Review: Storage System Issues

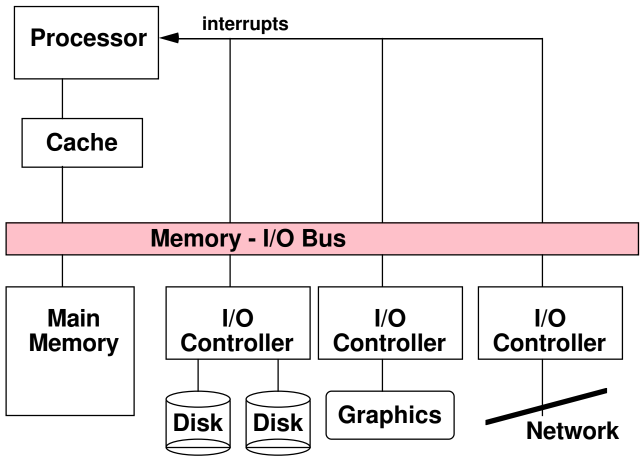
- Historical Context of Storage I/O
- Storage I/O Performance Measures
- Secondary and Tertiary Storage Devices
- A Little Queuing Theory
- Processor Interface Issues
- I/O & Memory Buses
- Show and Tell
- ABCs of UNIX File Systems
- I/O Benchmarks
- Comparing UNIX File System Performance
- RAID
- Tertiary Storage Possibilities

Review: RAID

RAID sales, ∍The Independent RAID Report May/June 1994, p. 15 (dwilmot@crl.com, 510-938-7425)

- 1993: $3.4 billion on 214,667 arrays ($15,000 / RAID)
- 1996 forecast: $11 billion
- 1997 forecast: $13 billion on 837,155 units
  - ≅ Source: DISK/TREND, 5/94 (415-961-6209)

I/O to external devices and other computers



ideal: high bandwidth, low latency

Networks

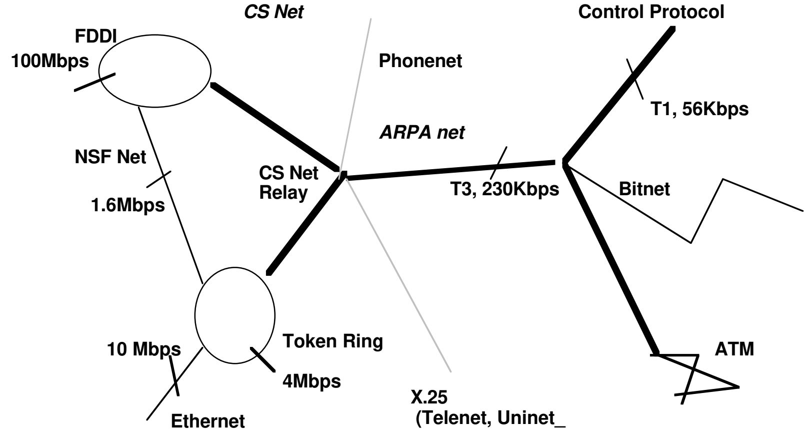
Goal: Communication between computers

Evenutal Goal: treat collection of computers as if one big computer

Theme: Different computers must agree on many things => Overriding importance of standards

Warning: Buzzword rich environment

Current Major Networks



Networks

Facets people talk a lot about

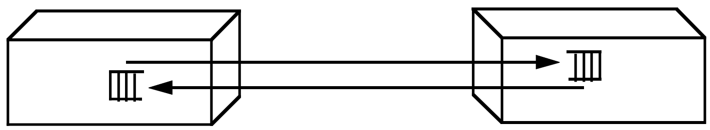
- ≅ direct vs indirect
- ≅ topology
- ≅ embeddings
- ≅ routing algorithm
- ≅ switching
- ≅ wiring

What matters

- ≅ latency
- ≅ bandwidth
- ≅ cost
- ≅ reliability

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**

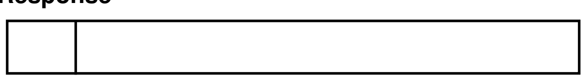
A Simple Example

What is format of packet?

- ≅ Fixed? Number bytes?

Request/
Response

Address/Data



1 bit

32 bits

0: Please send data from Address

1: Data corresponding to request

Questions about Simple Example

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?

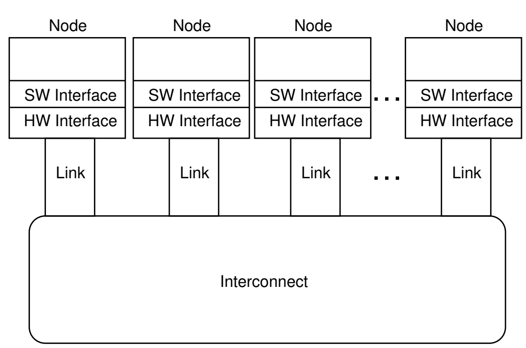
- Queue per process?

Questions such as these lead to more complex protocols and packet formats

Interconnections (Networks)

Examples

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



Interconnect Issues

- Implementation Issues
- Performance Measures
- Architecutal Issues
- Practical Issues

Implementation Issues

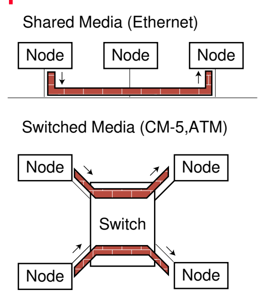
Interconnect	*MPP*	*LAN*	*WAN*
Example	CM-5	Ethernet	ATM
Maximum length between nodes	25 m	500 m; 5 repeaters	copper: 100 m optical: 1000 m
Number data lines	4	1	1
Clock Rate	40 MHz	10 MHz	B 155.5 MHz
Shared vs. Switch	Switch	Shared	Switch
Maximum number of nodes	2048	254	> 10,000
Media Material	Copper	Twisted pair copper wire or Coaxial cable	Twisted pair copper wire or optical fiber

Implementation Issues

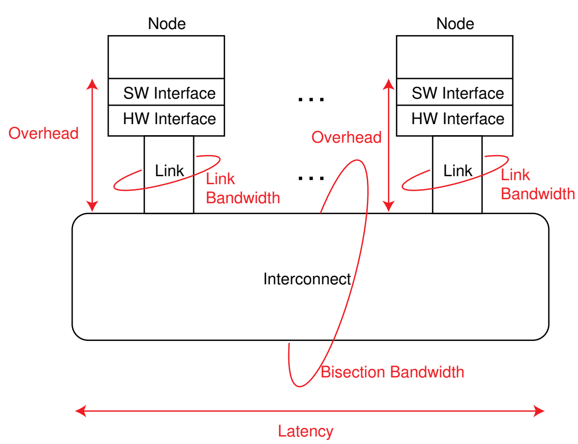
Advantages of Serial vs. Parallel lines:

- **No synchronizing signals**
- **Higher clock rate and longer distance than parallel lines. (e.g., 60 MHz x 256 bits x 0.5 m vs. 155 MHz x 1 bit x 100 m)**
    - **Imperfections in the copper wires or integrated circuit pad drivers can cause skew in the arrival of signals, limiting the clock rate, and the length and number of the parallel lines.**

Switched vs. Shared Media: pairs communicate at same time: ∍point-to-point connections



Network Performance Measures



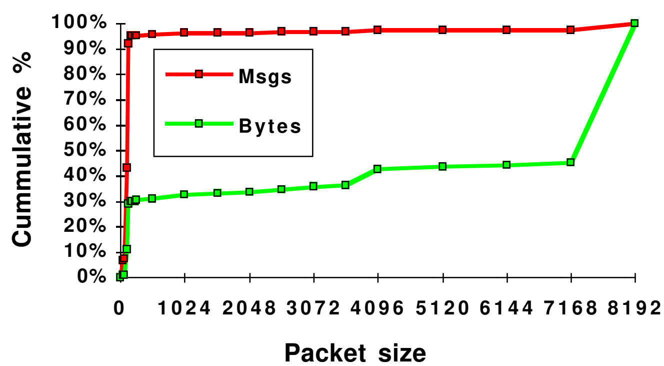
Overhead: latency of interface vs. Latency: network

Example Performance Measures

Interconnect	*MPP*	*LAN*	*WAN*
Example	CM-5	Ethernet	ATM
BisectionBW	Nx 5MB/s	1.125 MB/s	N x 10 MB/s
Int./Link BW	20 MB/s	1.125 MB/s	10 MB/s
Lantency	5 sec	15 sec	50 to 10,000 s
HW Overhead to/from	0.5/0.5 s	6/6 s	6/6 s
SW Overhead to/from	1.6/12.4 s	200/241 s	207/360 s

(TCP/IP on LAN/WAN)

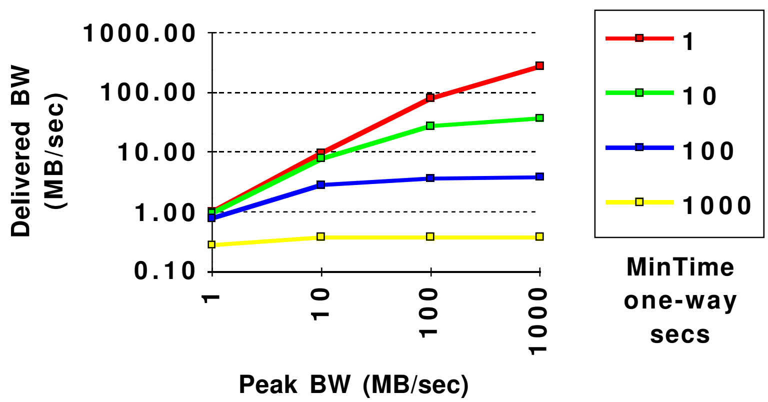
Measurement: Sizes of Message for NFS



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

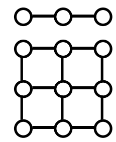
Topology

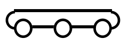
Structure of the interconnect

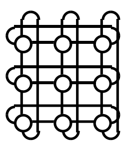
Determines

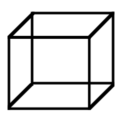
- ≅ **degree: number of links from a node**
- ≅ **diameter: max number of links crossed between nodes**
- ≅ **average distance: number of hops to random destination**
- ≅ **bisection**
  - ≅ **minimum number of links that separate the network into two halves**

Important Topologies





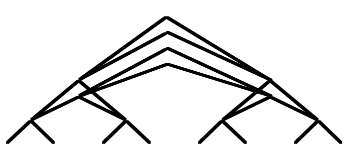


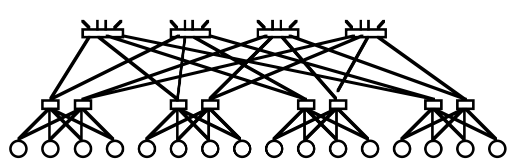


					N = 1024	
Type	Degree	Diameter	Ave Dist	Bisection	Diam	Ave D
1D mesh	2	N-1	N/3	1		
2D mesh	4	$2(N^{1/2} - 1)$	$2N^{1/2} / 3$	$N^{1/2}$	63	21
3D mesh	6	$3(N^{1/3} - 1)$	$3N^{1/3} / 3$	$N^{2/3}$	~30	~10
nD mesh	2n	$n(N^{1/n} - 1)$	$nN^{1/n} / 3$	$N^{(n-1) / n}$		
$(N = k^n)$						
Ring	2	N / 2	N/4	2		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

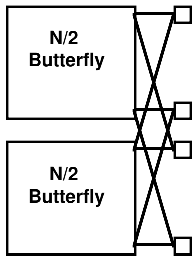
					N = 1024	
Type	Degree	Diameter	Ave Dist	Bisection	Diam	Ave D
2D Tree	3	$2Log_2 N$	$\sim 2Log_2 N$	1	20	~20
4D Tree	5	$2Log_4 N$	$2Log_4 N - 2/3$	1	10	9.33
kD	k+1	$Log_k N$				
2D fat tree	4	$Log_2 N$		N		
2D butterfly	4	$Log_2 N$		N/2	20	20

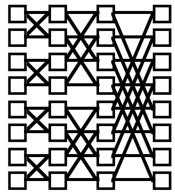




CM-5 Thinned Fat Tree

Butterfly



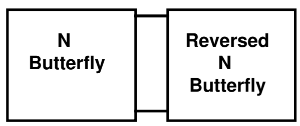


All paths equal length

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

Switching

w = wire width (flit)

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: $L = H(b + R_{sf})$

≅ cut-through: $L = HR_{ct} + b$

≅ typically, $R_{ct} = 32/W$

≅ most modern networks are about 1 small packet deep

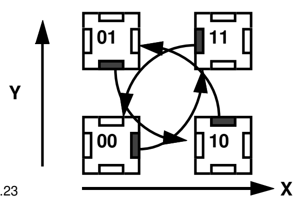
Routing

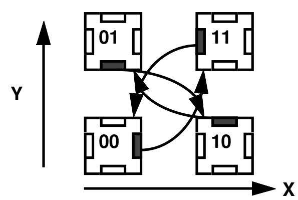
Deterministic - follows a pre-specified route

- **mesh: dimension-order routing**
  - $(x_1, y_1) \rightarrow (x_2, y_2)$
  - first $\Delta x = x_2 - x_1$,
  - then $\Delta y = y_2 - y_1$,
- **hypercube: e-cube routing**
  - $X = x_0 x_1 x_2 \ldots x_n \rightarrow Y = y_0 y_1 y_2 \ldots y_n$
  - 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





Virtual Channels

Multiplex several logical channels across each physical link

Can be used to break deadlock with adaptive routing

- **change channels on certain turns**
- **can also avoid deadlock by restricting turn combinations**

Improve link utilization

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

Generally, topologies with larger distance have better physical layout, so they have shorter, wider links and smaller routing delays. Hence, latencies are similar.

- **$L = HR_{ct} + M/w$**

What about bandwidth.

- **low degree networks have few links per processor**
- **if each message travels a large number of links**

=> few outstanding messages per processor

(similar to bisection bandwidth)

- **with skinny wires, message occupies path for a long time**

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 the alike; cost issues in topologies