18-742

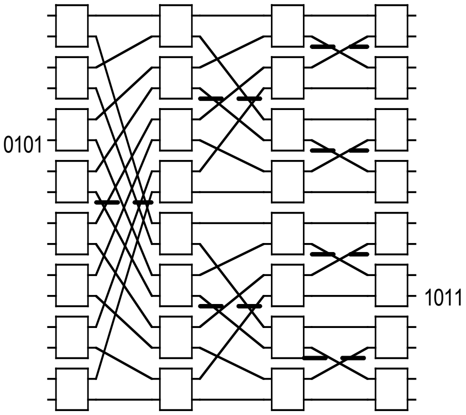
Lecture 20

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

Announcements

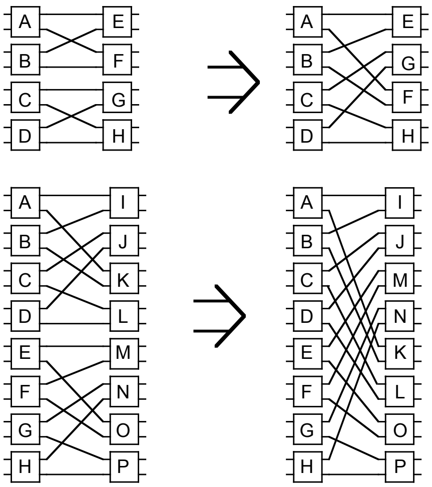
Clustered Multi-Threading: A Platform for Integrated Dynamic Thermal and Reliability Management

Tuesday April 5, 2005 in HH D-210 @ 4:00 pm

David Albonesi
Cornell University

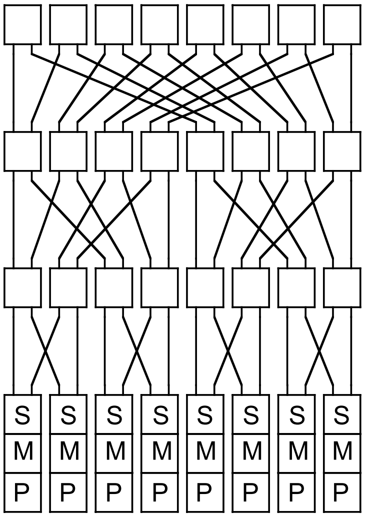
Multistage Nets, Equivalence

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



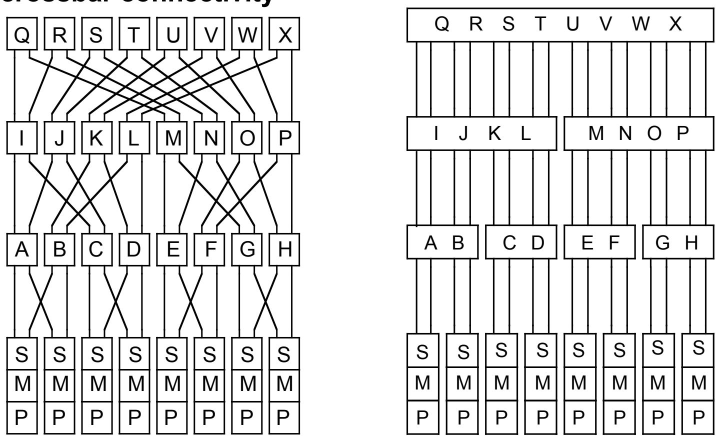
Fat Trees

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

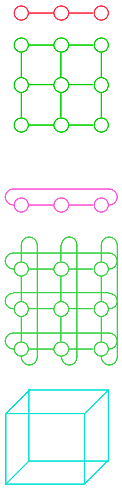


Fat Trees, cont.

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



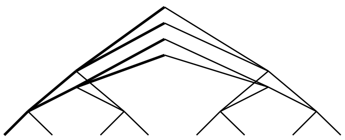
Important Topologies

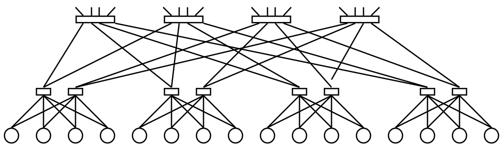


				N = 1024		
Type	Degree		Diameter	Ave Dist		
Bisection	Diam	Ave D				
1D mesh	2	N-1	2N/3	1		
2D mesh	4	$2(N^{1/2} - 1)$	$2N^{1/2} / 3$	$N^{1/2}$	63	21
3D mesh	6	$3(N^{1/3} - 1)$	$3N^{1/3} / 3$	$N^{2/3}$	~30	~10
nD mesh	2n	$n(N^{1/n} - 1)$	$nN^{1/n} / 3$	$N^{(n-1) / n}$		
$(N = k^n)$						
Ring	2	N / 2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

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	

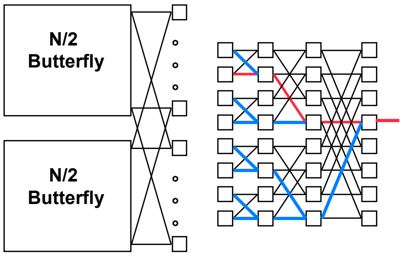




CM-5 Thinned Fat Tree

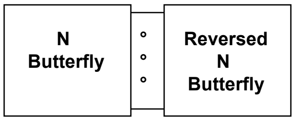
Butterfly

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



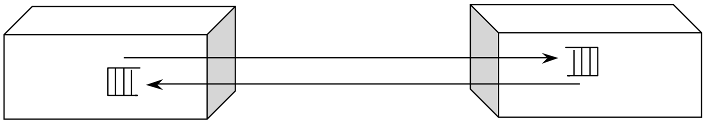
- Routes all permutations w/o conflict
- Notice similarity to Fat Tree (Fold in half)
- Randomization is major breakthrough

Outline

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

ABCs of Networks

- Starting Point: Send bits between 2 computers

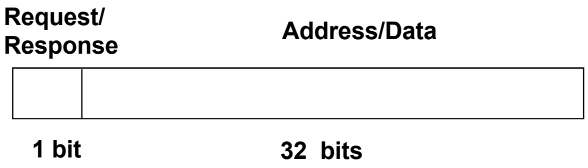


Queue on each end

- 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?



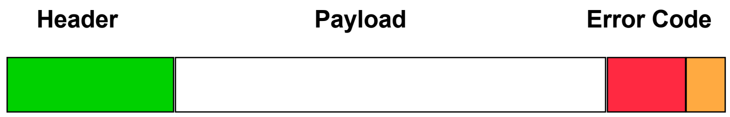
0: Please send data from Address
1: Packet contains data corresponding to request

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

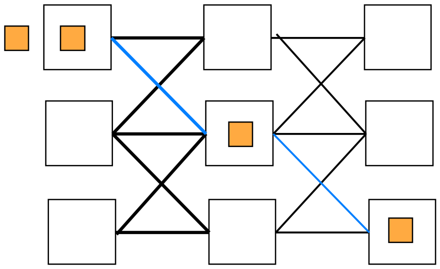
Packet Switched

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

Routing

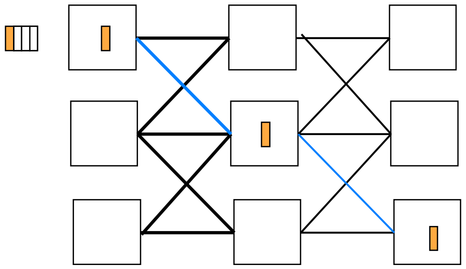
- Store-and-forward
- Cut-through
- Virtual cut-through
- Wormhole

Store and Forward



- 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



- Cut-through routing: switch examines the header, decides where to send the message, and then starts forwarding it immediately

Virtual Cut-Through

- What to do if output port is blocked?
- Lets the tail continue when the head is blocked, absorbing the whole message into a single switch.
  - Requires a buffer large enough to hold the largest packet.
- Degenerates to store-and-forward with high contention

Wormhole

- When the head of the message is blocked the message stays strung out over the network
  - Potentially blocks other messages (needs only buffer the piece of the packet that is sent between switches).
  - CM-5 used it, with each switch buffer being 4 bits per port.
  - Myrinet uses it
- Interaction with Pack Size
- Can cause tree saturation…

Store and Forward vs. Cut-Through

- Advantage
  - Latency reduces from function of:

    number of intermediate switches times the size of the packet

    to

    time for 1st part of the packet to negotiate the switches + the packet size ÷ interconnect BW

Routing Algorithm

- How do I know where a packet should go?
- Arithmetic
- Source-Based
- Table Lookup
- Adaptive—route based on network state (e.g., contention)

Arithmetic Routing

- For regular topology, simple arithmetic to determine route
- 2D Mesh (Also called NEWS network)
  - packet header contains signed offset to destination
  - switch ++ or -- one field of header (x or y dimension)
  - when x == 0 and y == 0, then at correct processor
- Requires ALU in switch
- Must recompute CRC

Source Based and Table Lookup Routing

Source Based

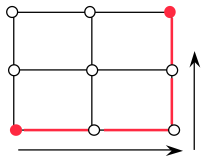
- Source specifies output port for each switch in route
- Very Simple Switches
  - no control state
  - strip output port off header
- Myrinet uses this

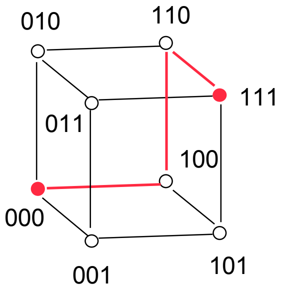
Table Lookup

- Very Small Header, index into table for output port
- Big tables, must be kept up to date...

Deterministic v.s. Adaptive Routing

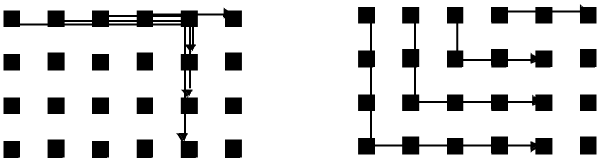
- Deterministic—follows a pre-specified route
  - mesh: dimension-order routing
    - » (x1, y1) -> (x2, y2)
    - » first Dx = x2 - x1,
    - » then Dy = y2 - y1,
  - hypercube: edge-cube routing
    - » X = x0x1x2 . . .xn -> Y = y0y1y2 . . .yn
    - » R = X xor Y
    - » Traverse dimensions of differing address in order
  - tree: common ancestor
- Adaptive—route determined by contention for output port





Adaptive Routing

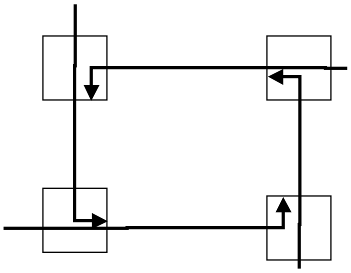
- Essential for fault tolerance
  - at least multipath
- Can improve utilization of the network
- Simple deterministic algorithms easily run into bad permutations



- Fully/partially adaptive, minimal/non-minimal
- Can introduce complexity or anomalies
- Little adaptation goes a long way!

Deadlock

- Break a Necessary Condition
  - Use more than one resource
  - Not willing to release resource in use
  - Cycle in order of recourse use

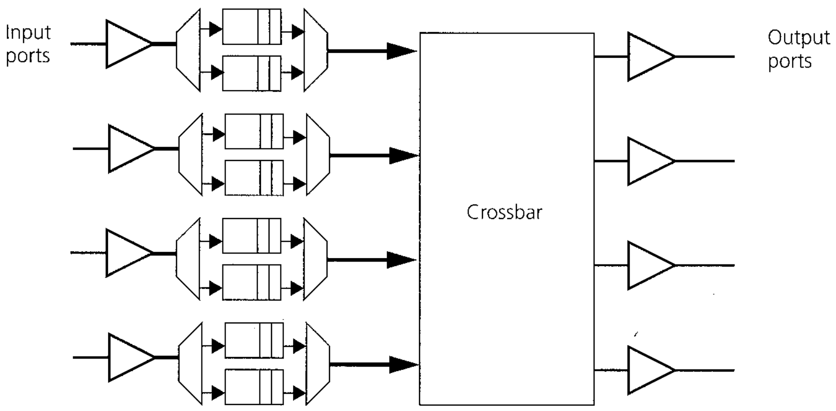


Deadlock Free Routing

- Virtual Channels
  - Not virtual cut-through
  - Add buffers so flits of wormhole packets can be interleaved
- Up*-Down*
  - Number switches: higher = farther away from processors
  - route up, make one turn, route down
- Turn Model Routing
  - Restrict order of turns
    - » West First
    - » North Last
    - » Negative First
  - Can increase number of hops

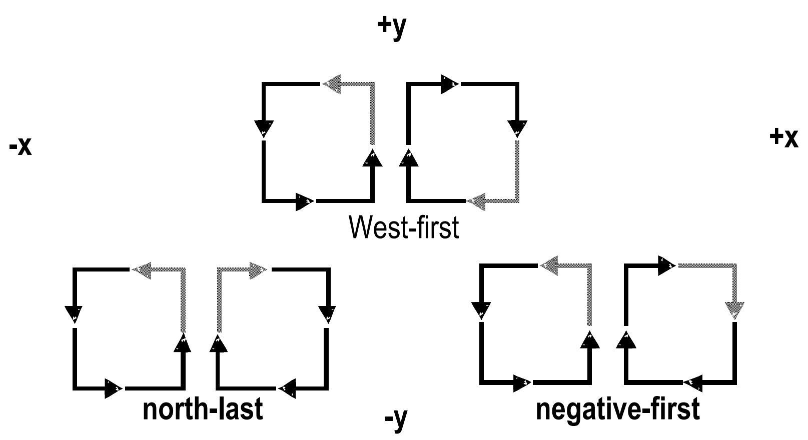
(C) 2005 Babak Falsafi from Adve, Falsafi,
Hill, Lebeck, Reinhardt, Smith & Singh

Virtual Channels

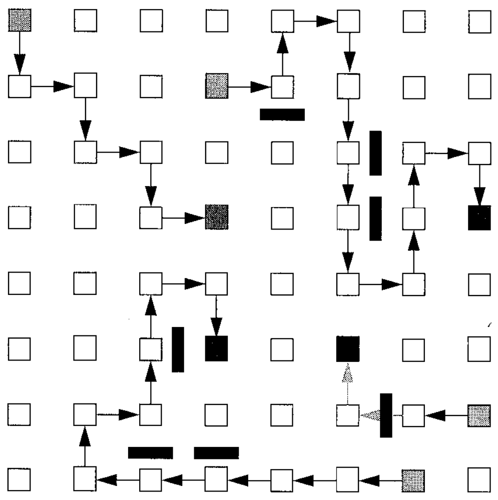


(C) 2005 Babak Falsafi from Adve, Falsafi,
Hill, Lebeck, Reinhardt, Smith & Singh

Minimal turn restrictions in 2D



West-First Routes

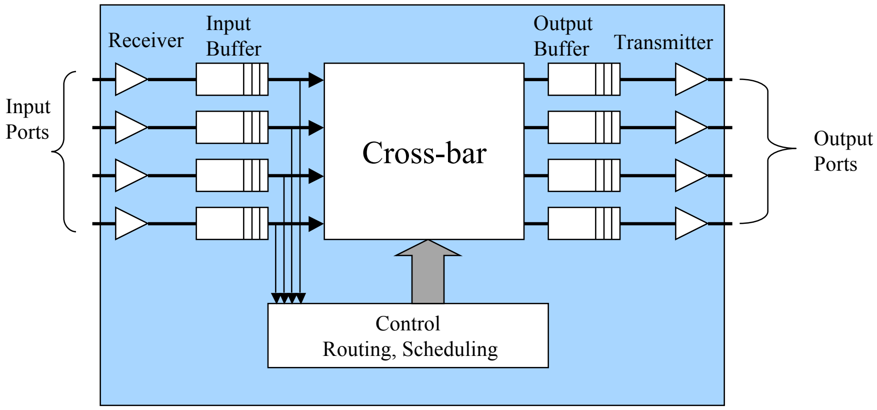


Outline

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

A Generic Switch

- At minimum, must route inputs to outputs



VLSI makes it easier to create larger fully connected switches

Switch Design Issues

- ports, pin limited
- data path (cross bar designs)
- non-blocking crossbar
- routing logic per input
  - ALU
  - table
  - Finite State Machine for cut-through

Switch Buffering

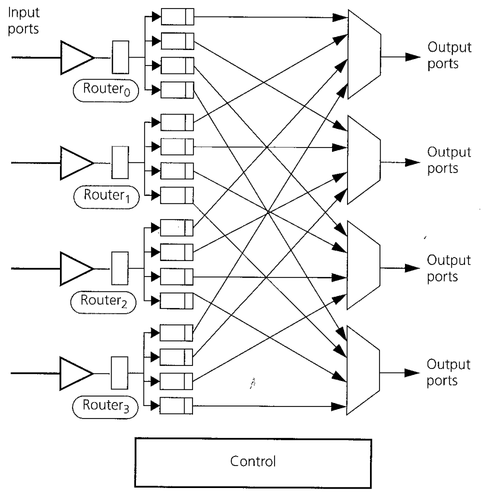
- need to absorb some transient peaks
- shared buffer pool
  - need high bandwidth
  - one output could hog all buffer space
- input buffers

Input Buffering

- buffer per port
- routing logic
- head of line (HOL) blocking
  - subsequent packet may be routed to unused output port

(C) 2005 Babak Falsafi from Adve, Falsafi,
Hill, Lebeck, Reinhardt, Smith & Singh

Avoiding HOL Blocking



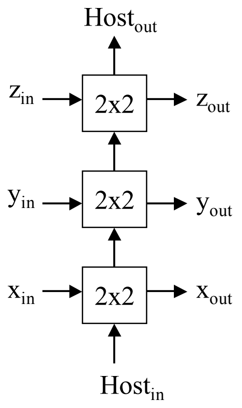
(C) 2005 Babak Falsafi from Adve, Falsafi,
Hill, Lebeck, Reinhardt, Smith & Singh

Output Buffering

- Buffers “logically” associated with output
  - split on either side of cross bar
- Arbitration for physical link (output scheduling)
  - static priority
  - random
  - round-robin
  - oldest-first
- Effects of adaptive routing?
  - Select output based on availability
  - requires feedback from output port

Stacked Dimension Switch

- Uses only 2x2 switch to build higher dimension switch


Host_out
z_in
2x2
z_out
y_in
2x2
y_out
x_in
2x2
x_out
Host_in