

Introduction to Scalable Interconnection Network Design

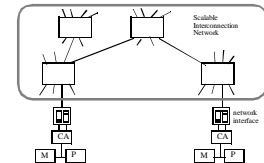
CS 258, Spring 99
David E. Culler
Computer Science Division
U.C. Berkeley

Scalable, High Perf. Interconnection Network

- At Core of Parallel Computer Arch.
- Requirements and trade-offs at many levels
 - Elegant mathematical structure
 - Deep relationships to algorithm structure
 - Managing many traffic flows
 - Electrical / Optical link properties

- Little consensus
 - interactions across levels
 - Performance metrics?
 - Cost metrics?
 - Workload?

=> need holistic understanding



3/19/99

CS258 S99

2

Requirements from Above

- Communication-to-computation ratio
 - => bandwidth that must be sustained for given computational rate
 - traffic localized or dispersed?
 - bursty or uniform?
 - Programming Model
 - protocol
 - granularity of transfer
 - degree of overlap (slackness)
- => job of a parallel machine network is to transfer information from source node to dest. node in support of network transactions that realize the programming model

3/19/99

CS258 S99

3

Goals

- latency as small as possible
- as many concurrent transfers as possible
 - operation bandwidth
 - data bandwidth
- cost as low as possible

3/19/99

CS258 S99

4

Outline

- Introduction
- Basic concepts, definitions, performance perspective
- Organizational structure
- Topologies

3/19/99

CS258 S99

5

Basic Definitions

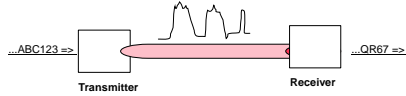
- Network interface
- Links
 - bundle of wires or fibers that carries a signal
- Switches
 - connects fixed number of input channels to fixed number of output channels

3/19/99

CS258 S99

6

Links and Channels



- **transmitter** converts stream of digital symbols into signal that is driven down the link
- **receiver** converts it back
 - tran/rcv share *physical protocol*
- trans + link + rcv form **Channel** for digital info flow between switches
- **link-level protocol** segments stream of symbols into larger units: packets or messages (**framing**)
- **node-level protocol** embeds commands for dest communication assist within packet

3/19/99

CS258 S99

7

Formalism

- network is a graph $V = \{\text{switches and nodes}\}$ connected by communication channels $C \subseteq V \times V$
- Channel has width w and signaling rate $f = 1/t$
 - channel bandwidth $b = wf$
 - phit (physical unit) data transferred per cycle
 - flit - basic unit of flow-control
- Number of input (output) channels is switch **degree**
- Sequence of switches and links followed by a message is a **route**
- Think streets and intersections

3/19/99

CS258 S99

8

What characterizes a network?

- **Topology** (what)
 - physical interconnection structure of the network graph
 - direct: node connected to every switch
 - indirect: nodes connected to specific subset of switches
- **Routing Algorithm** (which)
 - restricts the set of paths that msgs may follow
 - many algorithms with different properties
 - » gridlock avoidance?
- **Switching Strategy** (how)
 - how data in a msg traverses a route
 - circuit switching vs. packet switching
- **Flow Control Mechanism** (when)
 - when a msg or portions of it traverse a route
 - what happens when traffic is encountered?

3/19/99

CS258 S99

9

What determines performance

- Interplay of all of these aspects of the design

3/19/99

CS258 S99

10

Topological Properties

- **Routing Distance** - number of links on route
- **Diameter** - maximum routing distance
- **Average Distance**
- A network is **partitioned** by a set of links if their removal disconnects the graph

3/19/99

CS258 S99

11

Typical Packet Format



Sequence of symbols transmitted over a channel

- Two basic mechanisms for abstraction
 - encapsulation
 - fragmentation

3/19/99

CS258 S99

12

Communication Perf: Latency

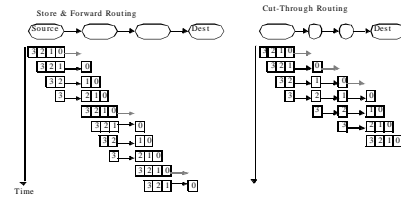
- $\text{Time}(n)_{s-d} = \text{overhead} + \text{routing delay} + \text{channel occupancy} + \text{contention delay}$
- $\text{occupancy} = (n + n_e) / b$
- Routing delay?
- Contention?

3/19/99

CS258 S99

13

Store&Forward vs Cut-Through Routing



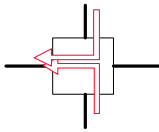
- $h(n/b + D)$ vs $n/b + h D$
- what if message is fragmented?
- wormhole vs virtual cut-through

3/19/99

CS258 S99

14

Contention



- Two packets trying to use the same link at same time
 - limited buffering
 - drop?
- Most parallel mach. networks block in place
 - link-level flow control
 - tree saturation
- Closed system - offered load depends on delivered

3/19/99

CS258 S99

15

Bandwidth

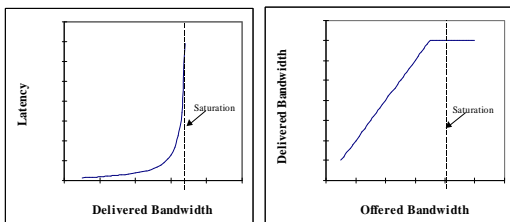
- What affects local bandwidth?
 - packet density $b \times n / (n + n_e)$
 - routing delay $b \times n / (n + n_e + w\Delta)$
 - contention
 - » endpoints
 - » within the network
- Aggregate bandwidth
 - bisection bandwidth
 - » sum of bandwidth of smallest set of links that partition the network
 - total bandwidth of all the channels: C_b
 - suppose N hosts issue packet every M cycles with ave dist
 - » each msg occupies h channels for $t = n/w$ cycles each
 - » C/N channels available per node
 - » link utilization $\rho = MC/Nht$

3/19/99

CS258 S99

16

Saturation



3/19/99

CS258 S99

17

Organizational Structure

- Processors
 - datapath + control logic
 - control logic determined by examining register transfers in the datapath
- Networks
 - links
 - switches
 - network interfaces

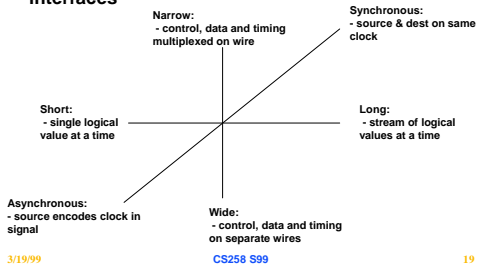
3/19/99

CS258 S99

18

Link Design/Engineering Space

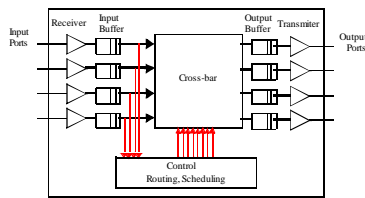
- Cable of one or more wires/fibers with connectors at the ends attached to switches or interfaces



Example: Cray MPPs

- **T3D: Short, Wide, Synchronous (300 MB/s)**
 - 24 bits
 - » 16 data, 4 control, 4 reverse direction flow control
 - single 150 MHz clock (including processor)
 - flit = phit = 16 bits
 - two control bits identify flit type (idle and framing)
 - » no-info, routing tag, packet, end-of-packet
- **T3E: long, wide, asynchronous (500 MB/s)**
 - 14 bits, 375 MHz, LVDS
 - flit = 5 phits = 70 bits
 - » 64 bits data + 6 control
 - switches operate at 75 MHz
 - framed into 1-word and 8-word read/write request packets
- **Cost = f(length, width) ?**

Switches



Switch Components

- **Output ports**
 - transmitter (typically drives clock and data)
- **Input ports**
 - synchronizer aligns data signal with local clock domain
 - essentially FIFO buffer
- **Crossbar**
 - connects each input to any output
 - degree limited by area or pinout
- **Buffering**
- **Control logic**
 - complexity depends on routing logic and scheduling algorithm
 - determine output port for each incoming packet
 - arbitrate among inputs directed at same output

Interconnection Topologies

- Class networks scaling with N
- **Logical Properties:**
 - distance, degree
- **Physical properties**
 - length, width
- **Fully connected network**
 - diameter = 1
 - degree = N
 - cost?
 - » bus => $O(N)$, but BW is $O(1)$ - actually worse
 - » crossbar => $O(N^2)$ for BW $O(N)$
- **VLSI technology determines switch degree**

Summary

Topology	Degree	Diameter	Ave Dist	Bisection	D (D ave) @ P=1024
1D Array	2	N-1	N / 3	1	huge
1D Ring	2	N/2	N/4	2	
2D Mesh	4	$2(N^{1/2} - 1)$	$2/3 N^{1/2}$	$N^{1/2}$	63 (21)
2D Torus	4	$N^{1/2}$	$1/2 N^{1/2}$	$2N^{1/2}$	32 (16)
k-ary n-cube	2n	nk/2	nk/4	nk/4	15 (7.5) @n=3
Hypercube	n = log N	n	n	n/2	N/2 10 (5)