Lecture 28: Networks & Interconnect—Architectural Issues

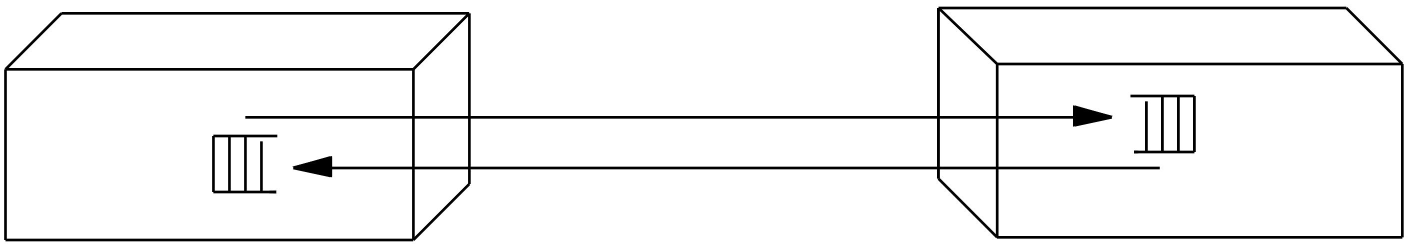
Professor Randy H. Katz

Computer Science 252

Spring 1996

Review: ABCs of Networks

- Starting Point: Send bits between 2 computers



- 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

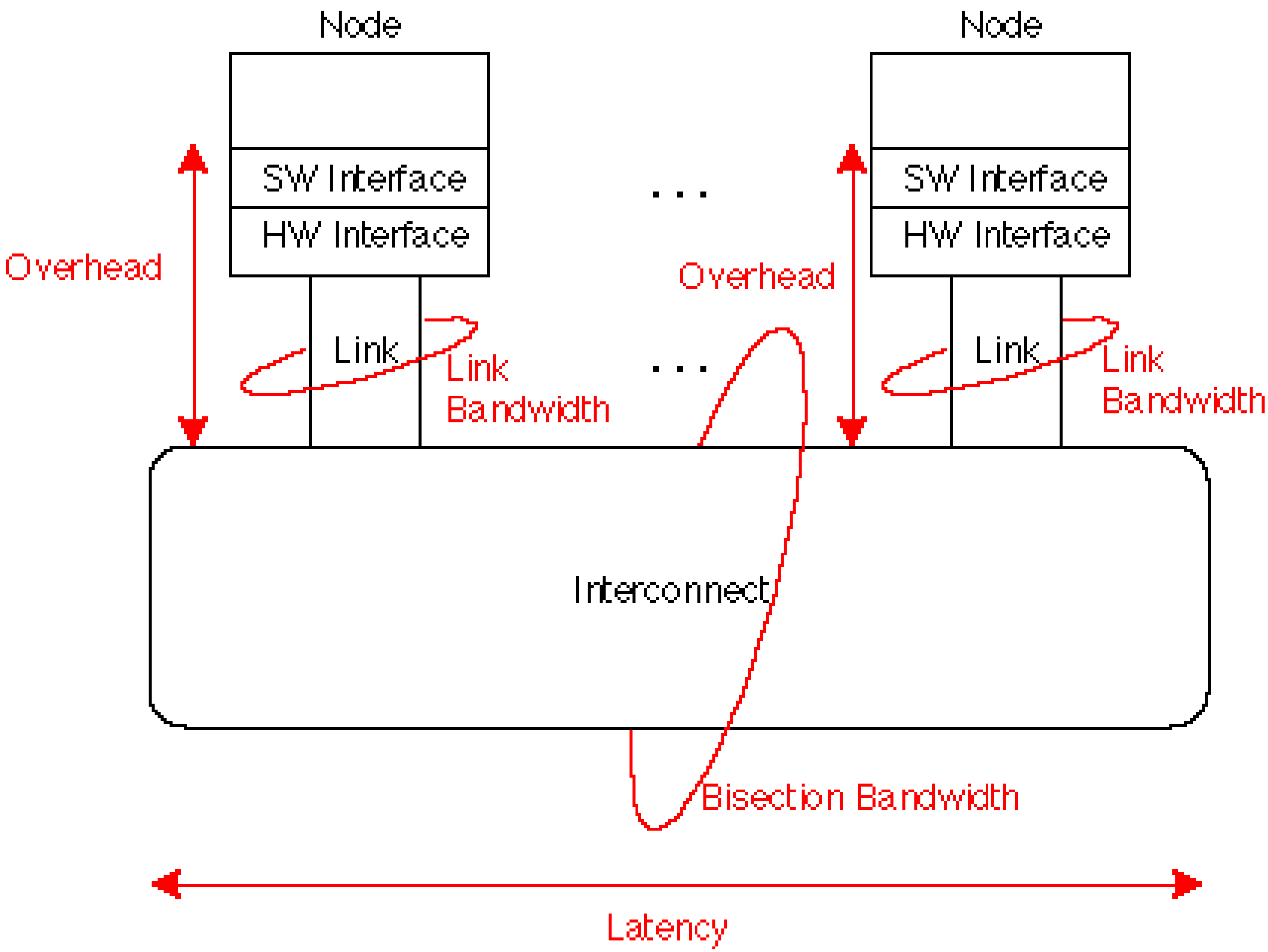
Review: Questions about Simple Network

- **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

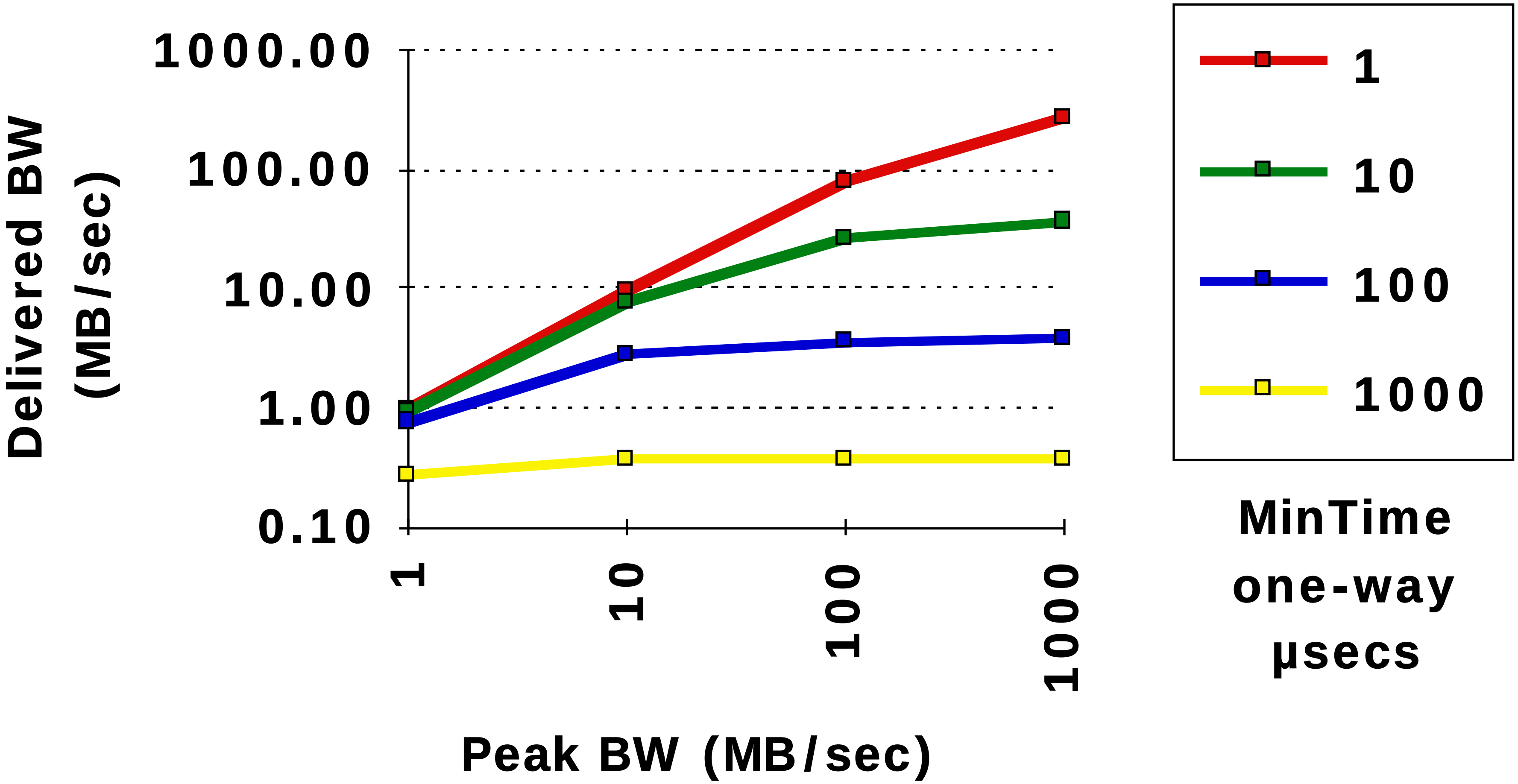
Interconnect	*MPP*	*LAN*	*WAN*
Example	CM-5	Ethernet	ATM
Maximum length between nodes	25 m	500 m; ≤5 repeaters	copper: 100 m optical(multimode): 2 km optical(single mode): 25 km
Number data lines	4	1	1
Clock Rate	40 MHz	10 MHz	≥ 155.5 MHz
Shared vs. Switch	Switch	Shared	Switch
Maximum number of nodes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

Review: Network Performance



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

Review: Impact of Overhead on Delivered BW



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

Review: Interconnect Issues

- Implementation Issues
- Performance Measures
- Architectual Issues
- Practical Issues

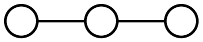
Example Architecture Measures

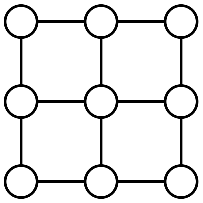
Interconnect	*MPP*	*LAN*	*WAN*
Example	CM-5	Ethernet	ATM
Topology	"Fat" tree	Line, Bus	Variable, constructed from multistage switches
Connection based?	No	No	Yes
Data Transfer Size	Variable: 4 to 20B	Variable: 0 to 1500B	Fixed: 48B
Store & Forward?	No	n.a.	Yes
Congestion control	At source: Flow control via back pressure	At source: Listen for E-net idle	Rate based via choke packets

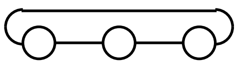
Topology

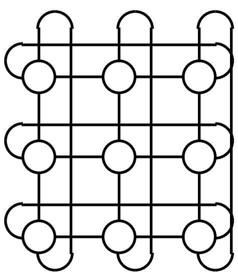
- **Structure of the interconnect**
- **Determines**
  - **Degree: number of links from a node**
  - **Diameter: max number of links crossed between nodes**
  - **Average distance: number of hops to random destination**
  - **Bisection: minimum number of links that separate the network into two halves**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

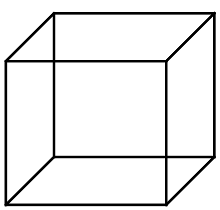
Important Topologies







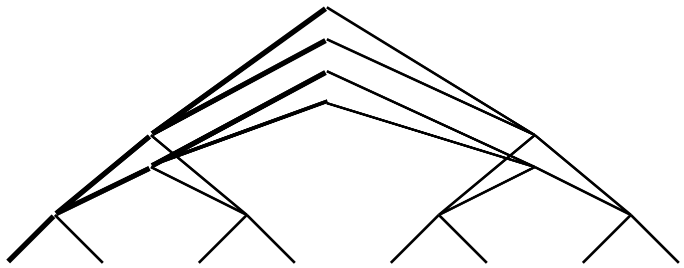


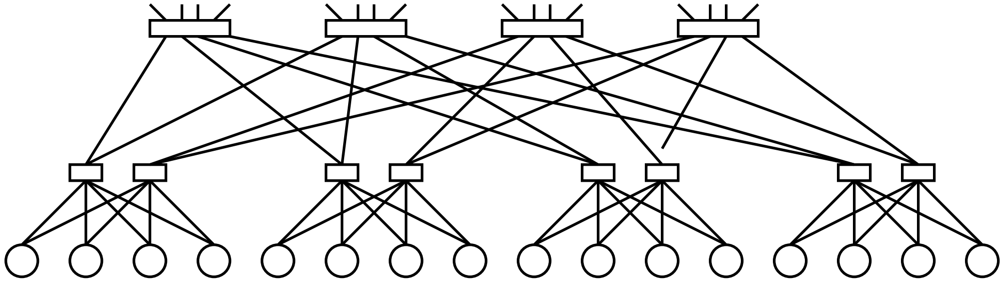


					N = 1024	
Type	**Degree**	**Diameter**	**Ave Dist**	**Bisection**	**Diam**	**Ave D**
1D mesh	≤ 2	N-1	N/3	1		
2D mesh	≤ 4	$2(N^{1/2} - 1)$	$2N^{1/2} / 3$	$N^{1/2}$	63	21
3D mesh	≤ 6	$3(N^{1/3} - 1)$	$3N^{1/3} / 3$	$N^{2/3}$	~30	~10
nD mesh	≤ 2n	$n(N^{1/n} - 1)$	$nN^{1/n} / 3$	$N^{(n-1) / n}$		
$(N = k^n)$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

Topologies (cont)

Type	Degree	Diameter	Ave Dist	Bisection	N = 1024 Diam	Ave D
2D Tree	3	$2Log_2 N$	$\sim 2Log_2 N$	1	20	~20
4D Tree	5	$2Log_4 N$	$2Log_4 N - 2/3$	1	10	9.33
kD	k+1	$Log_k N$				
2D fat tree	4	$Log_2 N$		N		
2D butterfly	4	$Log_2 N$		N/2	20	20

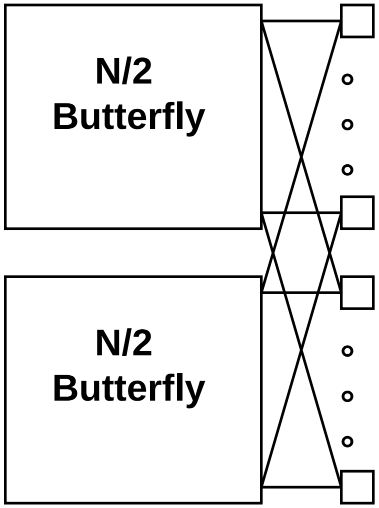


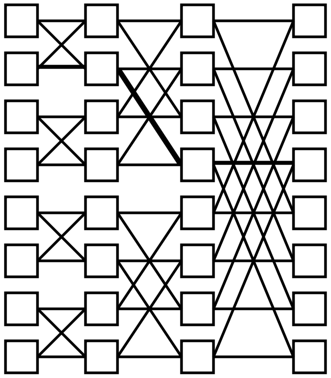


CM-5 Thinned Fat Tree

Butterfly

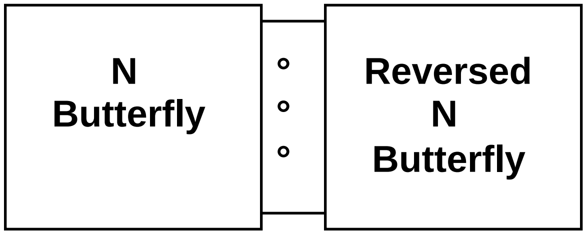
Multistage: nodes at ends, switches in middle





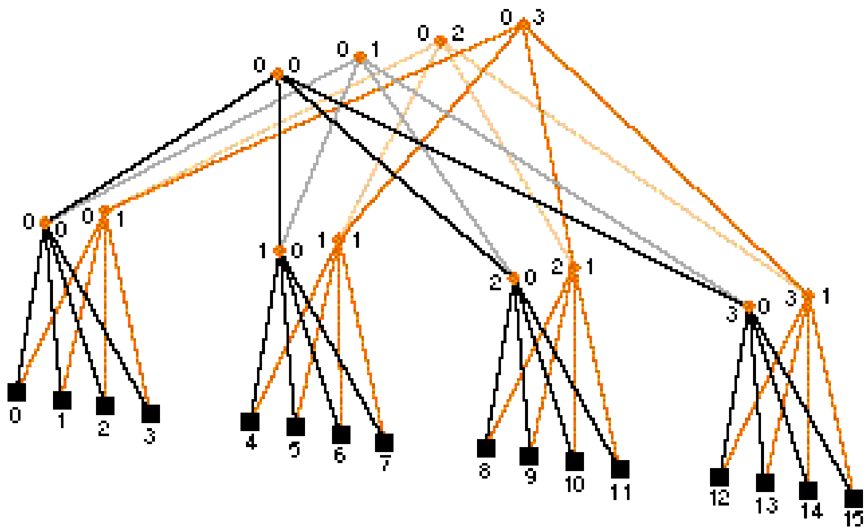
- All paths equal length
- Unique path from any input to any output
- Conflicts

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 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!

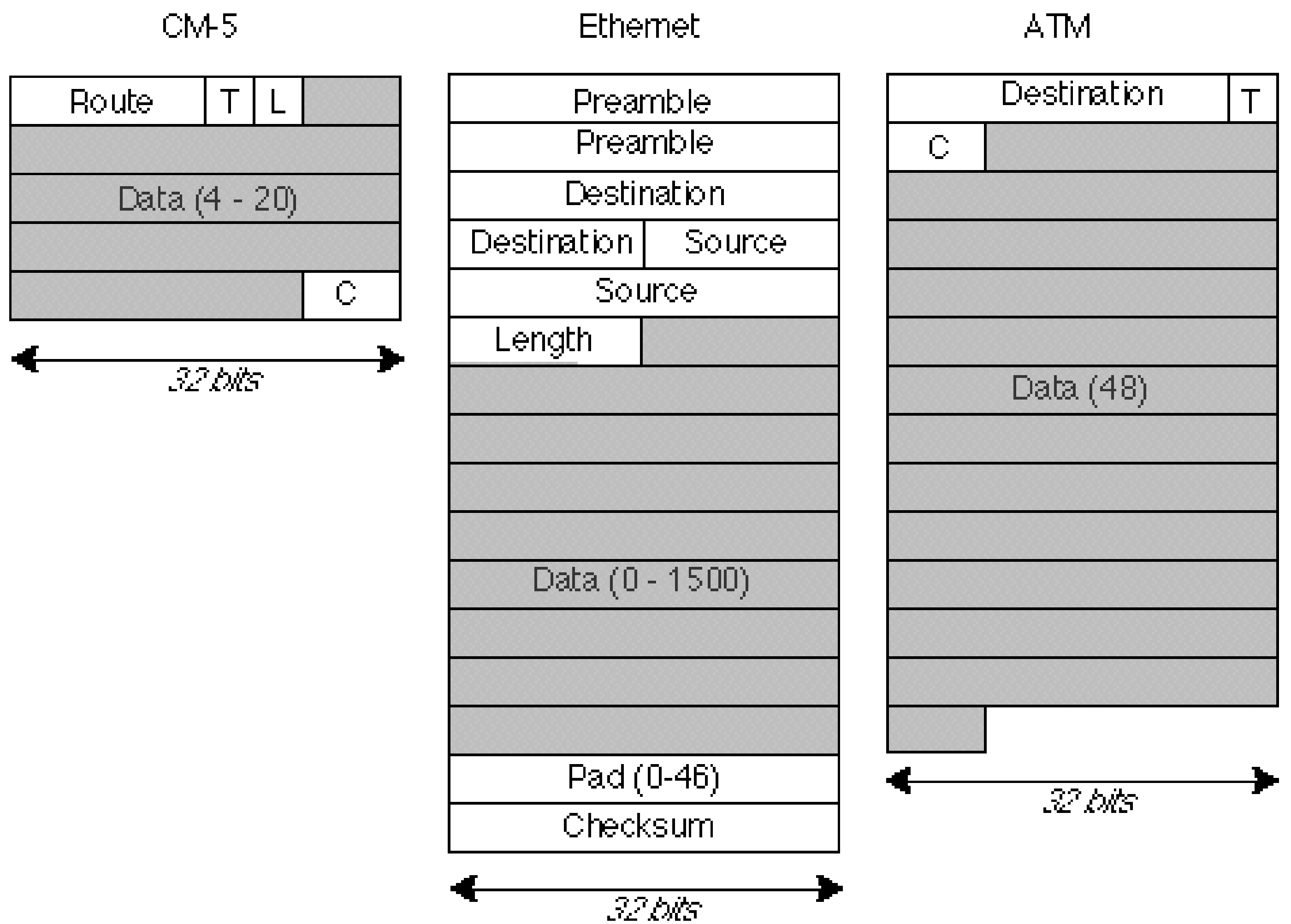
Connection-Based vs. Connectionless

- **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**
  - **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:** $L = H(b + R_{sf})$ (Link transfer time)
  - **Cut-through:** $L = HR_{ct} + b$
  - **Typically,** $R_{ct} = 32/w$
  - **Most modern networks are about 1 small packet deep**

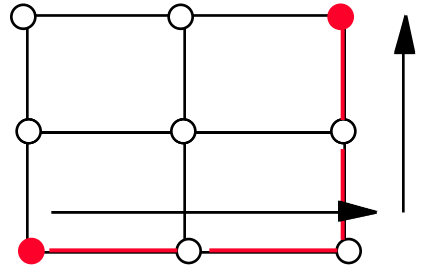
Examples

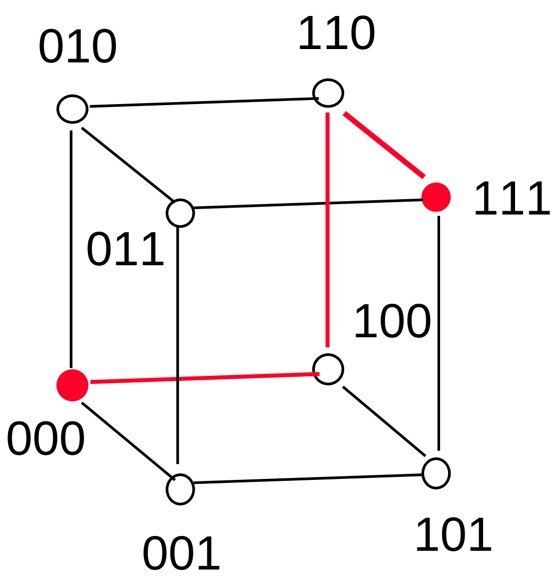
Machine	Network	W	R	rate MHz	MB/s per link
nCUBE/2	13-D hypercube	1	40	40	2.2
CM-5	4-D Thinned Fat Tree	4	8	40	20
Delta	2-D Mesh	8	2	~40	40
T3D	3-D Torus	16	?	150	300

- **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: $L = HR_{ct} + m/w$ (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**
    - » **$X = x_o x_1 x_2 \ldots x_n$ -> $Y = y_o y_1 y_2 \ldots y_n$**
    - » **R = X xor Y**
    - » **Traverse dimensions of differing address in order**
  - **tree: common ancestor**
- **Adaptive—route based on network state (e.g., contention)**
- **Deadlock free**





Practical Issues

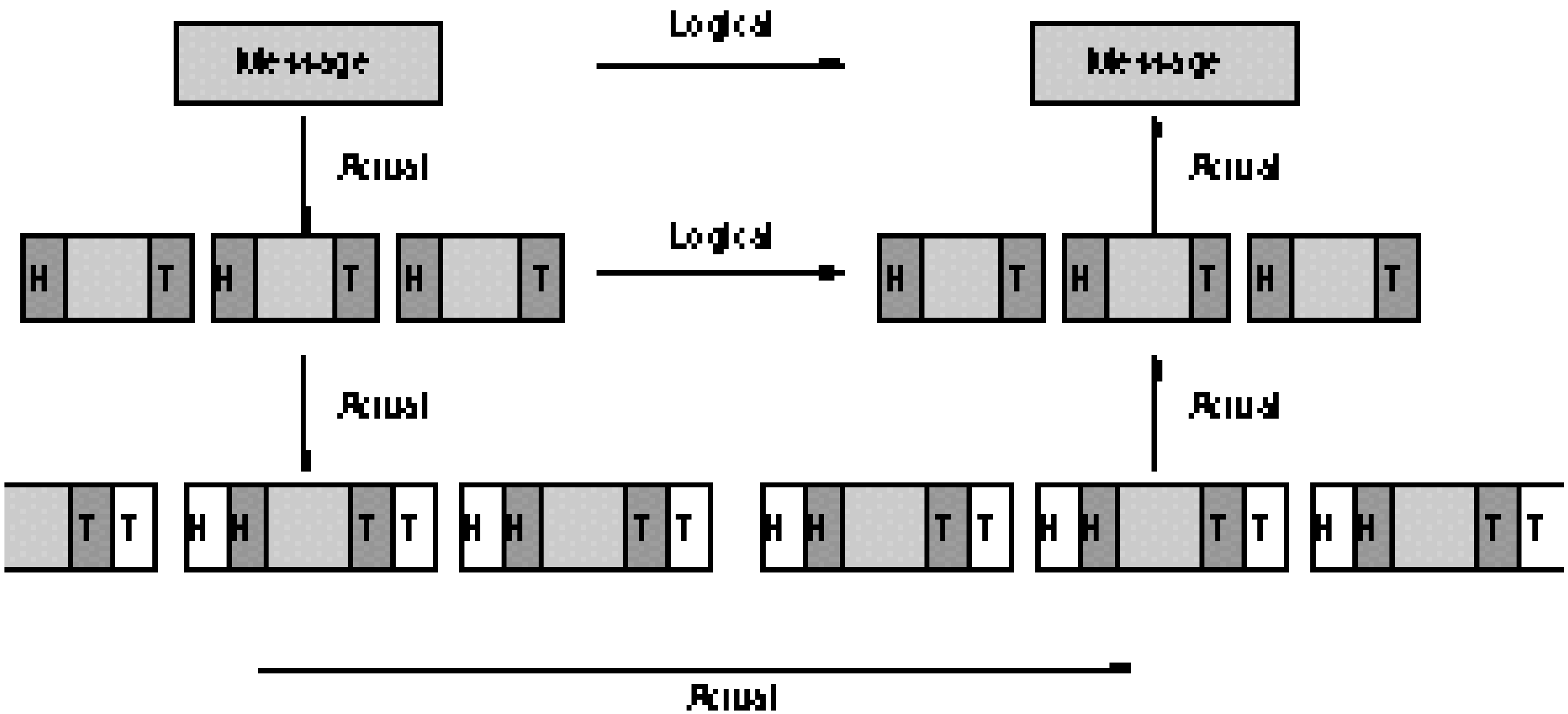
Interconnection	MPP	LAN	WAN
Example	CM-5	Ethernet	ATM
Standard	No	Yes	Yes
Fault Tolerance?	No	Yes	Yes
Hot Insert?	No	Yes	Yes

- **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!**

Protocols: HW/SW Interface

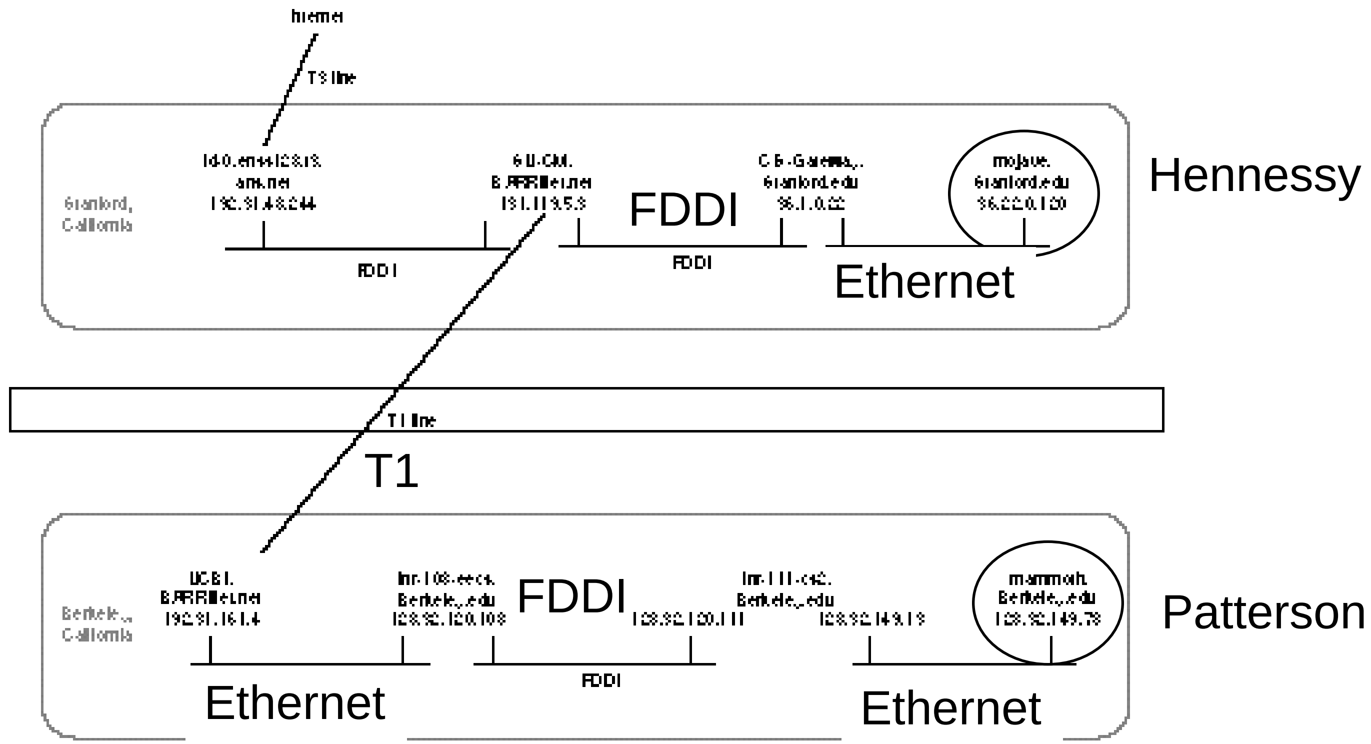
- **Internetworking: allows computers on independent and incompatible networks to communicate reliably and efficiently;**
  - **Enabling technologies: SW standards that allow reliable communications without reliable networks**
  - **Hierarchy of layers, giving each layer responsibility for portion of 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**