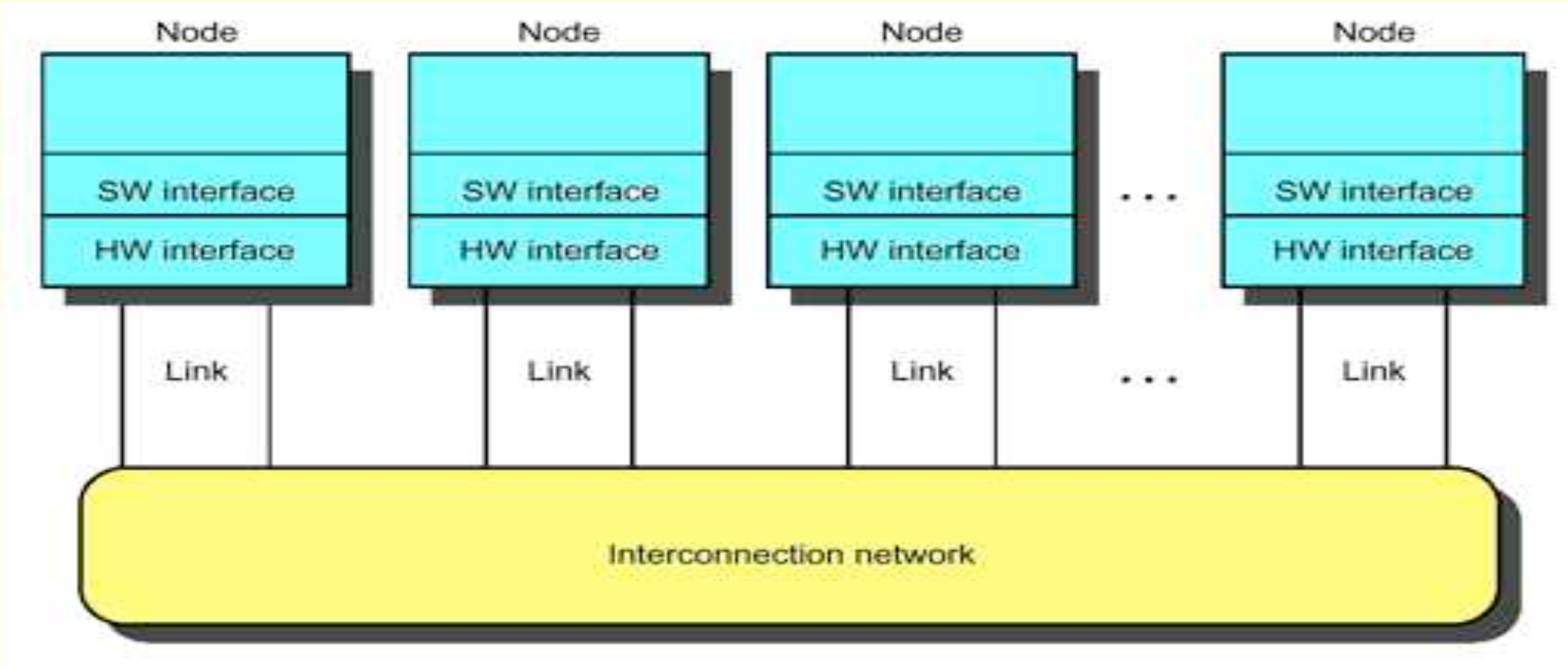
CMSC 611: Advanced Computer Architecture

Interconnection Networks

Interconnection Networks



Massively parallel processor networks (MPP)

- Thousands of nodes
- Short distance (<~25m)
- Traffic among all nodes

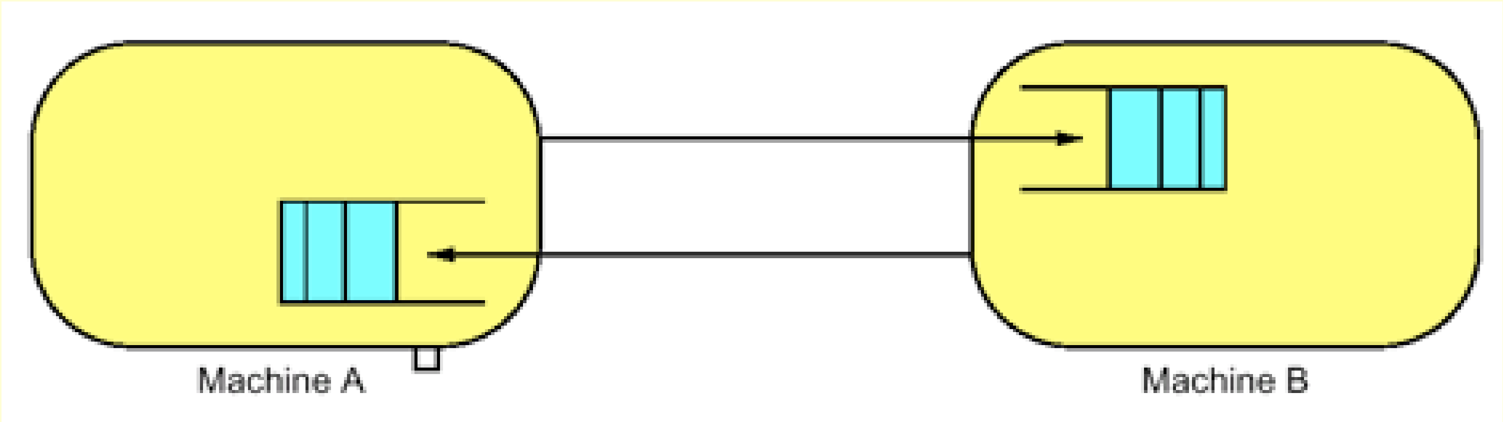
Local area network (LAN)

- Hundreds of computers
- A few kilometers
- Many-to-one (clients-server)

Wide area network (WAN)

- Thousands of computers
- Thousands of kilometers

ABCs of Networks

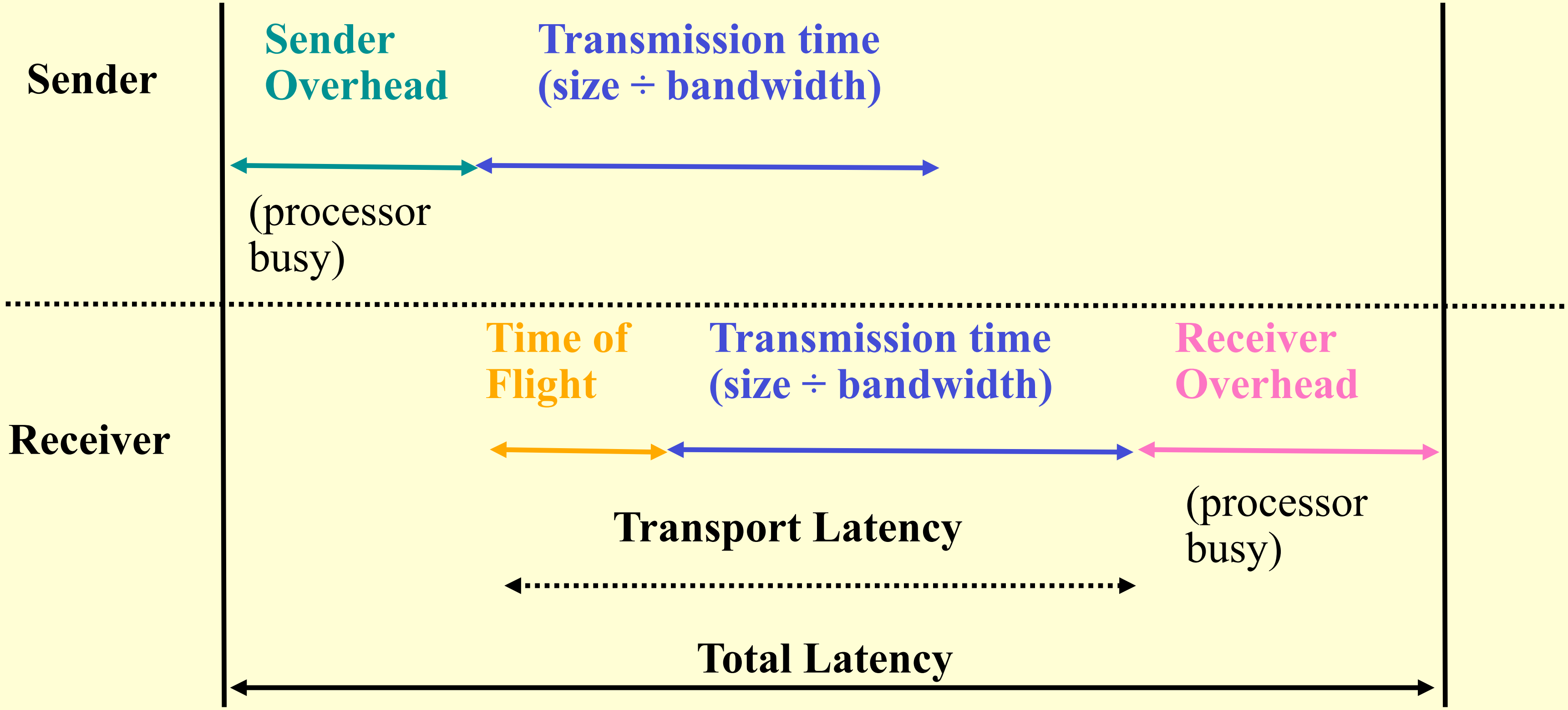


Rules for communication are called the "protocol", message header and data called a "packet"

- 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?
  - Time-out, retransmit; ACK & NACK
- What if multiple processes/machine?
  - Queue per process to provide protection

* Slide is partially a courtesy of Dave Patterson

Performance Metrics



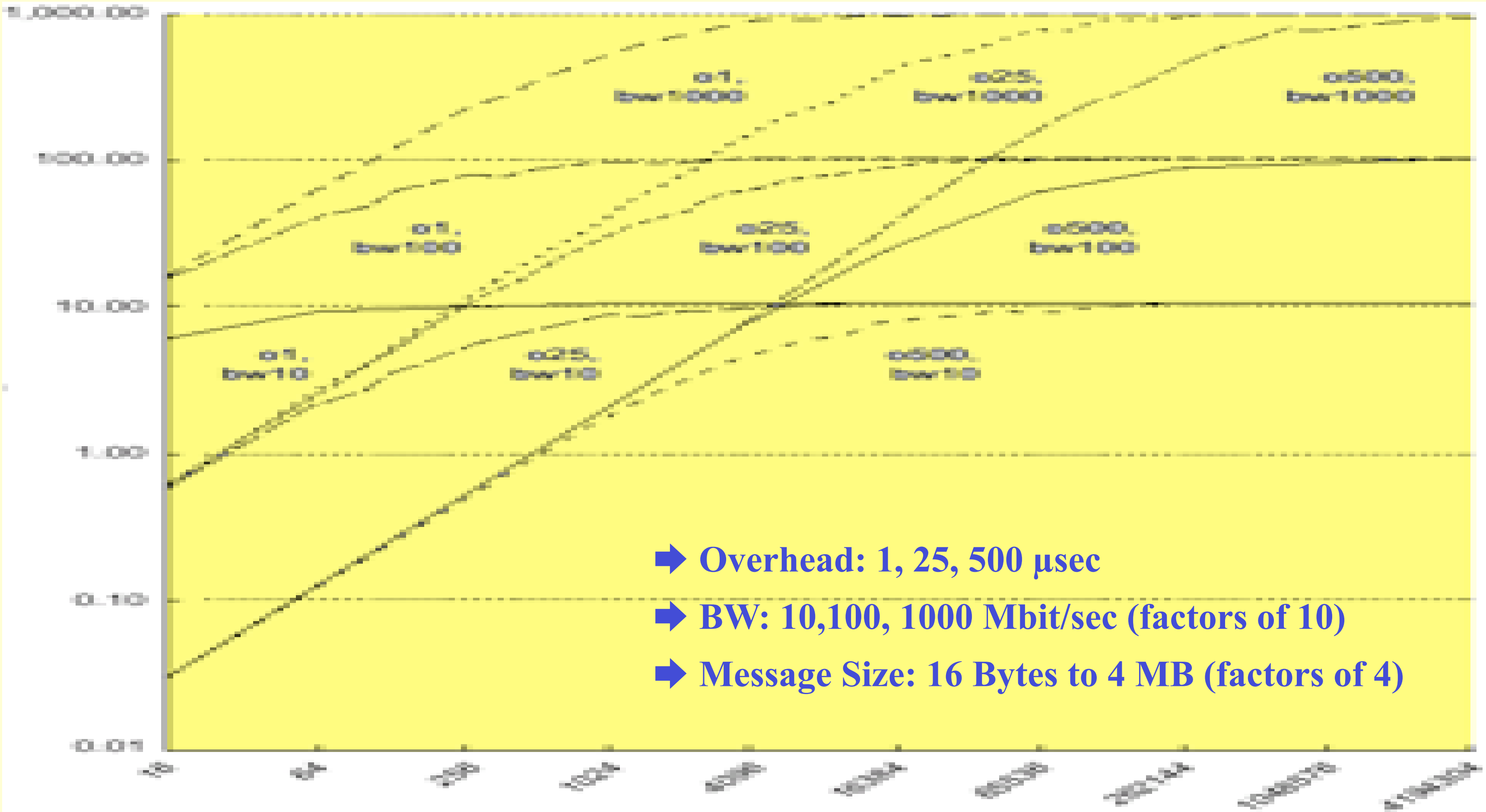
$$\text{Total latency} = \text{Sender Overhead} + \text{Time of flight} + \frac{\text{Message size}}{\text{Bandwidth}} + \text{Receiver overhead}$$

Bandwidth: maximum rate of propagating information

Time of flight: time for 1st bit to reach destination

Overhead: software & hardware time for encoding/ decoding, interrupt handling, etc.

Performance Measures



Effective Bandwidth (Mbit/sec)

Message Size (bytes)

- Overhead: 1, 25, 500 µsec
- BW: 10,100, 1000 Mbit/sec (factors of 10)
- Message Size: 16 Bytes to 4 MB (factors of 4)

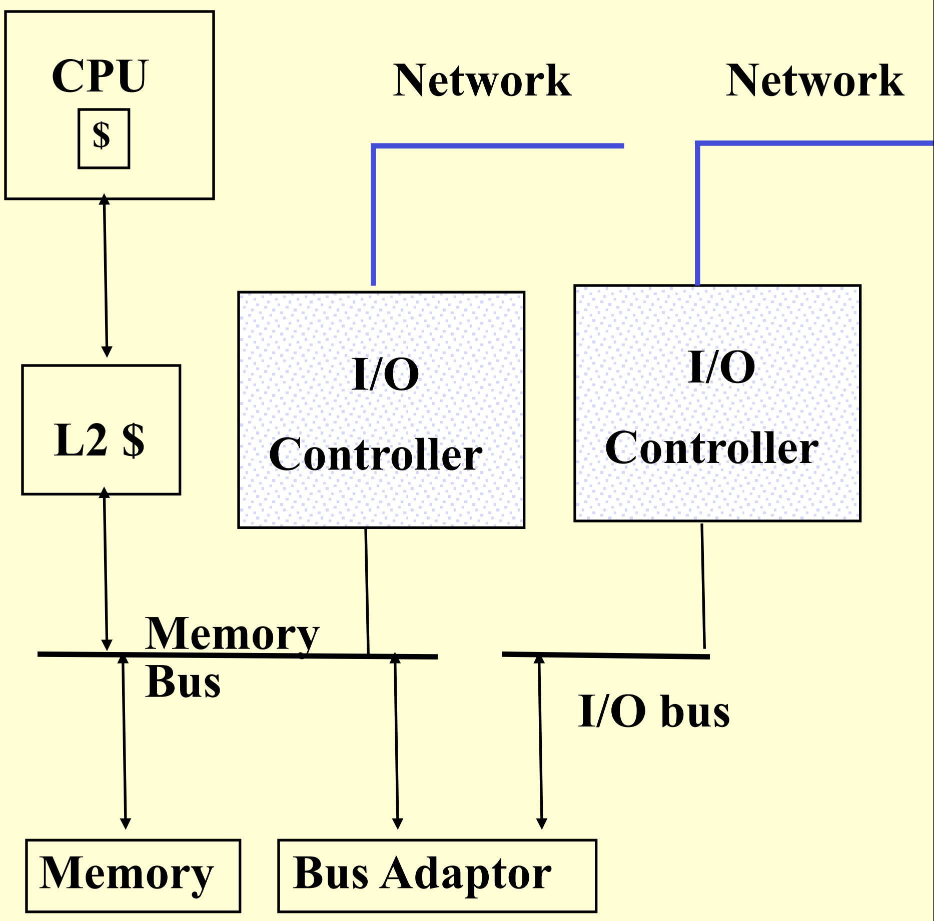
$$\text{Effective Bandwidth} = \frac{\text{Message Size}}{\text{Total Latency}}$$

Large messages needed to justify high overhead

Network Interface Issues

Where to connect network to computer?

- Cache consistency to avoid flushes (⇒ memory bus)
- Low latency and high bandwidth (⇒ memory bus)
- Standard interface card? (⇒ I/O bus)
- Typically, MPP uses memory bus; while LAN, WAN connect through I/O bus



Ideal: high bandwidth, low latency, standard interface

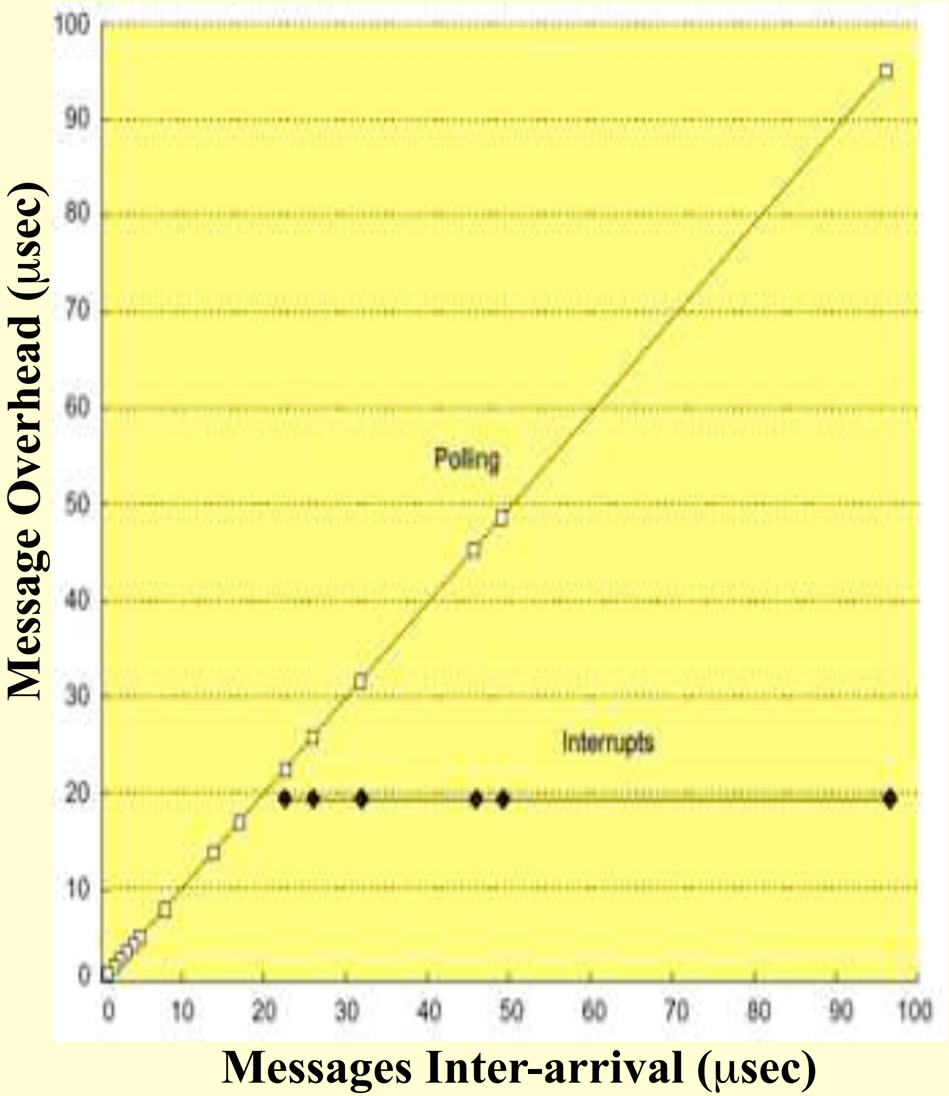
* Slide is a courtesy of Dave Patterson

Network Interface Issues

How to connect network to software?

- Programmed I/O (low latency)
- DMA? (best for large messages)
- Receiver interrupted or received polls?
- Avoid involving operating system in common case
- Avoid operating at non-cached memory speed (e.g., check network interface)

Example: CM-5 Software Interface



* Slide is a courtesy of Dave Patterson

CM-5 example (MPP)

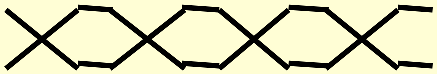
- Allows sending message without involving the operating system
- Receiver can poll or use interrupts to detect messages
- Time per polling 1.6 μsecs
- Time per interrupt 19 μsecs
- Minimum time to handle message: 0.5 μsecs
- Enable/disable 4.9/3.8 μsecs

As rate of messages arriving changes, use polling or interrupt?

- Avoid enabling and disabling interrupts due to high cost
- Always enable interrupts, have interrupt routine poll until no messages pending
  - Low rate => - interrupt
  - High rate => - polling

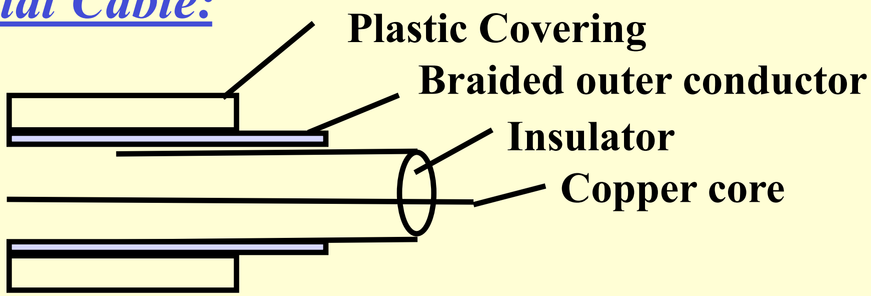
Network Media

Twisted Pair:



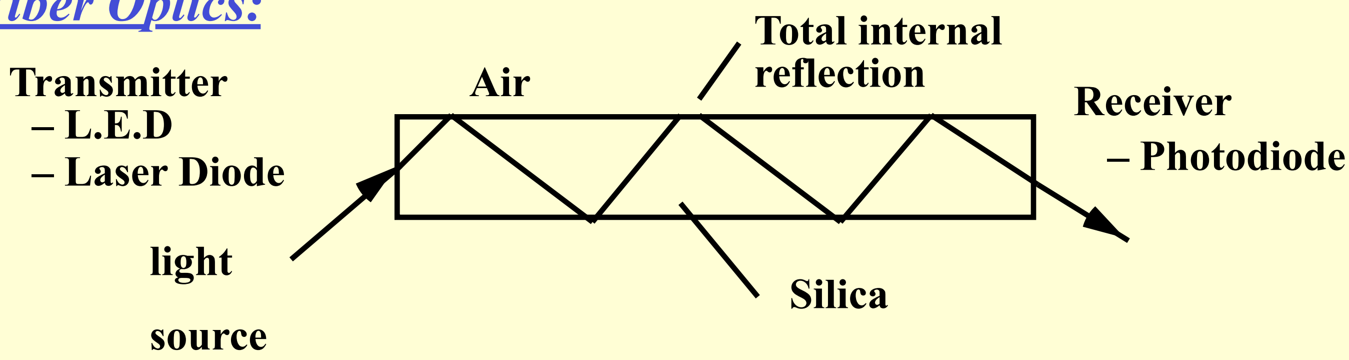
Copper, 1mm think, twisted to avoid antenna effect, suitable for telephone and LANs

Coaxial Cable:



Used by cable companies: high BW, good noise immunity, typically 10Mbit/sec over a kilometer

Fiber Optics:



Light: 3 parts are cable, light source, light detector

- Multimode light disperse (LED) allows 600 Mbit/sec for up to 2 Km
- Single mode single wave (laser) reaches gigabits/sec for hundreds of Km

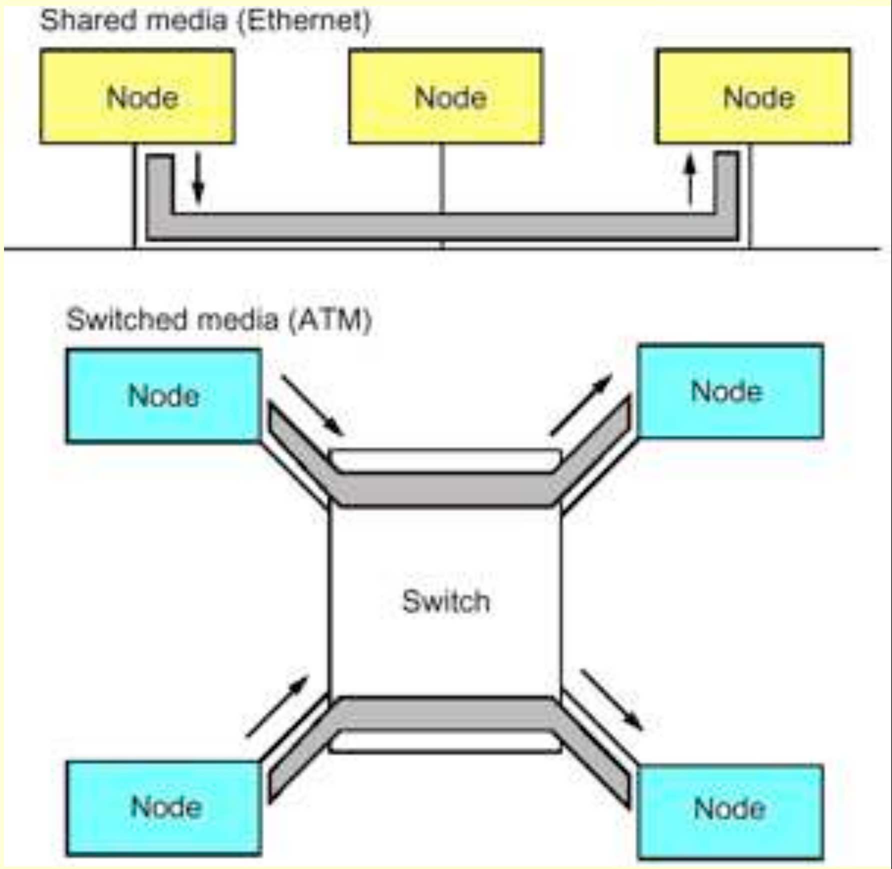
Connecting Multiple Computers

Shared Media vs. Switched

- Shared medium facilitates broadcasting and multicasting
- Aggregate BW in *switched* network is many times *shared*
  - point-to-point faster since no arbitration, simpler interface
  - switch increases latency

Shared network arbitration?

- 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)
- OK if low utilization

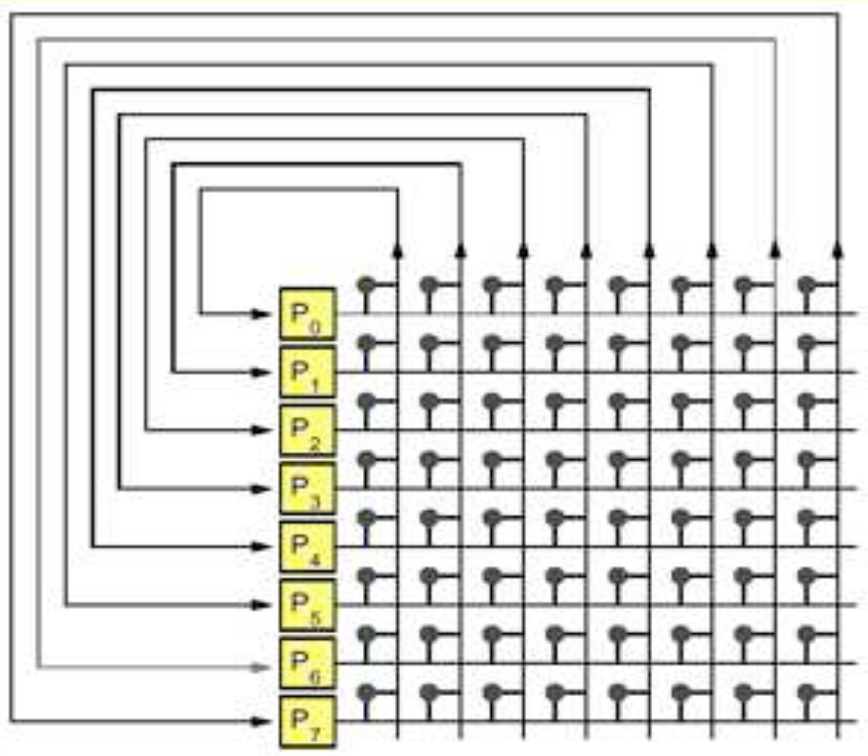


While all nodes have to share 10 Mbit/sec Ethernet connection, ATM can support multiple 155 Mbit/sec simultaneous transfers

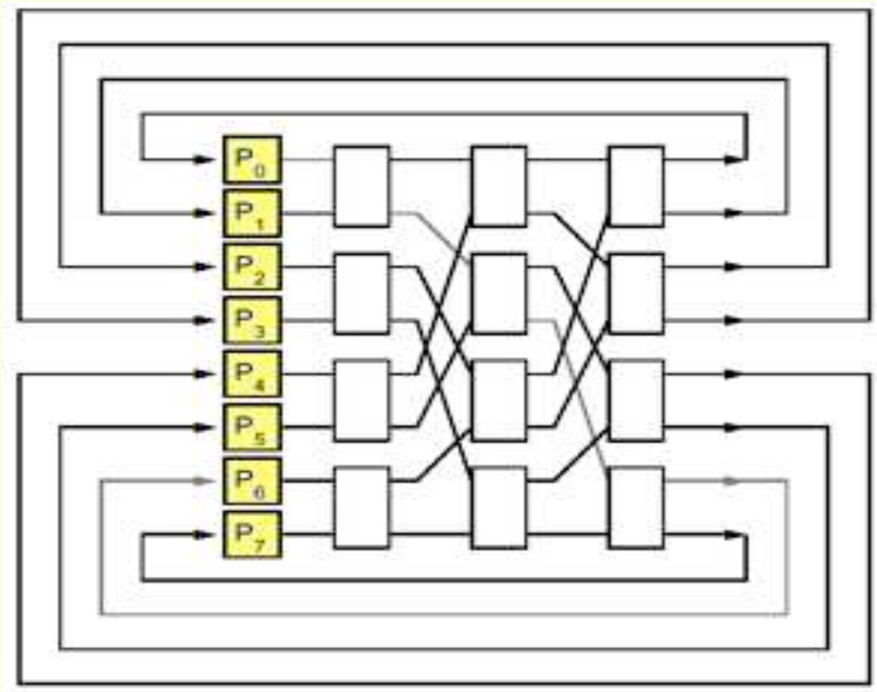
Switch Topology

Structure of the interconnect and determines

- **Degree**: number of links from a node
- **Diameter**: max number of links crossed between nodes
- **Average distance**: number of hops to random destination

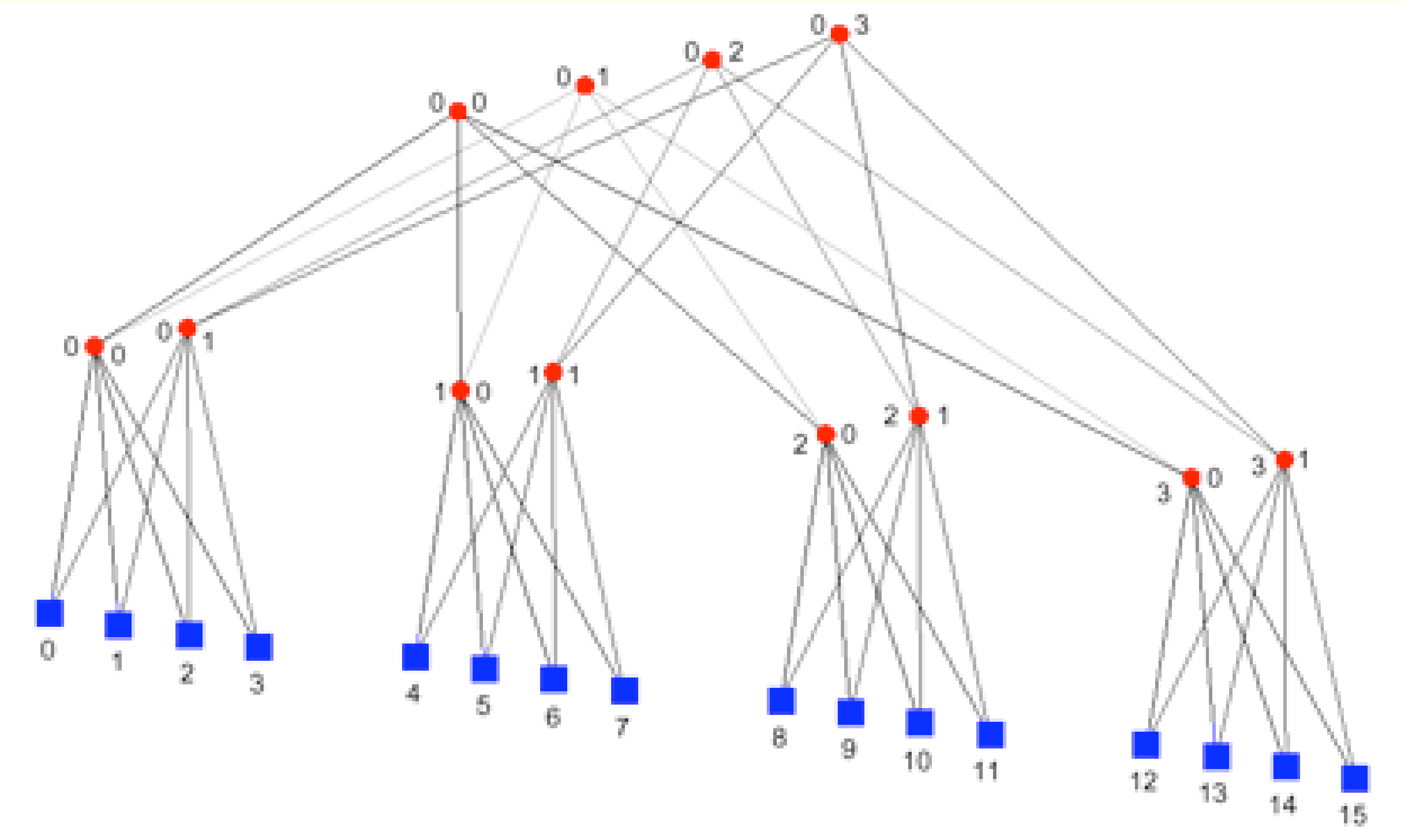


Cross bar uses n^2 switches and allows simultaneous routing of any permutation of traffic pattern among processor



Omega network uses ½ $n\log_2 n$ switches each uses 4 internal small switches (total is less than cross bar) but restrict routes

Example: Fat-Tree Topology

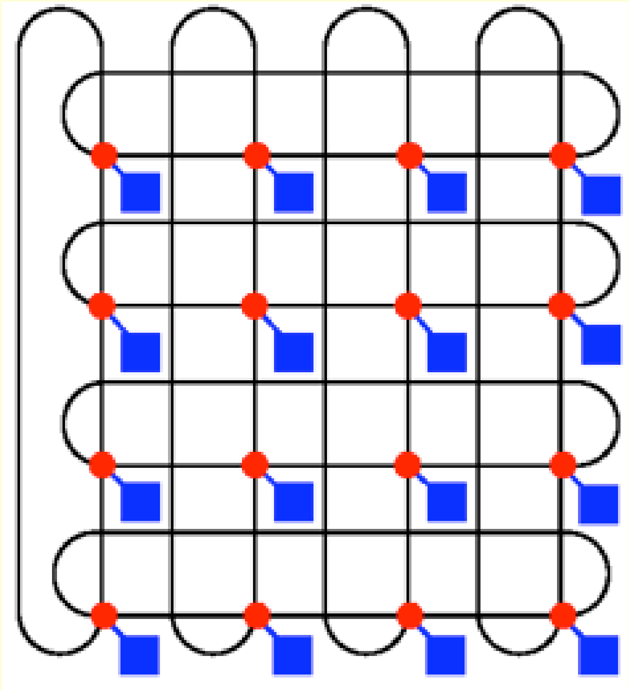


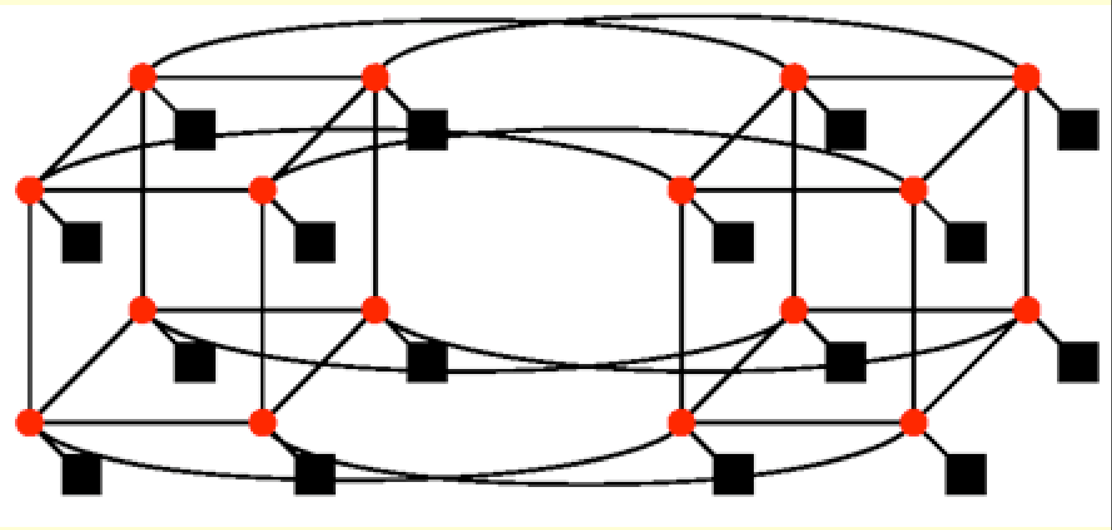
Increase the bandwidth via extra links at each level over a simple tree

Intermediate switches have two upward links and 4 downward links

Can handle multiple common communication patterns very well

Commercial MMP Topologies





2D torus of 16 nodes

- Ensures fully connected network
- Increases availability through redundant paths
- Enhances performance via splitting traffic and avoiding contention

Boolean hypercube tree of 16 nodes

Generally n-dimensional interconnect for 2^n nodes requiring n+1 ports per switches for the processor and nearest n neighbor nodes

Connection-based Communication

Telephone: operator sets up connection between the caller and the receiver

- Once the connection is established, conversation can continue for hours
- Generally use circuit switching to establish connection between communicating parties

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 ensures quality of service

Connectionless Communication

Every package of information must have an address

- **Packet**: one package of information

Each packet is routed to its destination by looking at its address

- Analogy, the postal system (sending a letter)

Also called “Statistical multiplexing” given the role of queuing theory in measuring performance

Circuit-based communication can be established on top of packet switched network

- TCP/IP

Packet-based communication can be established over a circuit-switched network

- e.g. UDP over ssh

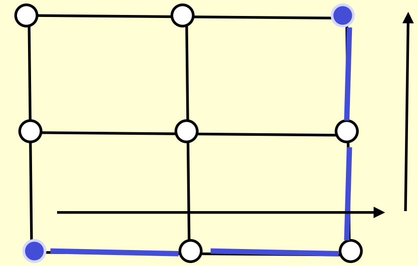
Routing Messages

Shared Media: broadcast to everyone and let the receiver pick it

Switched Media needs real routing since the path is not clear

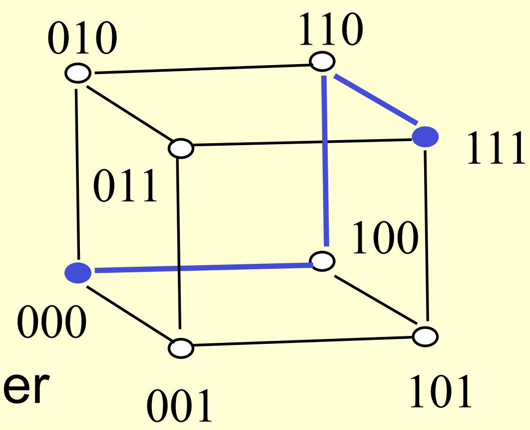
- **Source-based routing**: message specifies path to destination (provides directions)
- **Virtual Circuit**: circuits established from source to destination, message picks the circuit to follow
- **Destination-based routing**: message specifies destination, switch must pick the path
  - **deterministic**: always follow same path after establishing one
  - **adaptive**: pick different paths to avoid congestion, failures
  - **Randomized routing**: pick between several good paths to balance network load

Routing Examples



Mesh: dimension-order routing

- $(x_1, y_1) \rightarrow (x_2, y_2)$
- Deterministic
  - first x, then y
- Adaptive
  - At x,y, when $x \neq x_2$ and $y \neq y_2$
  - Pick least-congested direction



Hypercube: edge-cube routing

- $X = x_0x_1x_2 \ldots x_{n;}$ $Y = y_0y_1y_2 \ldots y_n$
- R = X xor Y
- Deterministic
  - Traverse dimensions of differing address in order
- Adaptive
  - Choose 1-bit in direction of least congestion

* Slide is a courtesy of Dave Patterson

Buffering Policy

Store-and-forward policy: each switch waits for the full packet to arrive in switch before sending to the next switch (good for WAN)

- Latency is function of: number of intermediate switches multiplied by the size of the packet

Cut-through routing or **worm-hole routing**: switch examines the header and then starts forwarding it immediately (common in MPP)

- **Worm hole**: when head of message is blocked, message stays strung out over the network, potentially blocking other messages (only buffer the piece of the packet that is sent between switches)
- **Cut through**: Tail continues when head is blocked, compressing the whole message into a single switch (Requires a buffer large enough to hold the largest packet)
- Latency is function of: time for 1st part of the packet to negotiate the switches + the packet size ÷ interconnect bandwidth

* Slide is a courtesy of Dave Patterson

Congestion Control

Connection based networks reserve bandwidth ahead of time and limit input to such capacity

Packet switched networks do not reserve bandwidth; this leads to contention

Contention not only increase latency unpredictably but also can cause deadlocks

Solution: prevent packets from entering until contention is reduced (e.g., freeway on-ramp metering lights)

* Slide is a courtesy of Dave Patterson

Congestion Control 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 (common in MPP)
- Window: give original sender right to send N packets before getting permission to send more; overlaps latency of interconnection with overhead to send & receive packet (e.g., TCP), adjustable window

Choke packets: 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

- Required for WAN and LAN but not MPP
  - \+ low cost (components used repeatedly)
  - \+ stability (many suppliers to chose from)
  - – Time for committees to agree
  - – When to standardize?
    - Before anything built? ⇒ Committee does design?
    - Too early suppresses innovation

Fault Tolerance: Can nodes fail and still deliver messages to other nodes?

- Required for WAN and LAN and difficult to ensure in MPP

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 and LAN

* Slide is a courtesy of Dave Patterson

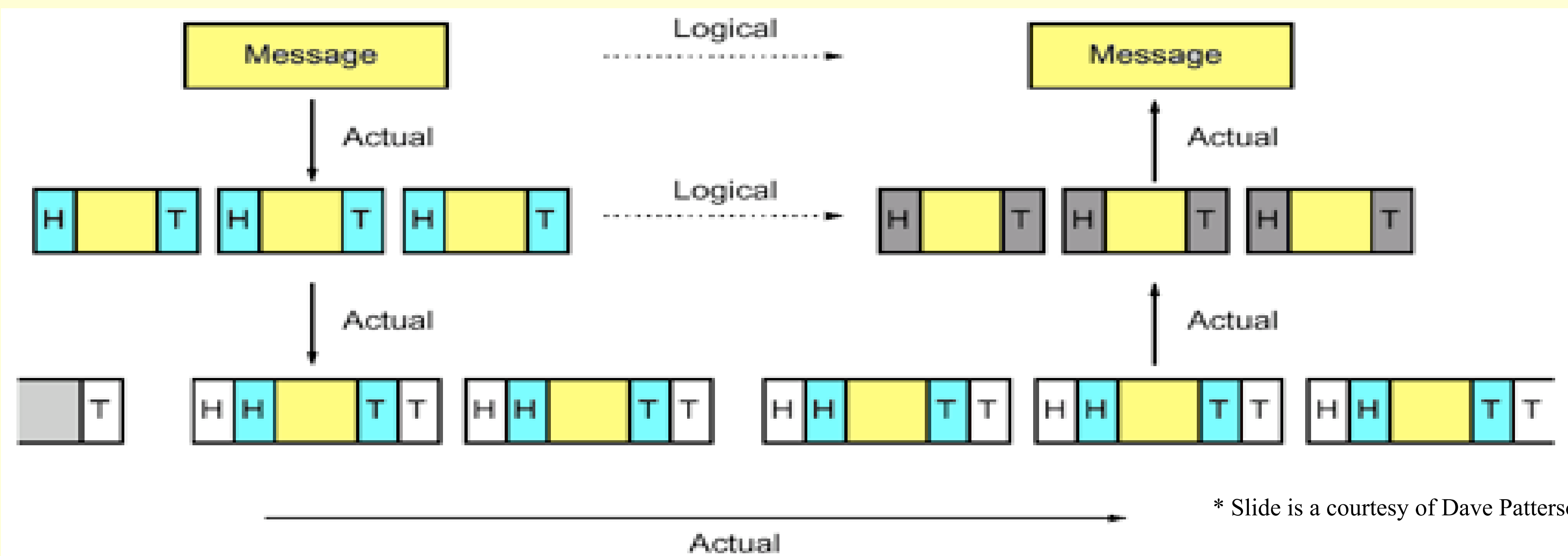
Examples

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

Internetworking

Internetworking allows computers on independent and incompatible networks to communicate

- Enabling technologies: software standards that allow reliable communications without reliable networks
- Hierarchy of SW layers (**protocol stack**), giving each layer responsibility for portion of overall communications task, called protocol families or protocol suites

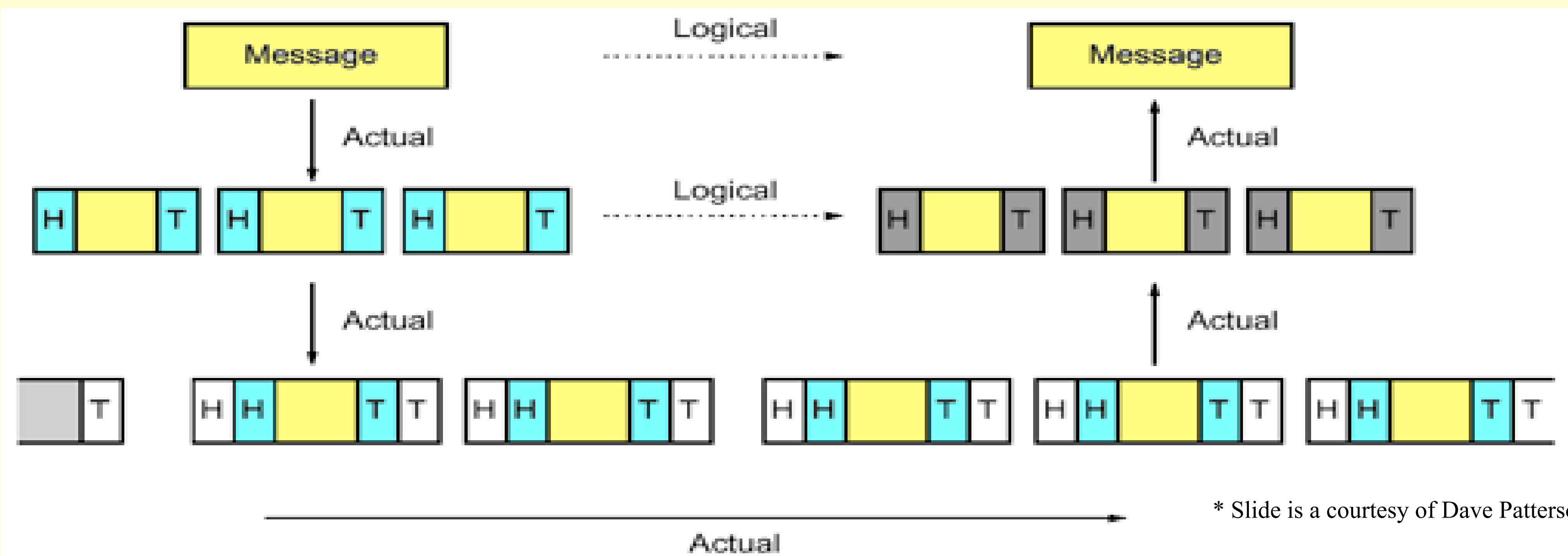


* Slide is a courtesy of Dave Patterson

Protocol Stack

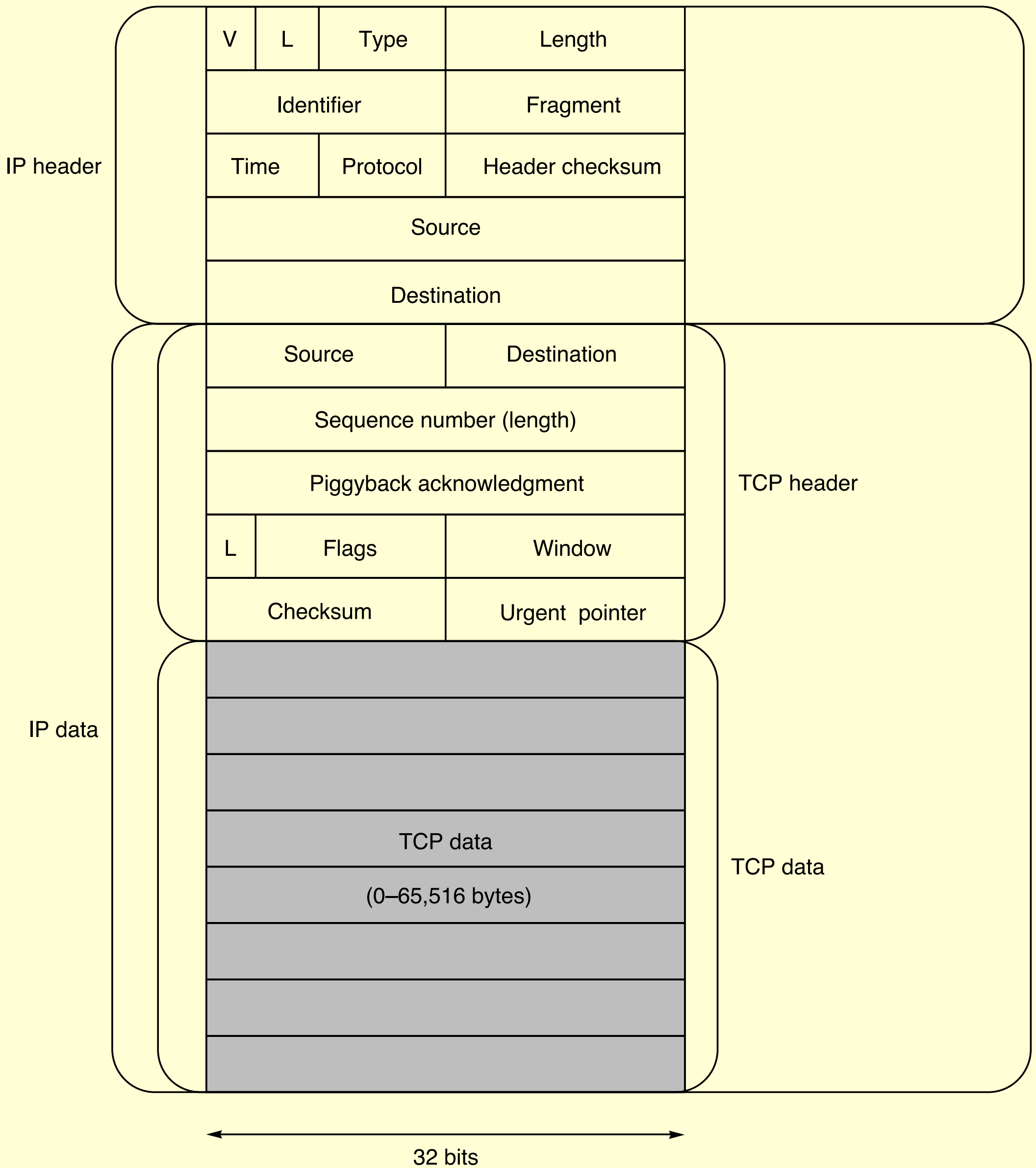
Communication occurs logically at the same level of the protocol, called peer-to-peer, but is implemented via services at the lower level

Danger is each level increases latency if implemented as hierarchy (e.g., multiple check sums)



* Slide is a courtesy of Dave Patterson

Protocol Stack



* Slide is a courtesy of Dave Patterson

OSI Layers

Open Systems Interconnect

- Application (HTTP, SMTP)
- Presentation (ntoh, hton)
- Session (Named pipes, RCP)
- Transport (TCP, UDP)
- Network (IP)
- Data Link (Ethernet)
- Physical (IEEE 802)

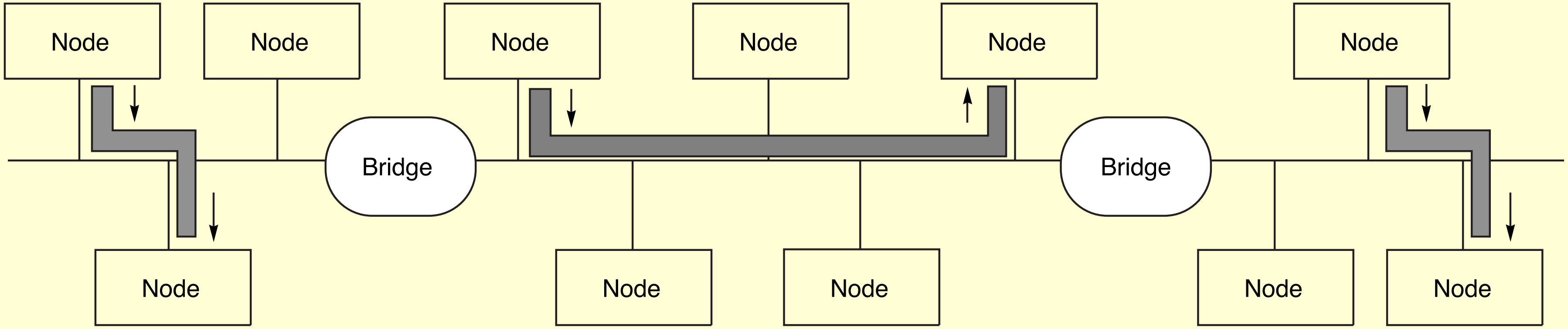
Connecting Networks

Bridges: connect LANs together, passing traffic from one side to another depending on the packet addresses

- operate at the **Ethernet protocol** level
- usually simpler and cheaper than routers

Routers or **Gateways**: connect networks 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



* Slide is a courtesy of Dave Patterson

Example Networks

	MPP IBM SP-2	LAN 100 Mb Ethernet	WAN 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
Peak Link BW	320	100	155/622
Latency (μsecs)	1	1.5	50
Send+Receive Overhead (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