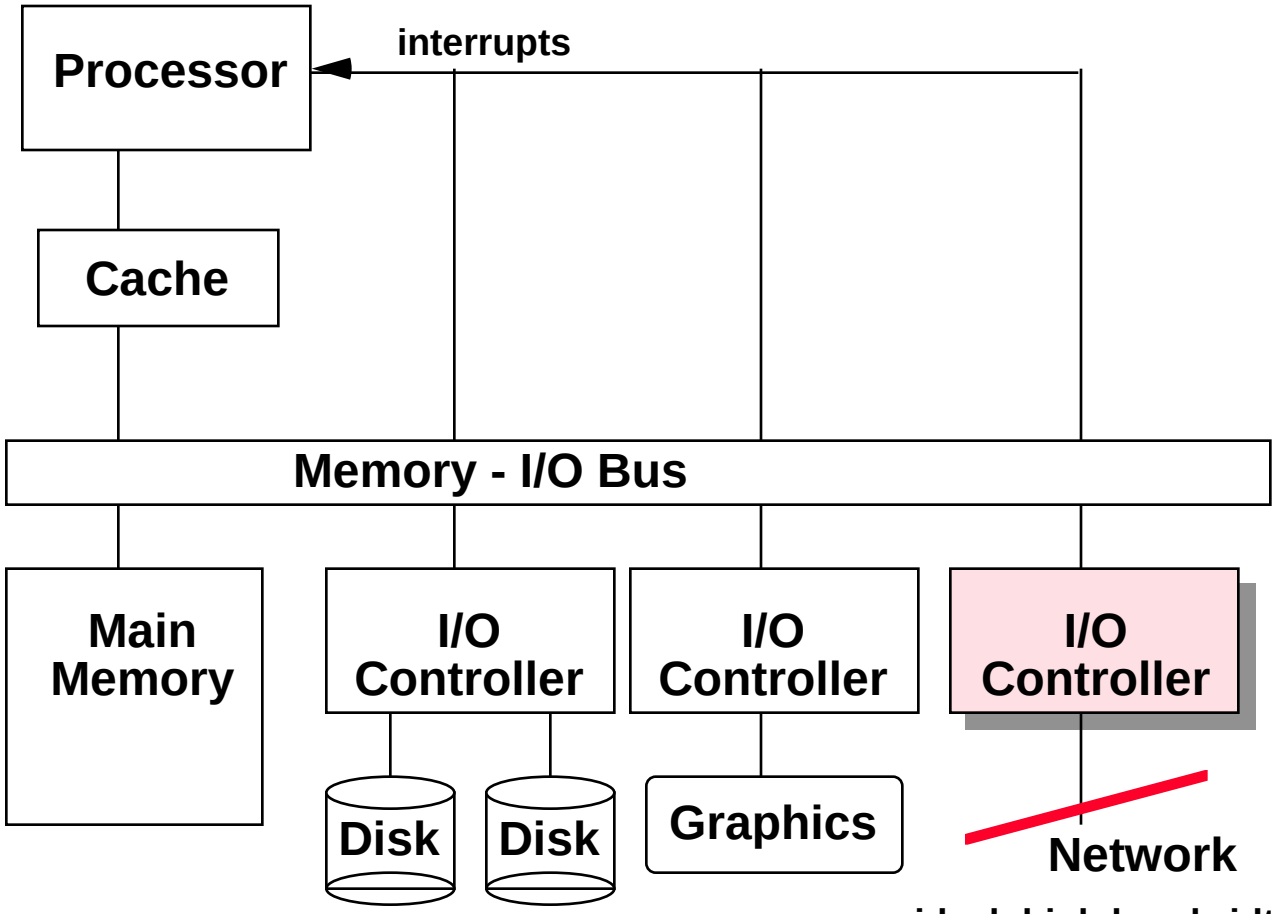
Review: Storage System Issues

- Historical Context of Storage I/O
- Secondary and Tertiary Storage Devices
- Storage I/O Performance Measures
- Processor Interface Issues
- A Little Queuing Theory
- Redundant Arrarys of Inexpensive Disks (RAID)
- ABCs of UNIX File Systems
- I/O Benchmarks

Review: I/O Benchmarks

- Scaling to track technological change
- TPC: price performance as nomalizing configuration feature
- Auditing to ensure no foul play
- Throughput with restricted response time is normal measure

I/O to External Devices and Other Computers

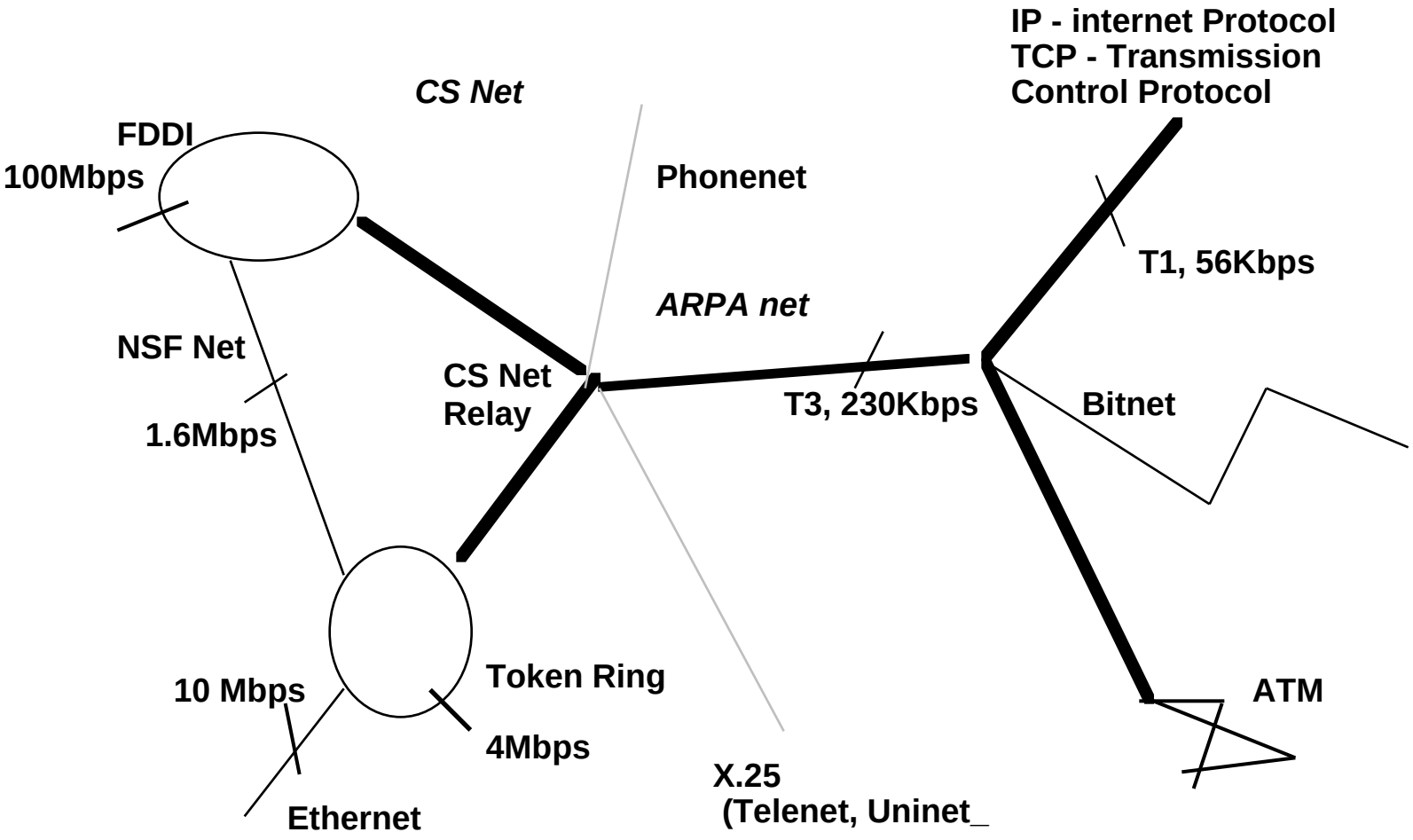


ideal: high bandwidth, low latency

Networks

- **Goal: Communication between computers**
- **Eventual Goal: treat collection of computers as if one big computer, distributed resource sharing**
- **Theme: Different computers must agree on many things**
  - **Overriding importance of standards and protocols**
  - **Fault tolerance critical as well**
- **Warning: Terminology-rich environment**

Example Major Networks



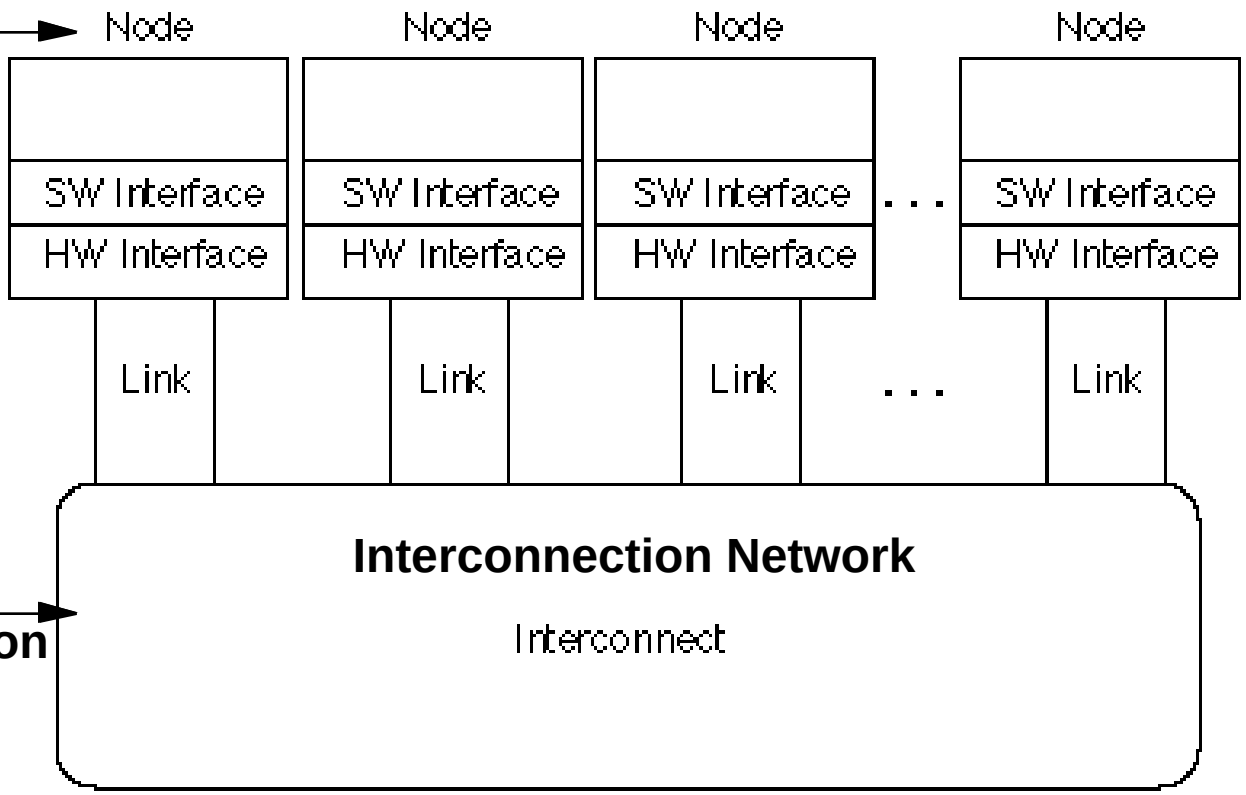
Networks

- **Facets people talk a lot about:**
  - **direct (point-to-point) vs. indirect (multi-hop)**
  - **topology (e.g., bus, ring, DAG)**
  - **routing algorithms**
  - **switching (aka multiplexing)**
  - **wiring (e.g., choice of media, copper, coax, fiber)**
- **What really matters:**
  - **latency**
  - **bandwidth**
  - **cost**
  - **reliability**

Interconnections (Networks)

- **Examples:**
  - **MPP networks (SP2): 100s 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



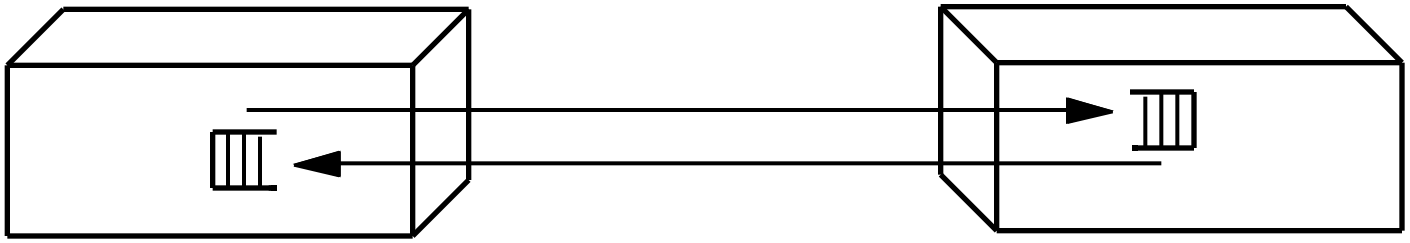
a.k.a. network, communication subnet

More Network Background

- Connection of 2 or more networks: Internetworking
- 3 cultures for 3 classes of networks
  - MPP: performance, latency and bandwidth
  - LAN: workstations, cost
  - WAN: telecommunications, phone call revenue
- Try for single terminology
- Motivate the interconnection complexity incrementally

ABCs of Networks

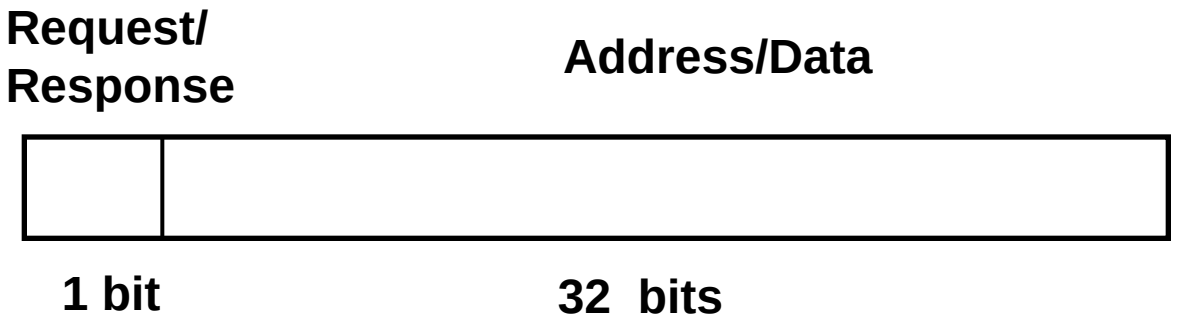
- **Starting Point: Send bits between 2 computers**



- **Queue (FIFO) on each end**
- **Information sent called a "message"**
- **Can send both ways ("Full Duplex")**
- **Rules for communication? "protocol"**
  - **Inside a computer:**
    - » **Loads/Stores: Request (Address) & Response (Data)**
    - » **Need Request & Response signaling**

A Simple Example

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



0: Please send data from Address
1: Packet contains data corresponding to request

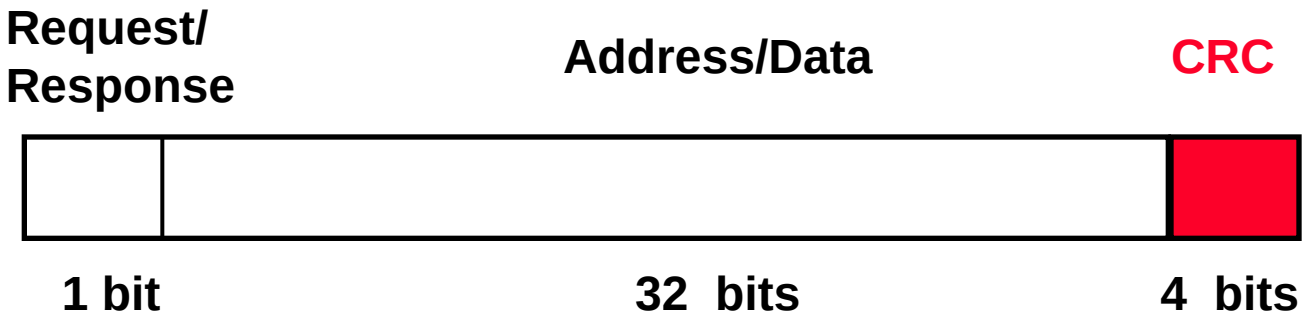
- **Header/Trailer: information to deliver a message**
- **Payload: data in message (1 word above)**

Questions About Simple Example

- What if more than 2 computers want to communicate?
  - Need computer "address field" (destination) in packet
- What if packet is garbled in transit?
  - Add "error detection field" in packet (e.g., CRC)
- What if packet is lost?
  - More "elaborate protocols" to detect loss (e.g., NAK, ARQ, time outs)
- What if multiple processes/machine?
  - Queue per process to provide protection
- Simple questions such as these lead to more complex protocols and packet formats => complexity

A Simple Example Revisted

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

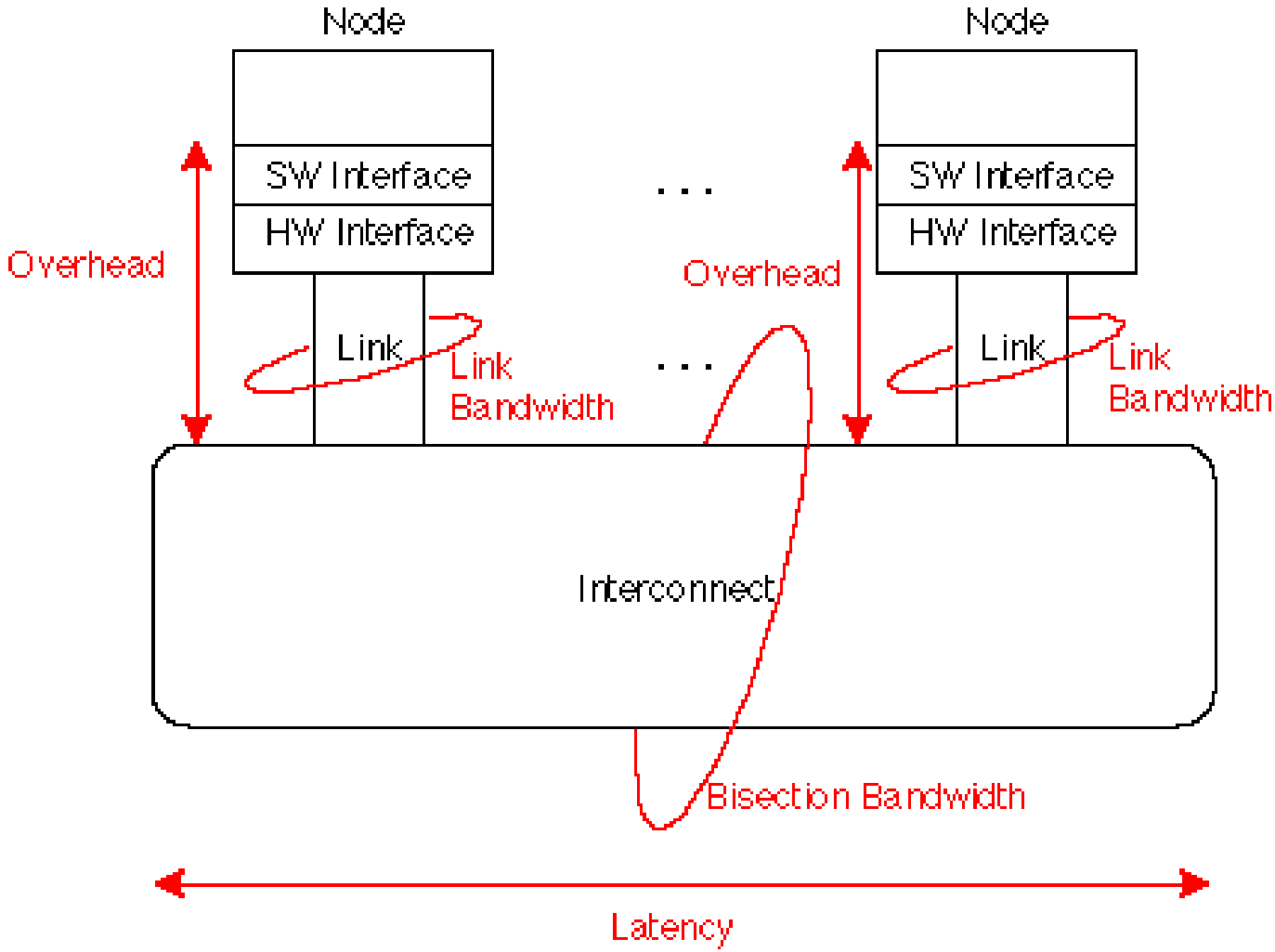


00: Request—Please send data from Address
01: Reply—Packet contains data corresponding to request
10: Acknowledge request
11: Acknowledge reply

Software to Send and Receive

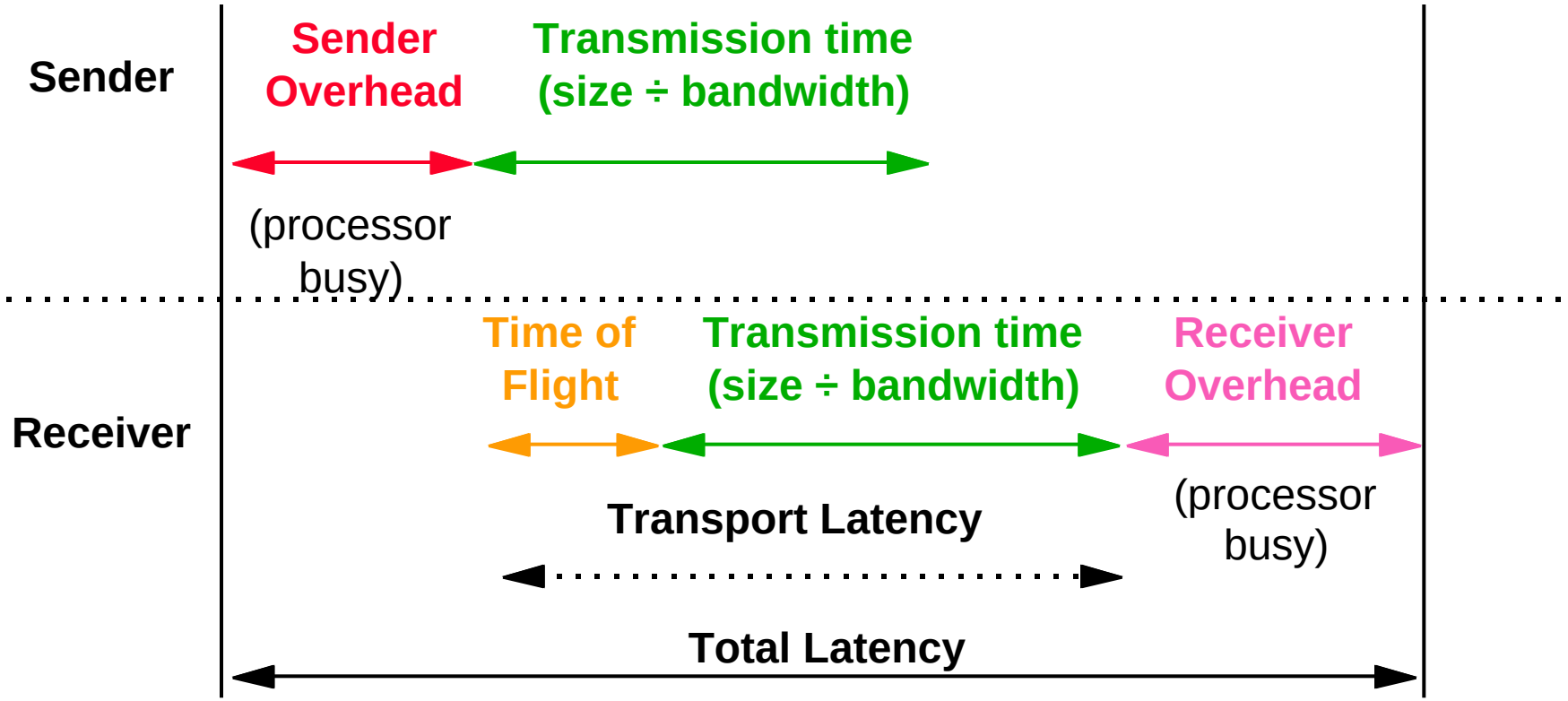
- **SW Send steps**
  - **1: Application copies data to OS buffer**
  - **2: OS calculates checksum, starts timer**
  - **3: OS sends data to network interface HW and says start**
- **SW Receive steps**
  - **3: OS copies data from network interface HW to OS buffer**
  - **2: OS calculates checksum, if matches send ACK; if not, *deletes message* (sender resends when timer expires)**
  - **1: If OK, OS copies data to user address space and signals application to continue**
- **Sequence of steps for SW: protocol**
  - **Example similar to UDP/IP protocol in UNIX**

Network Performance Measures



- **Overhead: latency of interface vs. Latency: network**

Universal Performance Metrics



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

Includes header/trailer in BW calculation?

Example Performance Measures

Interconnect	*MPP*	*LAN*	*WAN*
Example	CM-5	Ethernet	ATM
Bisection BW	N x 5 MB/s	1.125 MB/s	N x 10 MB/s
Int./Link BW	20 MB/s	1.125 MB/s	10 MB/s
Transport Latency	5 µsec	15 µsec	50 to 10,000 µs
HW Overhead to/from	0.5/0.5 µs	6/6 µs	6/6 µs
SW Overhead to/from	1.6/12.4 µs	200/241 µs	207/360 µs
		(TCP/IP on LAN/WAN)	

Software overhead dominates in LAN, WAN

CS 252 Administrivia

- **Upcoming events in CS 252**

23-Mar to 27-Mar Spring Break

Wed 1-Apr Networks

Thu 2-Apr Homework #3

Fri 3-Apr Networks

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)

Computers in the News

- **Plausible change in commercial software development?**
  - **See http://www.earthspace.net/~esr/writings/cathedral-bazaar/cathedral-bazaar.html by Eric S. Raymond**
  - **Cathedral model: programs built like cathedrals, carefully crafted by individual wizards or small bands of mages working in splendid isolation, with no beta to be released before its time**
  - **Bazaar model: One organinzation makes source code available– release early and often, delegate everything you can, be open to the point of promiscuity–resembling a great babbling bazaar of differing agendas and approaches (e.g., Linix)**
  - **Which model has more good people looking at code?**
  - **“Given enough eyeballs, all bugs are shallow”**
- **Open Source Foundation**
  - **New companies (e.g., Netscape, Sendmail, Scriptics) make some source code available, in return companies that make changes must return changes**
- **Impact on Computer Architecture?**

Total Latency Example

- 10 Mbit/sec., sending overhead of 230 µsec & receiving overhead of 270 µsec.
- a 1000 byte message (including the header), allows 1000 bytes in a single message.
- 2 situations: distance 0.1 km vs. 1000 km
- Speed of light = 299,792.5 km/sec (1/2 in media)
- $\text{Latency}_{0.1\text{km}}$ =

- $\text{Latency}_{1000\text{km}}$ =

- Long time of flight => complex WAN protocol

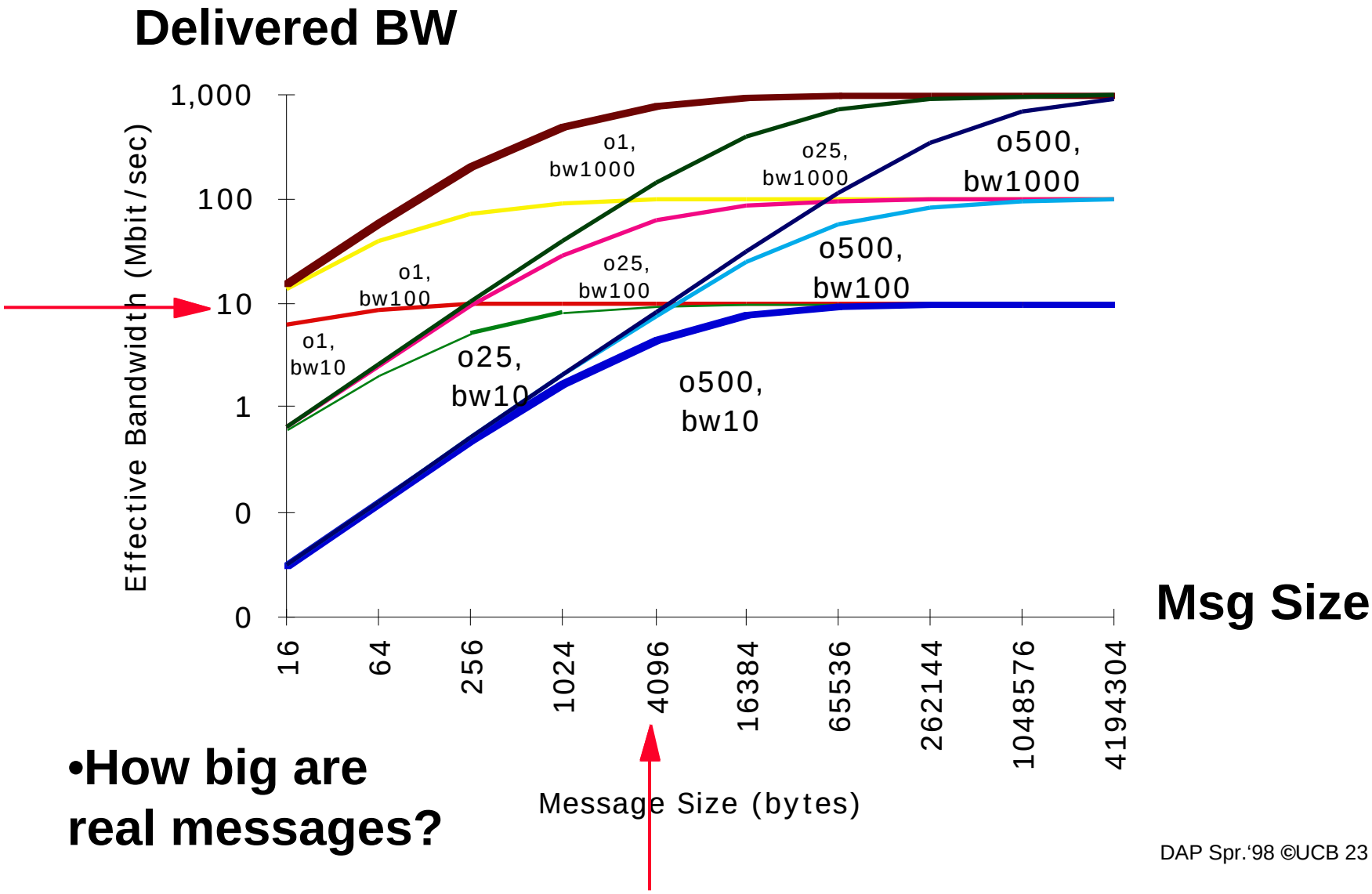
Total Latency Example

- 10 Mbit/sec., sending overhead of 230 µsec & receiving overhead of 270 µsec.
- a 1000 byte message (including the header), allows 1000 bytes in a single message.
- 2 situations: distance 100 m vs. 1000 km
- Speed of light = 299,792.5 km/sec
- $Latency_{0.1km}$ = 230 + 0.1km / (50% x 299,792.5) + 1000 x 8 / 10 + 270
- $Latency_{0.1km}$ = 230 + 0.67 + 800 + 270 = 1301 µsec
- $Latency_{1000km}$ = 230 + 1000 km / (50% x 299,792.5) + 1000 x 8 / 10 + 270
- $Latency_{1000km}$ = 230 + 6671 + 800 + 270 = 7971 µsec
- Long time of flight => complex WAN protocol

Simplified Latency Model

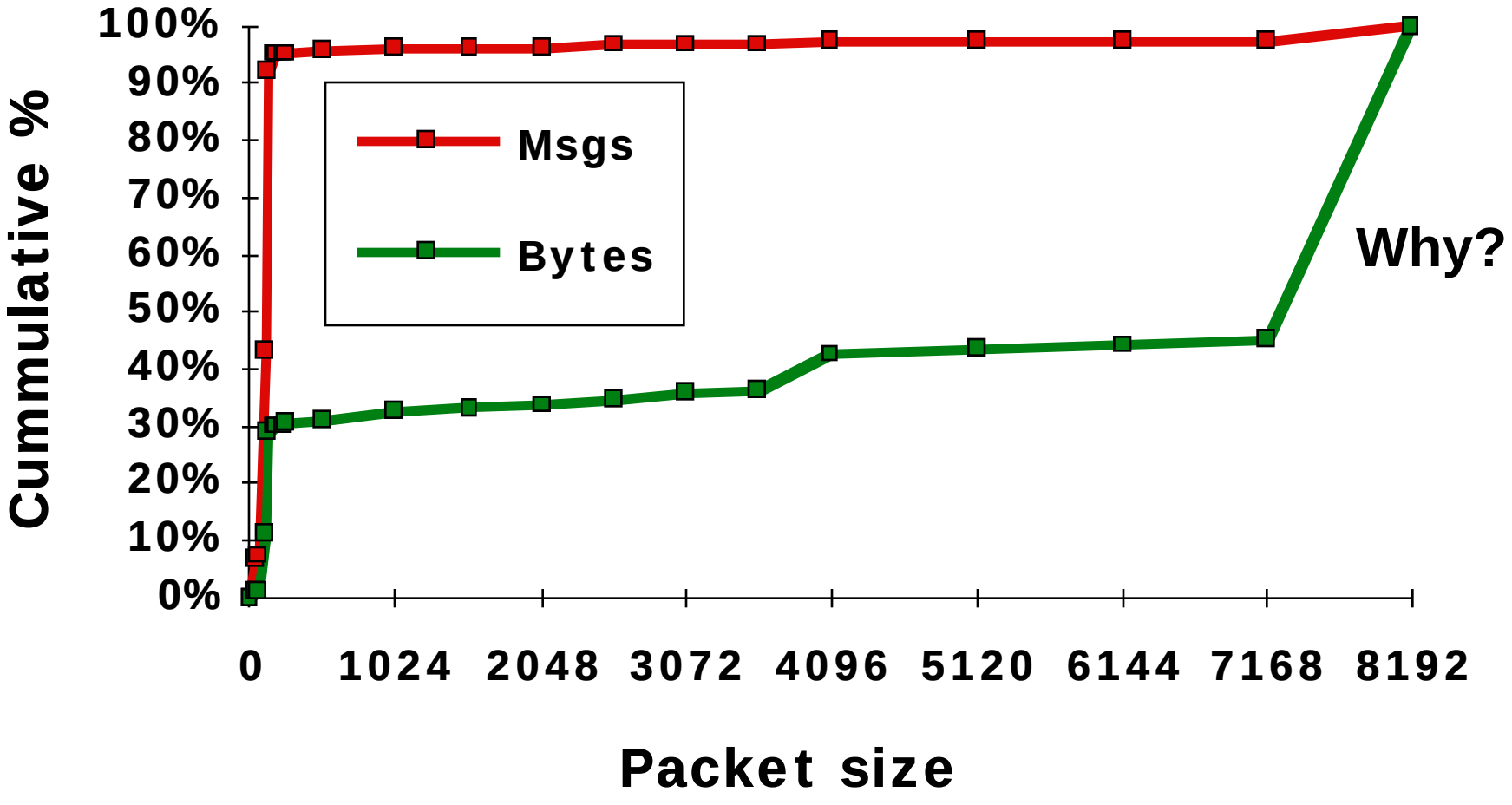
- **Total Latency ≈ Overhead + Message Size / BW**
- **Overhead = Sender Overhead + Time of Flight + Receiver Overhead**
- **Example: show what happens as vary**
  - **Overhead: 1, 25, 500 µsec**
  - **BW: 10,100, 1000 Mbit/sec (factors of 10)**
  - **Message Size: 16 Bytes to 4 MB (factors of 4)**
-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

Overhead, BW, Size



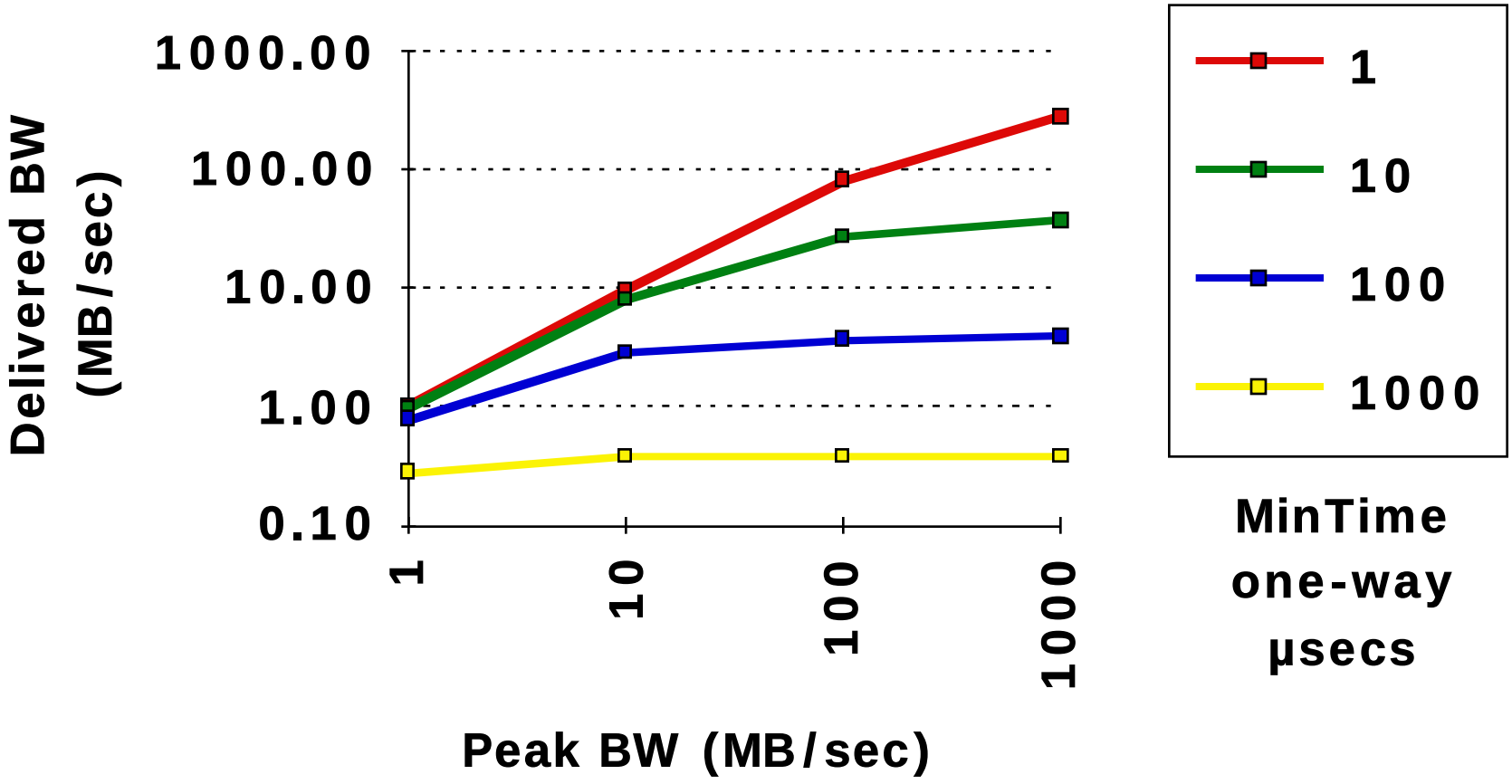
- **How big are real messages?**

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



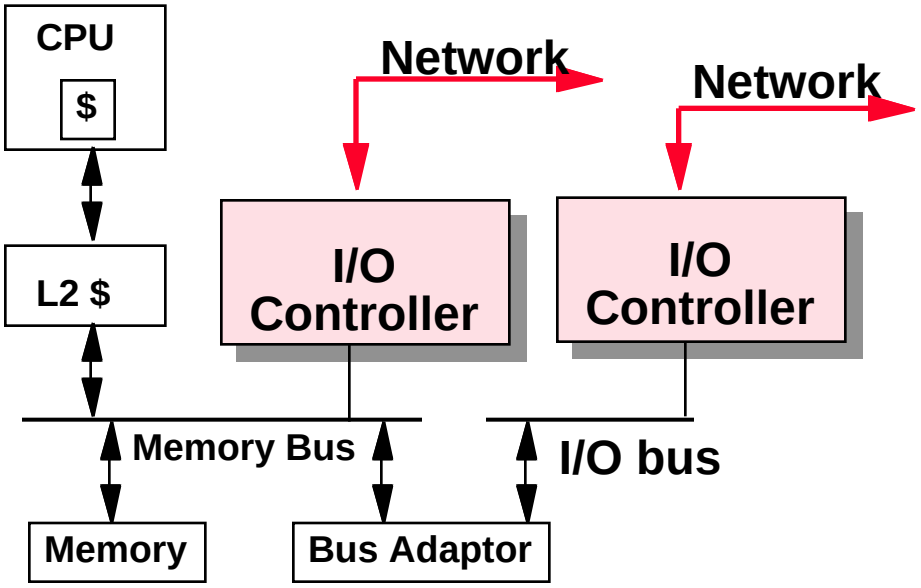
- BW model: Time = overhead + msg size/peak BW
- > 50% data transfered in packets = 8KB

Interconnect Issues

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

HW Interface Issues

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



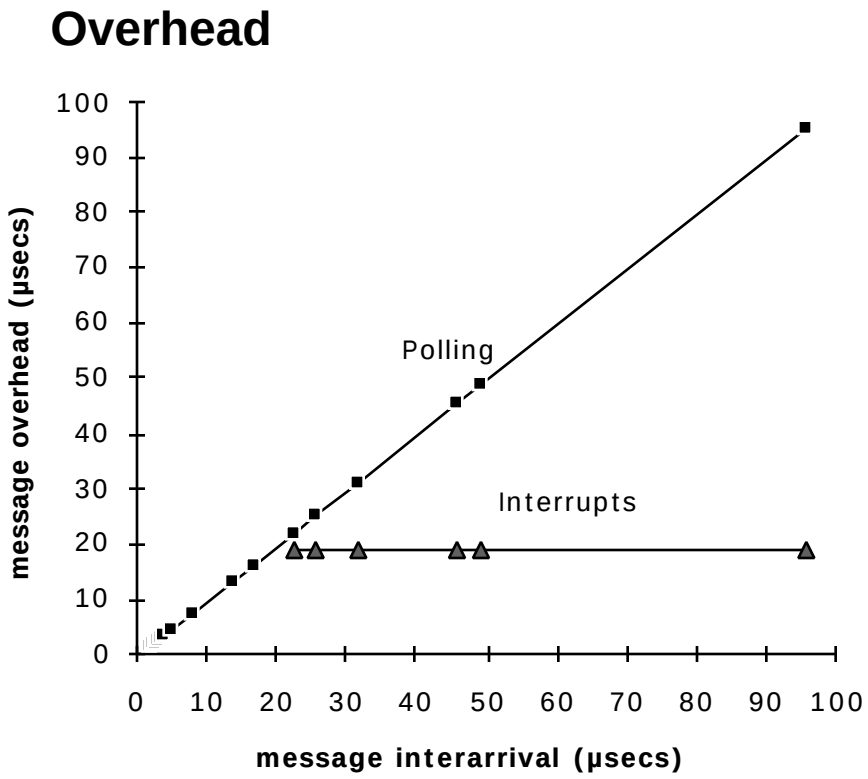
ideal: high bandwidth, low latency, standard interface

SW 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 uncached 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**

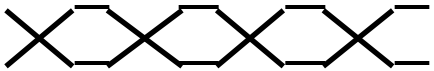


Interconnect Issues

- Performance Measures
- Interface Issues
- Network Media

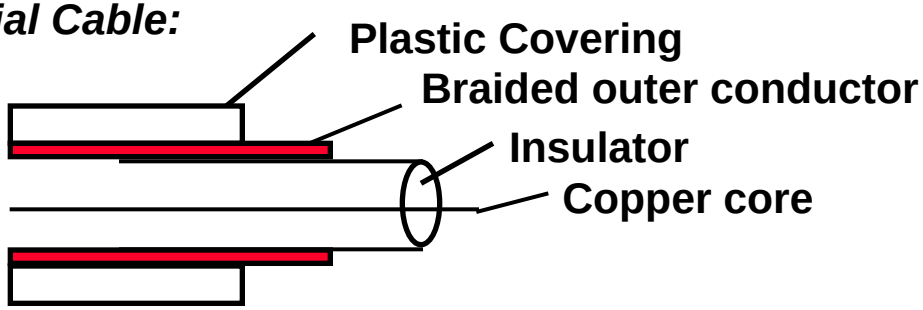
Network Media

Twisted Pair:



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

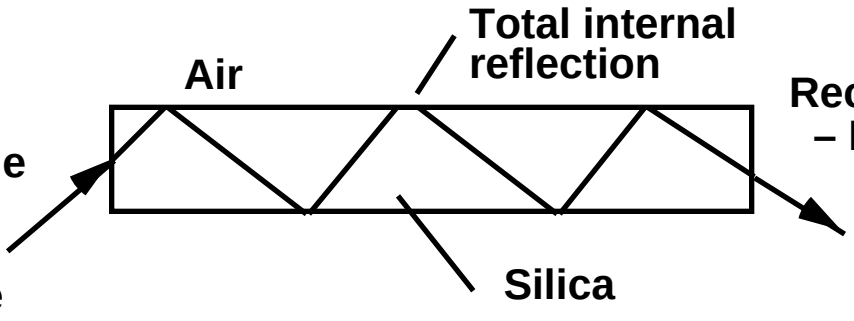
Coaxial Cable:



Used by cable companies: high BW, good noise immunity

Fiber Optics

Transmitter
– L.E.D
– Laser Diode



Receiver
– Photodiode

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

Costs of Network Media (1995)

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

Note: more elaborate signal processing allows higher BW from copper (ADSL)

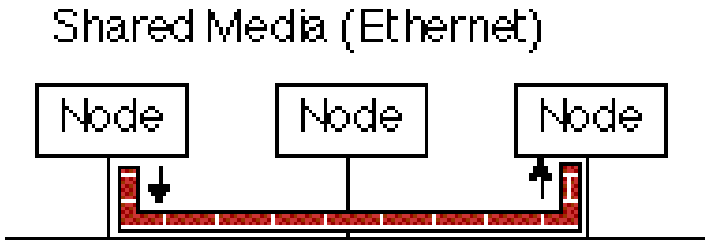
Single mode Fiber measures: BW * distance as 3X/year

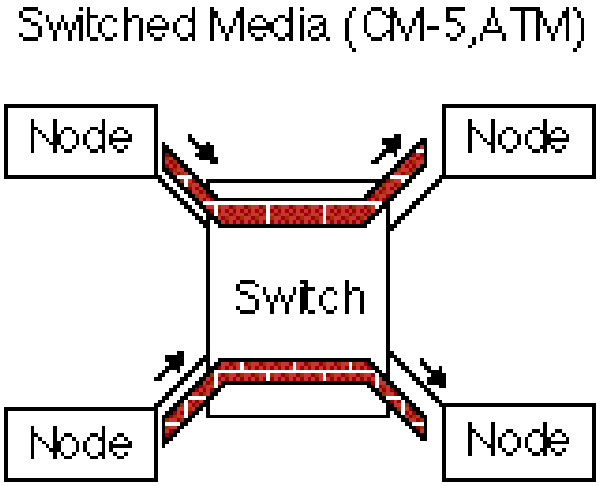
Interconnect Issues

- 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, simpler interface**
- **Arbitration in Shared network?**
  - **Central arbiter for LAN?**
  - **Listen to check if being used ("Carrier Sensing")**
  - **Listen to check if collision ("Collision Detection")**
  - **Random resend to avoid repeated collisions; not fair arbitration;**
  - **OK if low utilization**





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

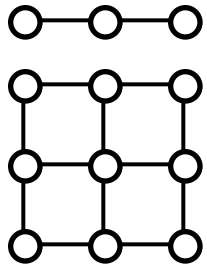
Example Interconnects

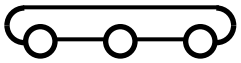
Interconnect	*MPP*	*LAN*	*WAN*
Example	CM-5	Ethernet	ATM
Maximum length between nodes	25 m	500 m; ≤5 repeaters	copper: 100 m optical: 2 km—25 km
Number data lines	4	1	1
Clock Rate	40 MHz	10 MHz	≥ 155.5 MHz
Shared vs. Switch	Switch	Shared	Switch
Maximum number of nodes	2048	254	> 10,000
Media Material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

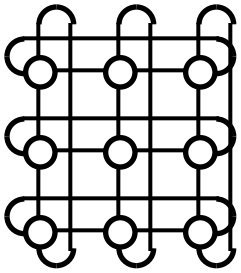
Switch Topology

- 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 (worst case)
- 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 (largely 2D)
- Networks should not be interesting!

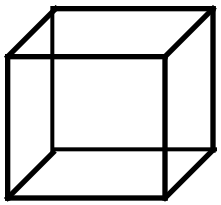
Important Topologies







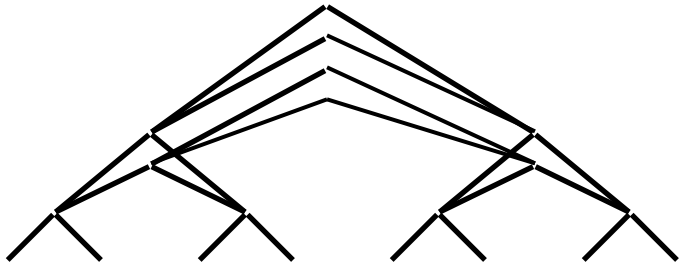
Type	Degree	Diameter	Ave Dist	Bisection	N = 1024 Diam	Ave D
1D mesh	≤ 2	N-1	N/3	1		
2D mesh	≤ 4	$2(N^{1/2} - 1)$	$2N^{1/2} / 3$	$N^{1/2}$	63	21
3D mesh	≤ 6	$3(N^{1/3} - 1)$	$3N^{1/3} / 3$	$N^{2/3}$	~30	~10
nD mesh ($N = k^n$)	$\leq 2n$	$n(N^{1/n} - 1)$	$nN^{1/n} / 3$	$N^{(n-1) / n}$		
Ring	2	N / 2	N/4	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						



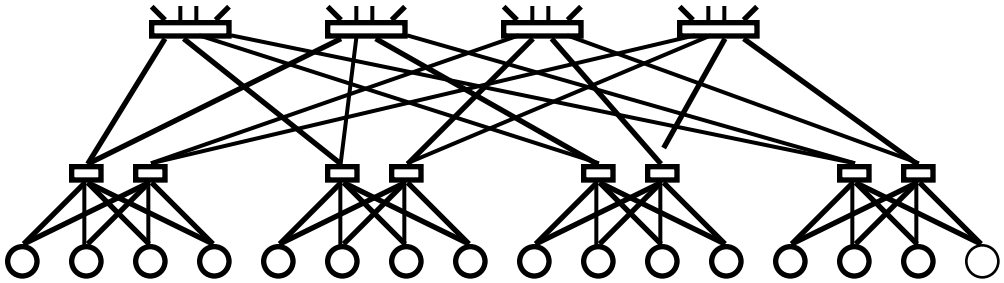
Hypercude 2^3

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



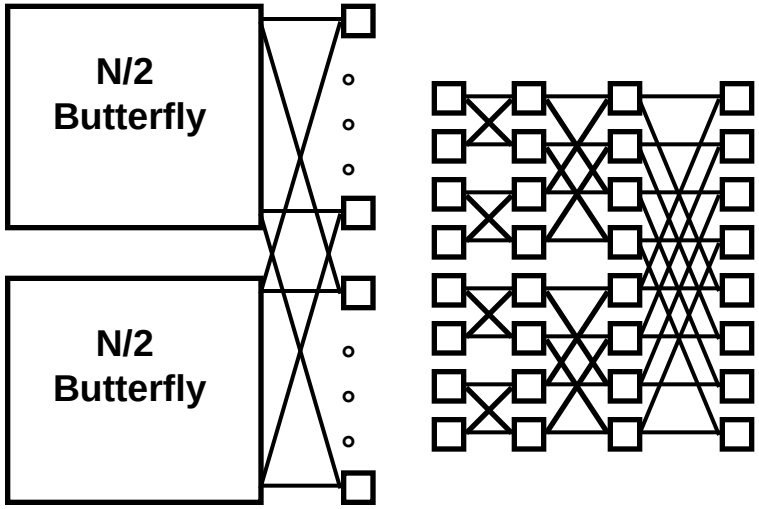
Fat Tree



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 that try to avoid
- Don’t want algortihm to have to know paths

Example MPP 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 MPP topology!

Summary: 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. 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, not algorithms