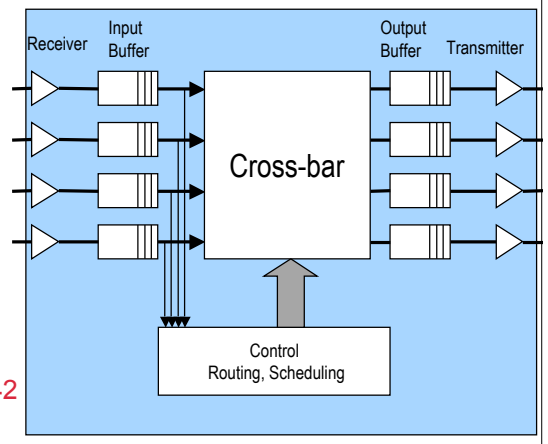
18-742

Lecture 21

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 7

- **L. Hammond, B. Hubbert, M. Siu, M. Prabhu, M. Chen, and K. Olukotun, *The Stanford Hydra CMP*, IEEE Micro, March-April 2000, pp. 71-84.**
- **L. A. Barroso, K. Gharachorloo, R. McNamara, A. Nowatzyk, S. Qadeer, B. Sano, S. Smith, R. Stets, and B. Verghese, *Piranha: a scalable architecture based on single-chip multiprocessing*, ISCA 2000.**

Announcement

Network-on-Chip - towards communication-centric System-on-Chip design

TODAY in HH D-210 @ 4:30 pm

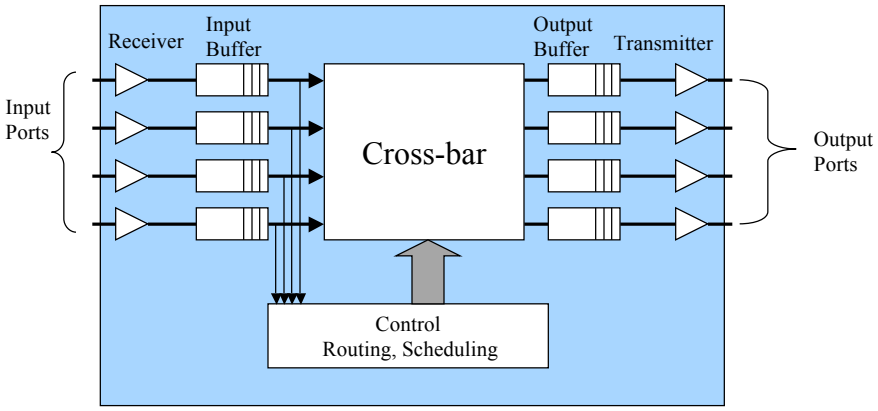
Tobias Bjerregaard
Technical University of Denmark

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**

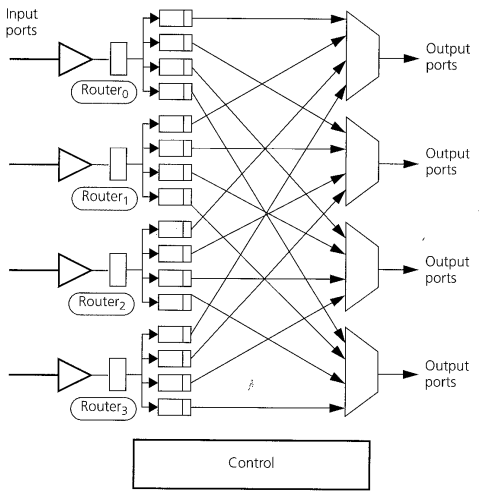
Switch Buffering

- 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

Avoiding HOL Blocking

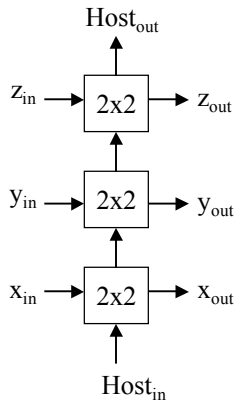


Output Buffering

- **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**



Outline

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

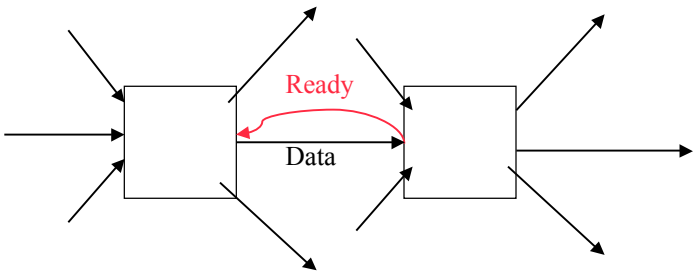
Congestion Control

- 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:
  - End-to-end Flow Control
  - Link-level Flow Control

Flow Control

- Packet discarding: If a packet arrives at a switch and there is no room in the buffer, the packet is discarded
  - no communication between switches, requires higher level protocol
- Flow control: between pairs of receivers and senders; use feedback to tell the sender when it is allowed to send the next packet

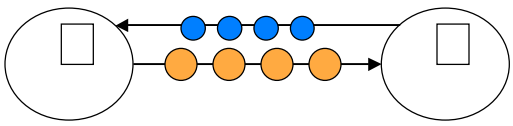
Link-Level Flow Control



- Transfer single flit when receiver is ready
- Could have long links with many flits in flight

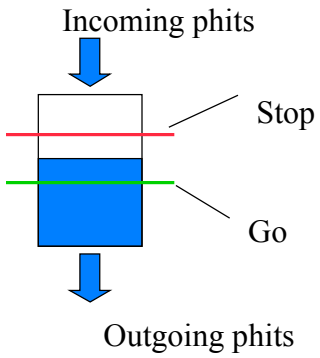
Credit-based (Window) Flow Control

- **Receiver gives N credits to sender**
  - **sender decrements count**
  - **stops sending if zero**
  - **receiver sends back credit as it drains its buffer**
  - **bundle credits to reduce overhead**
- **Must account for link latency**



Water Level

- high water, low water
- stop & go back to source switch (Myrinet)
- can send redundant stop/go



Outline

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

Case Study Cray T3D

- 1024 switch nodes each connected to 2 processors
- 3D Torus, bidirectional, 300 MB/s
- Link: 16 bits, 8 control bits
- Variable size packet (multiple of 16 bits)
- Logical request & response networks
  - 2 virtual channels each for deadlock
- Stacked dimension routing
- Wormhole for large packets, virtual cut-through for small packets

IBM SP-2 (Vulcan)

- Switch has eight bidirectional 40 MB/s links
- Link: 8 data bits, 1 tag, 1 reverse flow-control
- Flit is 16 bits, phit is 8
- input FIFO + output FIFO + central buffer 128 8-byte segments

SGI Origin (Spider)

High-end networking

4.8 GB/s:

- **Across chips within chasis**
- **Across chasis up to 5m apart**

Data link layer:

- **128 bits of data + 8 bits of sideband (flow control info)**
- **Error-free in 16 byte quantities**
- **A go-back-n sliding window for recovery**
- **Retransmission upon error**

Spider: Message Layer & Format

256-byte buffer per virtual channel

Header micropacket:

- **23 bits for destination & routing**
  - **9-bit destination identifier**
  - **4-bit exit port (on the next chip)**
  - **2-bit congestion control info**
  - **8-bit message "age" info for arbitration**

Spider: Routing

Distributed routing tables

- **Static routing (per given table info)**
- **Programmable tables (in case routing must change)**
- **Hierarchical to save space**

Topology:

- **Hierarchical fat hypercube**
- **Constant bisection bandwidth of 800 MB/s**

Alpha 21364 Network

Integrated network interface & hw

22.4 GB/s router bandwidth: 1.2 GHz router clock

128-processor DSM-ready

Topology: 2D torus

39-bit flits:

- **32-bit payload (data)**
- **7-bit CRC**

Packet sizes:

- **1, 2, 3 flits (control)**
- **18, and 19 flits (for 64-byte cache blocks)**

Alpha: Routing

Virtual cut-through:

- **Buffer space of up to 316 packets**

Adaptive routing:

- **Minimum rectangle**
- **With start/end points as diagonally opposite vertices**

Deadlock avoidance: hierarchical virtual channels

- **Coherence protocol -> protocol-specific VCs with ordering assignments**
- **Adaptive routing -> VCs further subdivided into dimension-specific VCs to break depenendences**

Programmable router tables

Real Machines

Machine	Topology	Cycle Time (ns)	Channel Width (bits)	Routing Delay (cycles)	Flit (data bits)
nCUBE/2	Hypercube	25	1	40	32
TMC CM-5	Fat-Tree	25	4	10	4
IBM SP-2	Banyan	25	8	5	16
Intel Paragon	2D Mesh	11.5	16	2	16
Meiko CS-2	Fat-Tree	20	8	7	8
CRAY T3D	3D Torus	6.67	16	2	16
DASH	Torus	30	16	2	16
J-Machine	3D Mesh	31	8	2	8
Monsoon	Butterfly	20	16	2	16
SGI Origin	Hypercube	2.5	20	16	160
Myricom	Arbitrary	6.25	16	50	16

- **Wide links, smaller routing delay**
- **Tremendous variation**