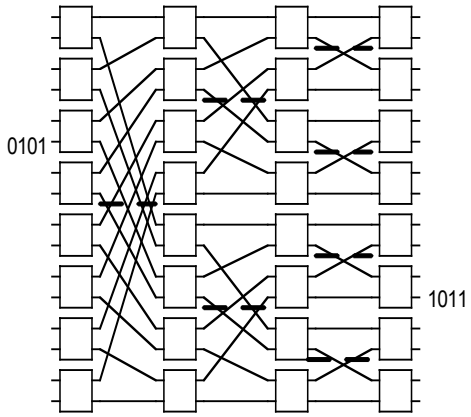
18-742

Lecture 19

Interconnection Networks



Spring 2005
Prof. Babak Falsafi
http://www.ece.cmu.edu/~ece742

Slides developed in part by Profs. Adve, Falsafi, Hill, Lebeck, Reinhardt, Smith, and Singh of University of Illinois, Carnegie Mellon University, University of Wisconsin, Duke University, University of Michigan, and Princeton University.

Readings

Chapter 10 of Culler & Singh

Reader 6:

- **W. J. Dally and C. L. Seitz, *The Torus Routing Chip*, Technical Report 5208:TR:86, California Institute of Technology, 1986.**
- **S. S. Mukherjee, P. Bannon, S. Lang, A. Spink, and D. Webb, *The Alpha 21364 Network Architecture*, IEEE Micro, January-February 2002, pp. 45-54.**

Announcements

Can Parallel Computing Finally Impact Mainstream Computing?

Friday April 1, 2005
Intel Research Pittsburgh
417 S. Craig Street, 3rd Floor
10:30 am

Uzi Vishkin
University of Maryland

Interconnection Networks

- **Goal: Communication between computers**
- **Warning: Terminology-rich environment**
- **Focus on Networks for Parallel Computing**
  - **today's System Area Networks exhibit many of the same properties**
- **Traffic Patterns**
  - **Arbitrary (bisection bandwidth matters)**
  - **Near-neighbor**
  - **Permutations (e.g., FFT)**
  - **Multicasts or Broadcasts?**
  - **Latency or Bandwidth Critical?**

Terms

Network characterized by

- **Topology**
  - **physical structure of the graph**
- **Routing Algorithm**
  - **through which paths can message flow**
- **Switching Strategy**
  - **How data in message traverses its route**
  - **Circuit Switched vs Packet Switched**
- **Flow Control**
  - **When does a packet (or portions of it) move along its route**

More Terms

- **Given topology constructed by linking switches and network interfaces, must deliver packet from node A to node B**
- **Link: cable with connectors on each end**
  - **connect switches to other switches or network interfaces**
- **Switch: N inputs N outputs (degree N)**
- **Phit: Minimum # of bits physically moved across link in one cycle**
- **Flit: Minimum # of bits move across link as a single unit**
- **Packet: unit that requires routing information, some number of flits**

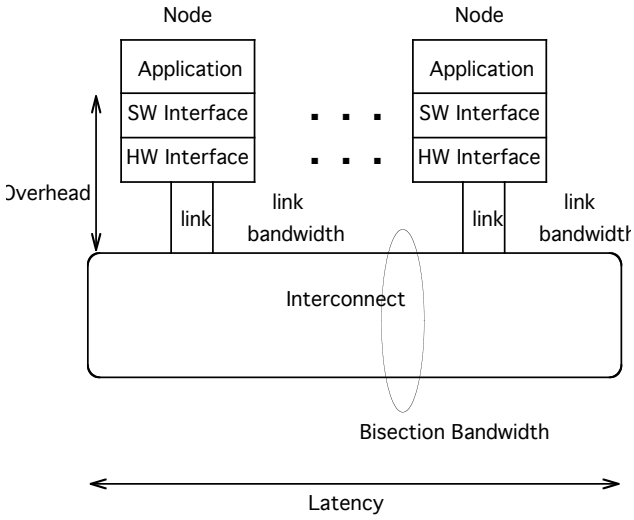
Outline

- Topology
- Switching, Routing, & Deadlock
- Switch Design
- Flow Control
- Case Studies

Topology

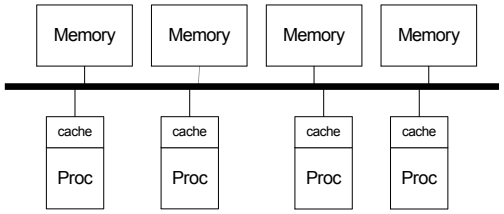
- Structure of the interconnect
- Determines
  - Switch Degree: number of links from a node
  - Diameter: number of links crossed between nodes on maximum shortest path
  - Average distance: number of hops to random destination
  - Bisection: minimum number of links that separate the network into two halves
- Don't forget Network Interface (NI)
  - "On-" and "off-ramp" to the interconnect "freeway"
  - NI latency can dominate interconnect latency (reducing the importance of topology)

Topology



Buses

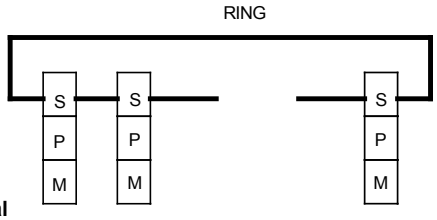
- **Direct interconnect**
- **Performance/Cost:**
  - **Switch cost: N**
  - **Wire cost: const**
  - **Avg. latency: const**
  - **Bisection B/W: const**
  - **Not neighbor optimized**
  - **May be local optimized**



- **Capable of broadcast (good for MP coherence)**
- **Bandwidth not scalable => major problem**
  - **Hierarchical buses?**
    - » **Bisection B/W remains constant**
    - » **Becomes neighbor optimized**

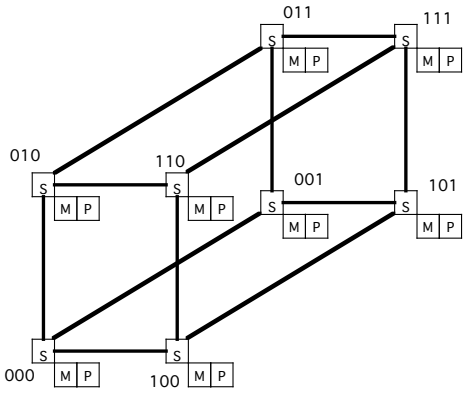
Rings

- **Direct interconnect**
- **Performance/Cost:**
  - **Switch cost: N**
  - **Wire cost: N**
  - **Avg. latency: N / 2**
  - **Bisection B/W: const**
  - **neighbor optimized, if bi-directional**
  - **probably local optimized**
- **Not easily scalable**
  - **Hierarchical rings?**
    - » **Bisection B/W remains constant**
    - » **Becomes neighbor optimized**



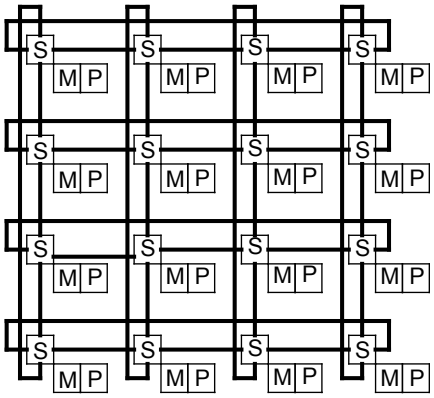
Hypercubes

- **n-dimensional unit cube**
- **Direct interconnect**
- **Performance/Cost:**
  - **Switch cost: N log2 N**
  - **Wire cost: (N log2 N) /2**
  - **Avg. latency: (log2 N) /2**
  - **Bisection B/W: N /2**
  - **neighbor optimized**
  - **probably local optimized**
- **latency and bandwidth scale well, BUT**
  - **individual switch complexity grows**
  - **=> max size is built in**



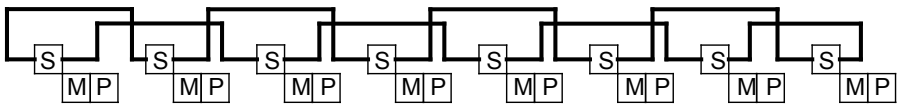
2D Torus

- **Direct interconnect**
- **Performance/Cost:**
  - **Switch cost: N**
  - **Wire cost: 2 N**
  - **Avg. latency: N 1/2**
  - **Bisection B/W: 2 N 1/2**
  - **neighbor optimized**
  - **probably local optimized**



2D Torus, cont.

- **Cost scales well**
- **Latency and bandwidth do not scale as well as hypercube, BUT**
  - **difference is relatively small for practical-sized systems**
- **Weave nodes to make inter-node latencies const.**
- **Routing can be tricky (deadlocks)**
- **2D Mesh similar, but without wraparound**



3D Torus

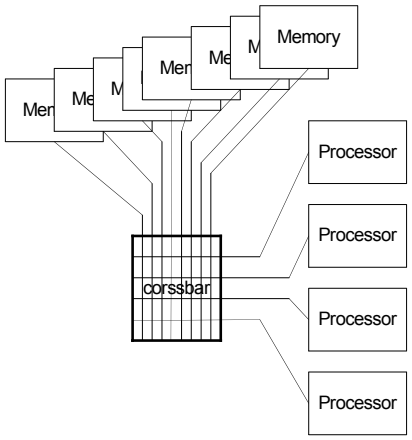
- Direct interconnect
- Performance/Cost:
  - Switch cost: N
  - Wire cost: 3 N
  - Avg. latency: $3(N^{1/3} / 2)$
  - Bisection B/W: $2 N^{2/3}$
  - neighbor optimized
  - probably local optimized

3D Torus, cont.

- Cost scales well
- Latency and bandwidth do not scale as well as hypercube, BUT
  - difference is relatively small for practical-sized systems
- Seems to have become an interconnect of choice:
  - Cray, Intel, Tera, DASH, etc.
- Routing can be tricky (deadlocks)
- 3D Mesh similar, but without wraparound

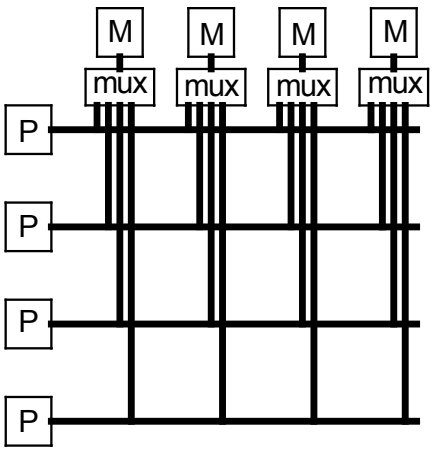
Crossbars

- **Indirect interconnect**
- **Performance/Cost:**
  - **Switch cost: N 2**
  - **Wire cost: 2 N**
  - **Avg. latency: const**
  - **Bisection B/W: N**
  - **Not neighbor optimized**
  - **Not local optimized**



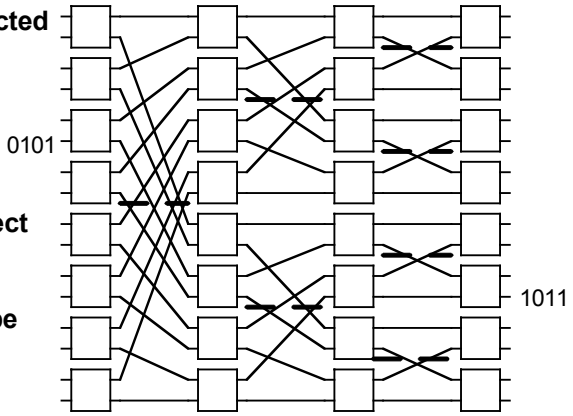
Crossbar, cont.

- **Capable of broadcast**
- **No network conflicts**
- **Cost does not scale well**
- **Often implemented with muxes**



Multistage Interconnection Networks (MINs)

- AKA: Banyan, Baseline, Omega networks, etc.
- Indirect interconnect
- Crossbars interconnected with shuffles
- Can be viewed as overlapped MUX trees
- Destination address specifies the path
- The shuffle interconnect routes addresses in a binary fashion
- This can most easily be seen with MINs in the form at right

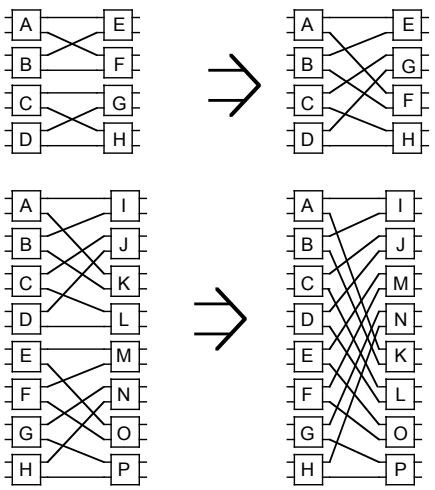


MINs, cont.

- f switch outputs => decode $\log_2$ f bits in switch
- Performance/Cost:
  - Switch cost: (N $\log_f$ N) / f
  - Wire cost: N $\log_f$ N
  - Avg. latency: $\log_f$ N
  - Bisection B/W: N
  - Not neighbor optimized
    - » Problem (in combination with latency growth)
  - Not local optimized
- Capable of broadcast
- Commonly used in large UMA systems

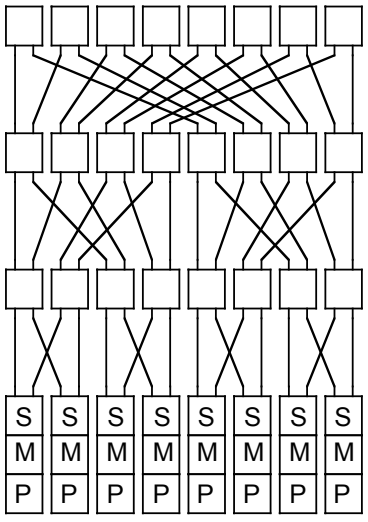
Multistage Nets, Equivalence

- **By rearranging switches, multistage nets can be shown to be equivalent**



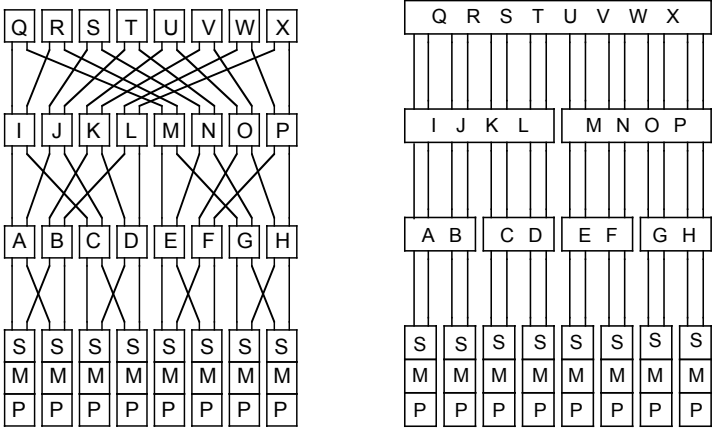
Fat Trees

- **Tree-like, with constant bandwidth at all levels**
- **Closely related to MINs**
- **Indirect interconnect**
- **Performance/Cost:**
  - **Switch cost: N logf N**
  - **Wire cost: f N logf N**
  - **Avg. latency: approx 2 logf N**
  - **Bisection B/W: f N**
  - **neighbor optimized**
  - **may be local optimized**
- **Capable of broadcast**

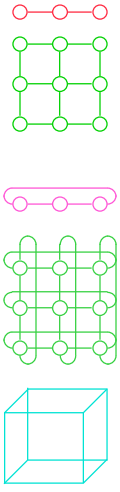


Fat Trees, cont.

- **The MIN-derived Fat Tree, is, in fact, a Fat Tree:**
- **However, the switching "nodes" in effect do not have full crossbar connectivity**



Important Topologies

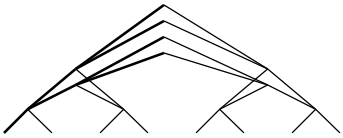


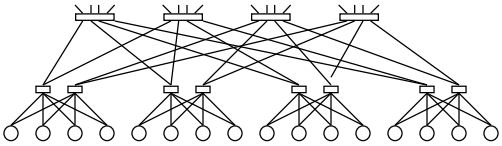
Type	Degree	Diameter	Ave Dist	Bisection	N = 1024 Diam	Ave D
1D mesh	2	N-1	2N/3	1		
2D mesh	4	$2(N^{1/2} - 1)$	$2N^{1/2} / 3$	$N^{1/2}$	63	21
3D mesh	6	$3(N^{1/3} - 1)$	$3N^{1/3} / 3$	$N^{2/3}$	~30	~10
nD mesh (N = k^n)	2n	$n(N^{1/n} - 1)$	$nN^{1/n} / 3$	$N^{(n-1)/n}$		
Ring	2	N / 2	N/4	2		
2D torus	4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

5

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	$Log_2 N$	$Log_2 N$	N/2	20	20

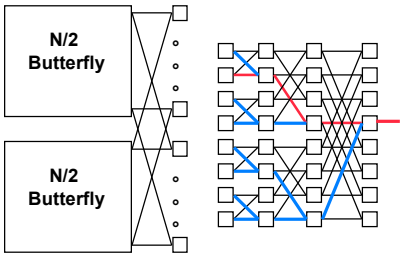




CM-5 Thinned Fat Tree

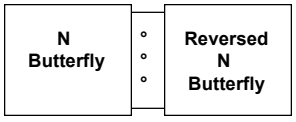
Butterfly

Multistage: nodes at ends, switches in middle



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

Benes Network



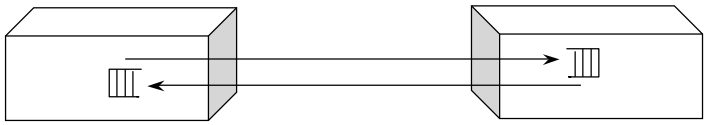
- Routes all permutations w/o conflict
- Notice similarity to Fat Tree (Fold in half)
- Randomization is major breakthrough

Outline

- Topology
- Switching, Routing, & Deadlock
- Switch Design
- Flow Control
- Case Studies

ABCs of Networks

- Starting Point: Send bits between 2 computers

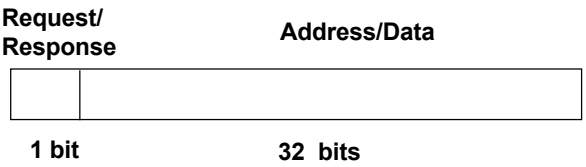


Queue on each end

- Can send both ways ("Bi-directional, Full Duplex")
- Rules for communication? "protocol"
  - Synchronous send
    - » Need Request & Response signaling
  - Name for standard group of bits sent: Packet

A Simple Example

- What is the packet format?
  - Fixed? (for HW Interpretation)
  - Number bytes?



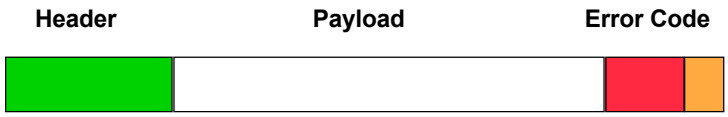
0: Please send data from Address
1: Packet contains data corresponding to request

Questions About Simple Example

- What if more than 2 computers want to communicate?
  - Need node identifier field (destination) in packet
  - Routing and topology
- 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, time outs)
- What if multiple processes/machine?
  - Dispatch
  - Queue per process
- Questions such as these lead to more complex protocols and packet formats

General Packet Format

- **Header**
  - routing and control information
- **Payload**
  - carries data (non HW specific information)
  - can be further divided (framing, protocol stacks…)
- **Error Code**
  - generally at tail of packet so it can be generated on the way out



Message vs. Packet

- A Message may be composed of several packets
- Applications reason about messages
- Network transfers packets
- Small fixed size packets. Problems?
  Fragmentation and reassembly (SW overhead)
- Variable Size packets. Problems?
  Congestion

Packet Switched vs Circuit Switched

Circuit Switched

- Establish Route then Send Data
- Telephone system

Packet Switched

- Route each packet individually
- Delivery Guarantees
  - Reliable
  - In order, what if not?