EE382C
Lecture 1

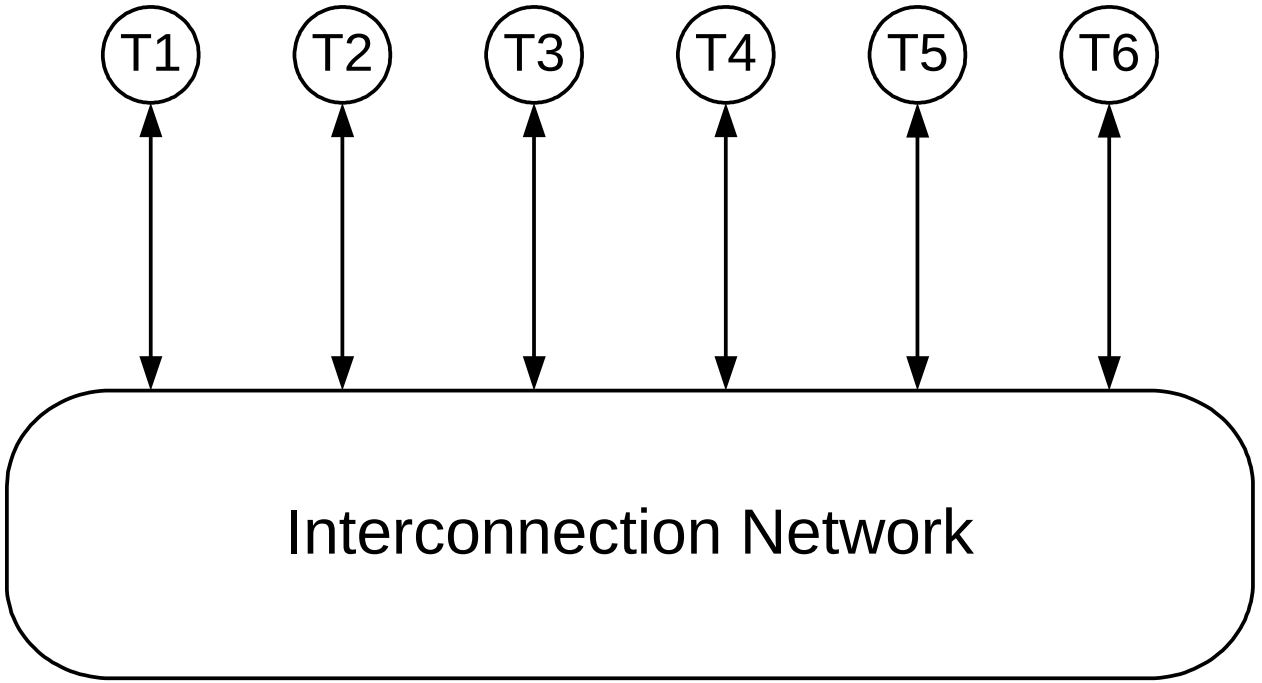
Bill Dally

3/29/11

Logistics

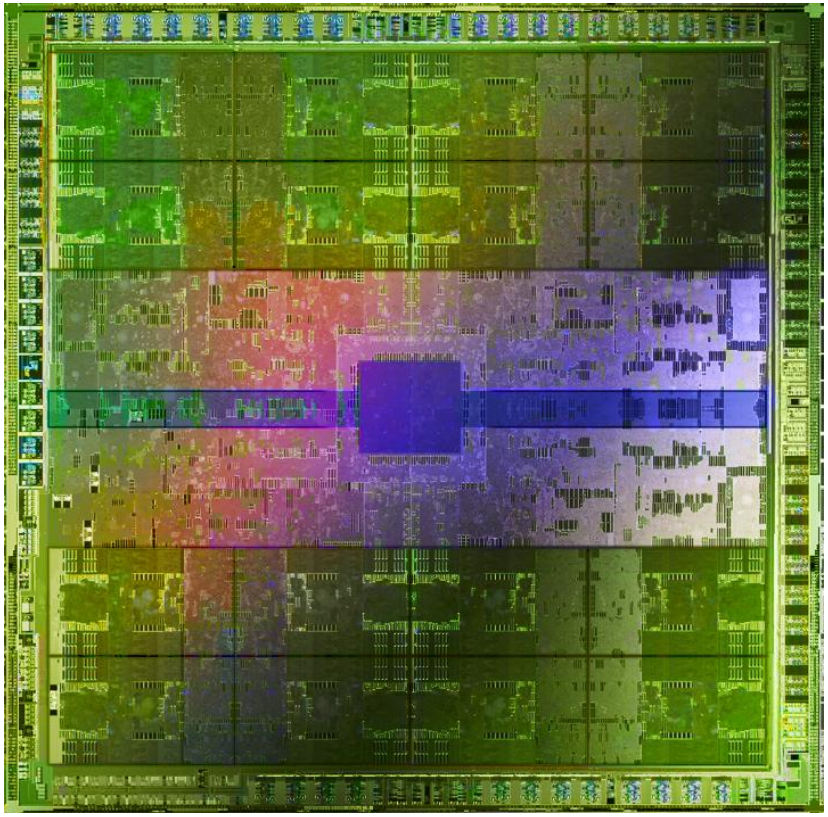
- Handouts
  - Course policy sheet
  - Course schedule
- Assignments
  - Homework
  - Research Paper
  - Project
- Midterm

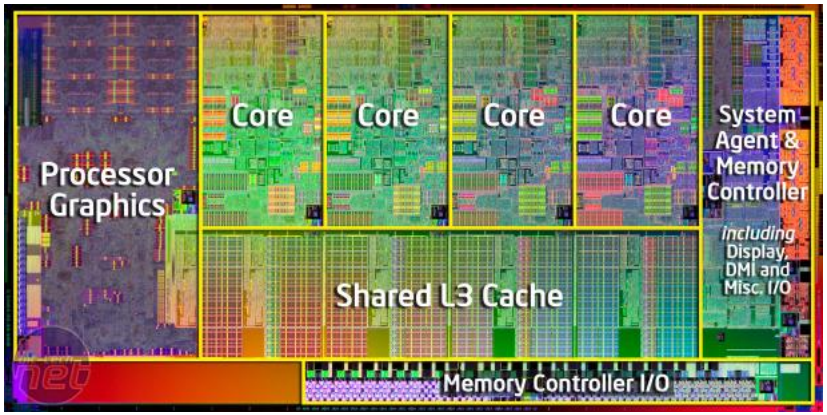
What is an interconnection network?



Where do you find interconnection networks?

Network on Chip for Many-Core Processors

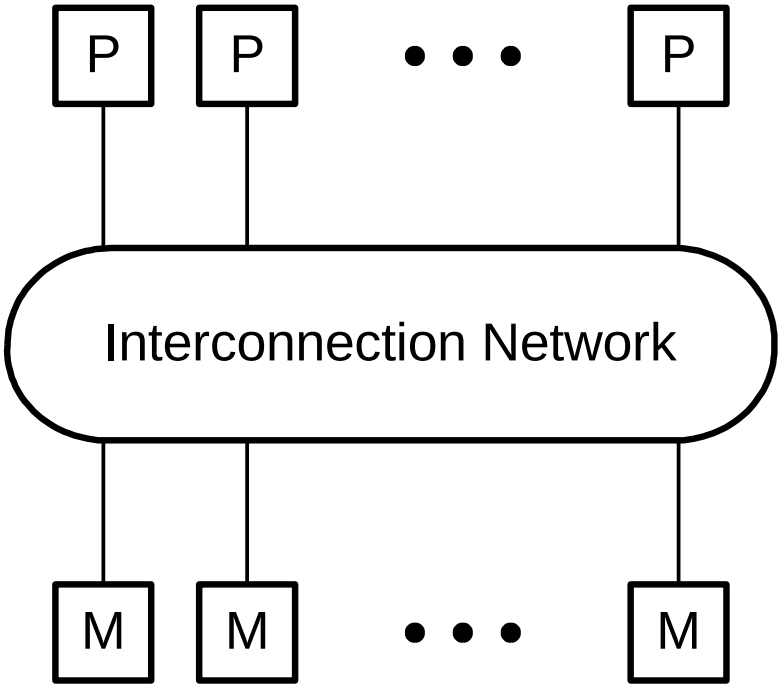


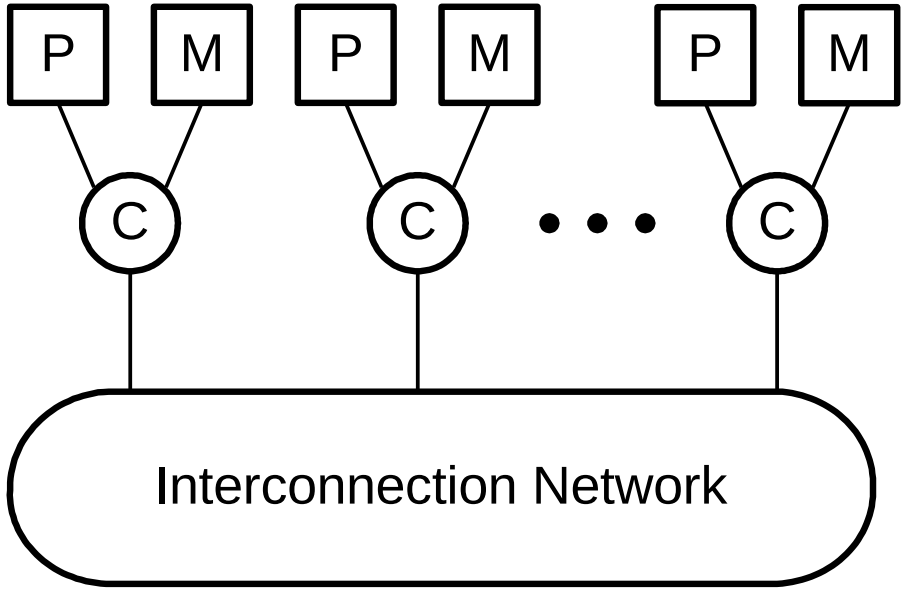


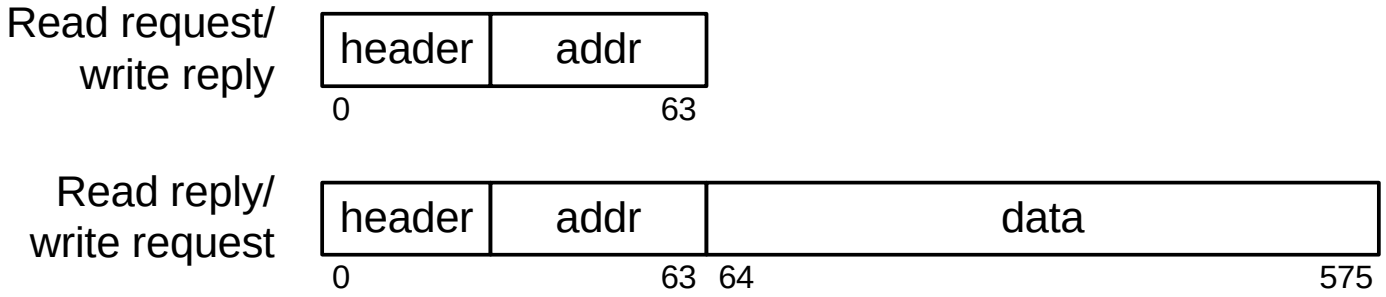
Processor-Memory and Processor-Processor Interconnect for Supercomputers

Data Center Networks
RPC, Map-Reduce, BigFile, …

Processor-Memory Interconnect



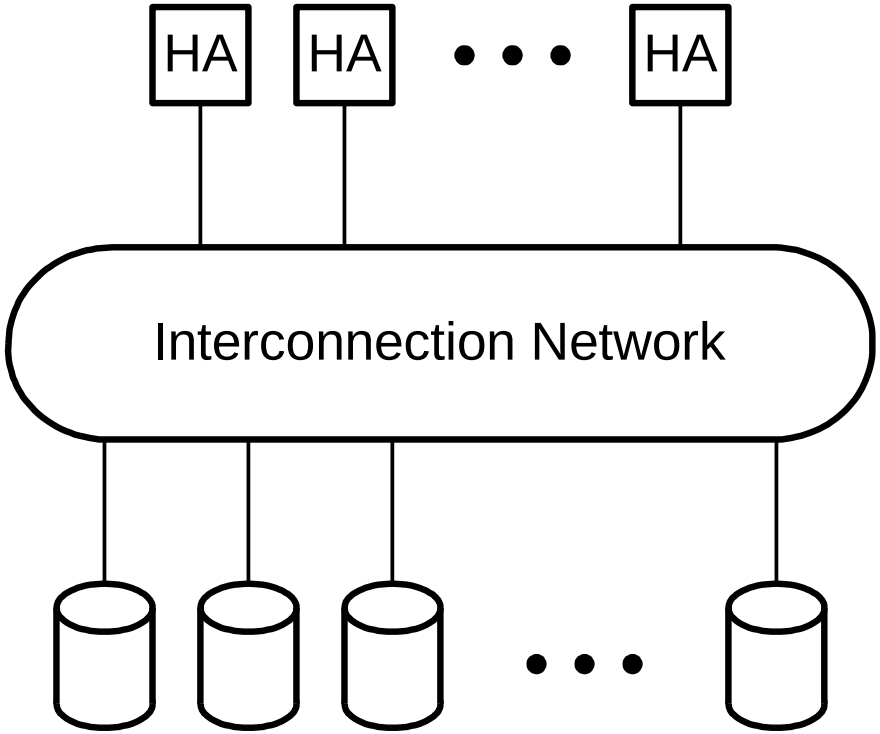




Specifications for P-M Interconnect

Parameter	Value
Processor Ports	1-2,048
Memory Ports	0-4,096
Peak Bandwidth	8GBytes/s
Average Bandwidth	400MBytes/s
Message Latency	100ns
Message Size	64 or 576 bits
Traffic Patterns	arbitrary
Quality of Service	none
Reliability	no message loss
Availability	0.999 to 0.99999

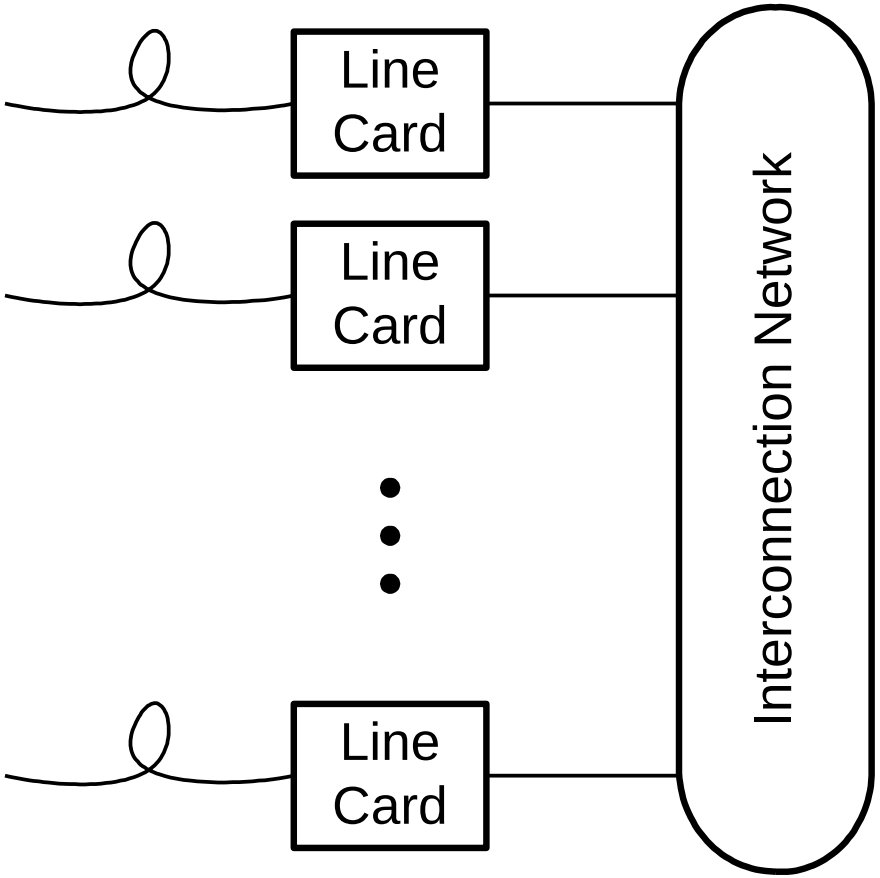
I/O Interconnect



Specification for I/O Interconnect

Parameter	Value
Device Ports	1-4,096
Host Ports	1-64
Peak Bandwidth	200MBytes/s
Average Bandwidth	1MBytes/s (devices) 64MBytes/s (hosts)
Message Latency	$10\mu s$
Message Size	32Bytes or 4KBytes
Traffic Patterns	arbitrary
Reliability	no message loss[a]
Availability	0.999 to 0.99999

Switch/Router Fabric



Specification for Switch Fabric

Parameter	Value
Ports	4–512
Peak Bandwidth	10Gb/s
Average Bandwidth	7Gb/s
Message Latency	10μs
Packet Payload Size	40–64KBytes
Traffic Patterns	arbitrary
Reliability	$< 10^{-15}$ loss rate
Quality of Service	needed
Availability	0.999 to 0.99999

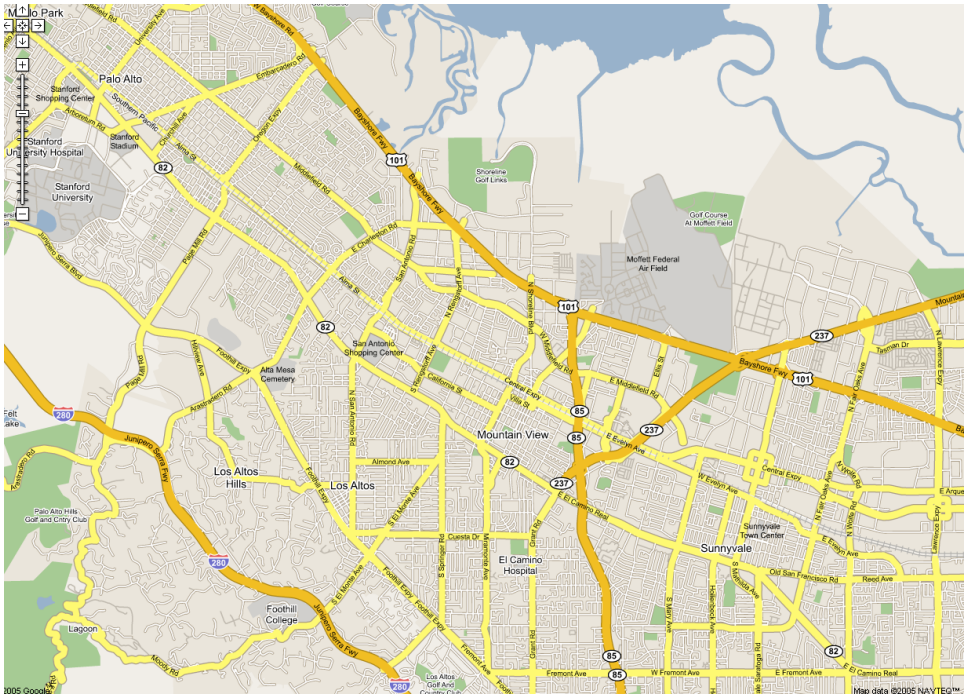
Different applications have different requirements that drive different design decisions

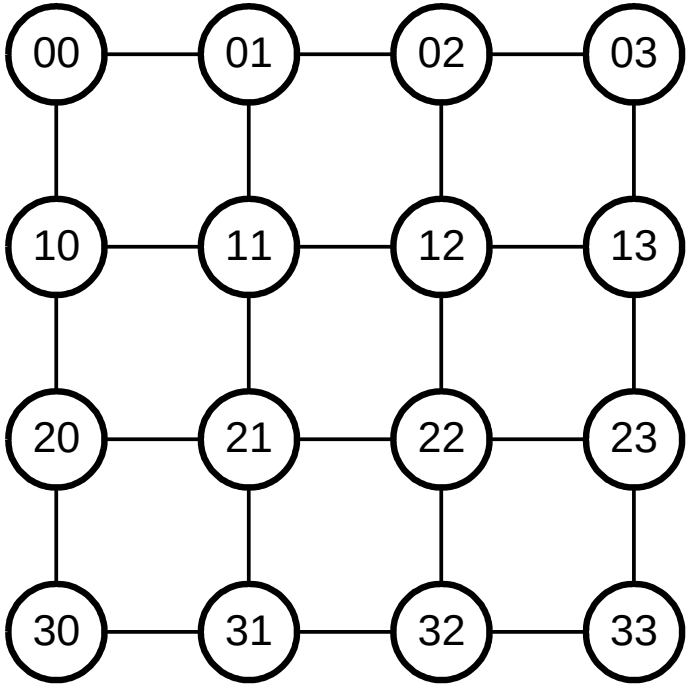
One size does not fit all.

But much technology applies to all!

Interconnection network basics

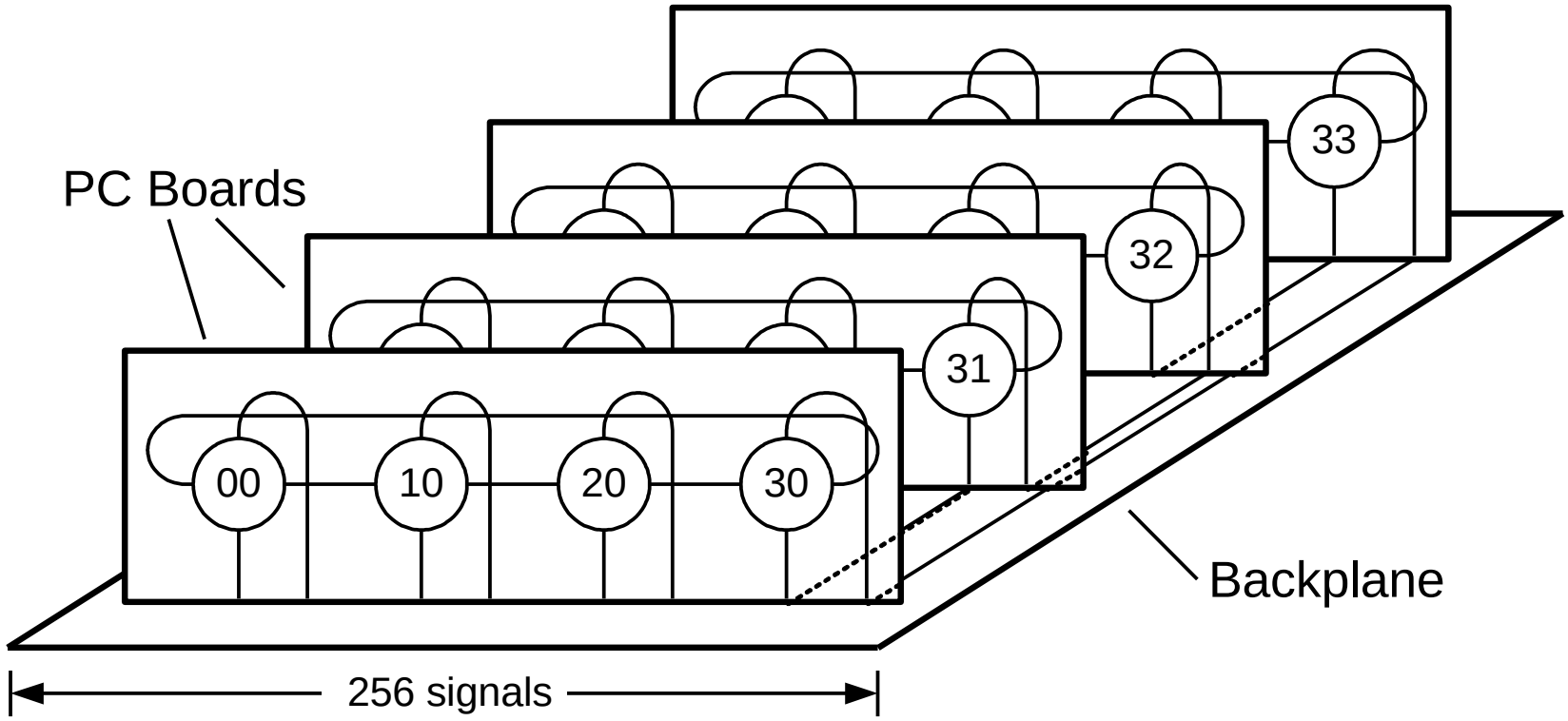
Topology


Menlo Park
Palo Alto
Stanford Shopping Center
Stanford University Hospital
Stanford Stadium
Stanford University
Shoreline Golf Links
Golf Course At Moffett Field
Moffett Federal Air Field
San Antonio Shopping Center
Alta Mesa Cemetery
Mountain View
Los Altos Hills
Los Altos
Sunnyvale
Sunnyvale Town Center
El Camino Hospital
Foothill College
Lagoon
Map data ©2005 NAVTEQ™

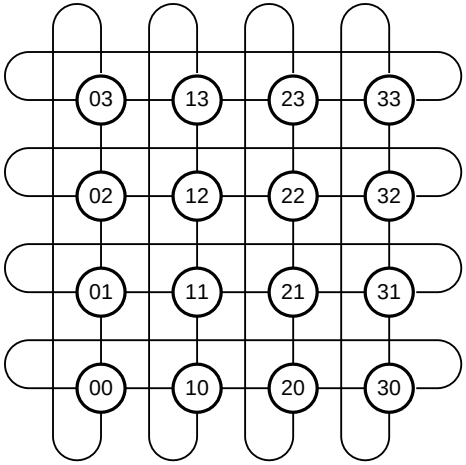

00
01
02
03
10
11
12
13
20
21
22
23
30
31
32
33

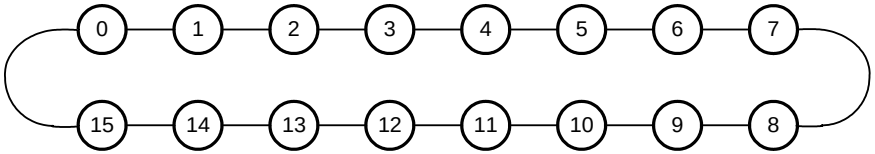
Topology is Constrained by Packaging

Chip pin count
Board pin count
Cable/backplane bisection

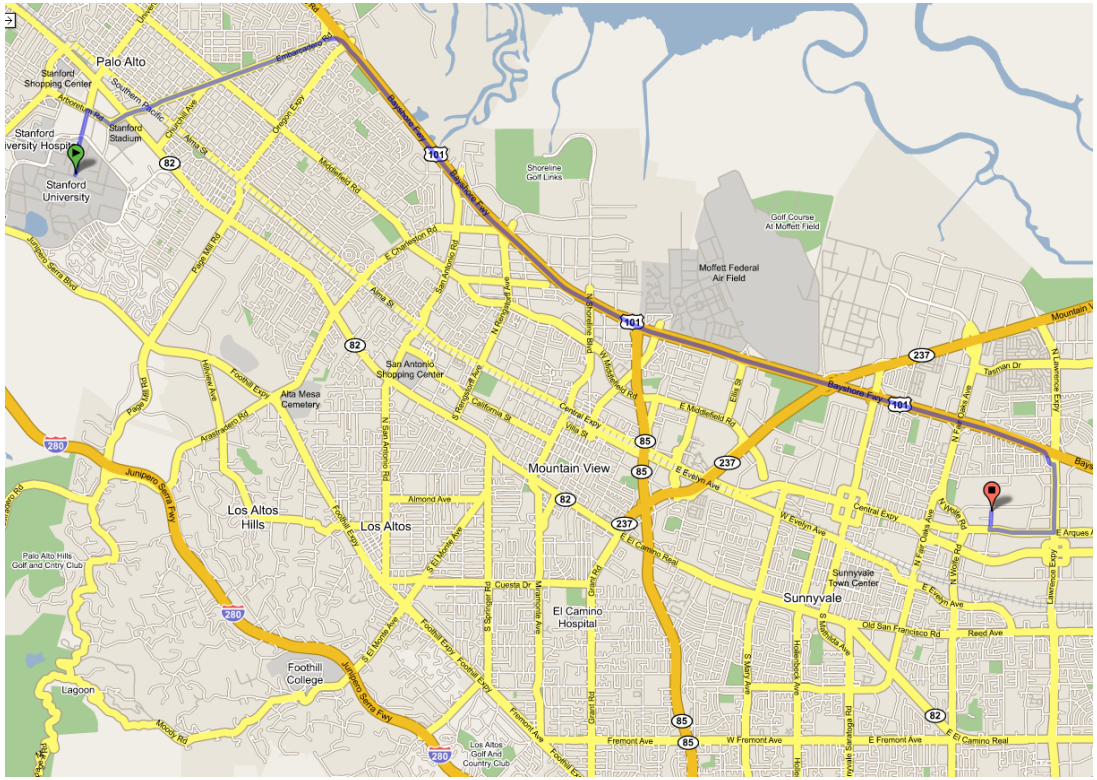


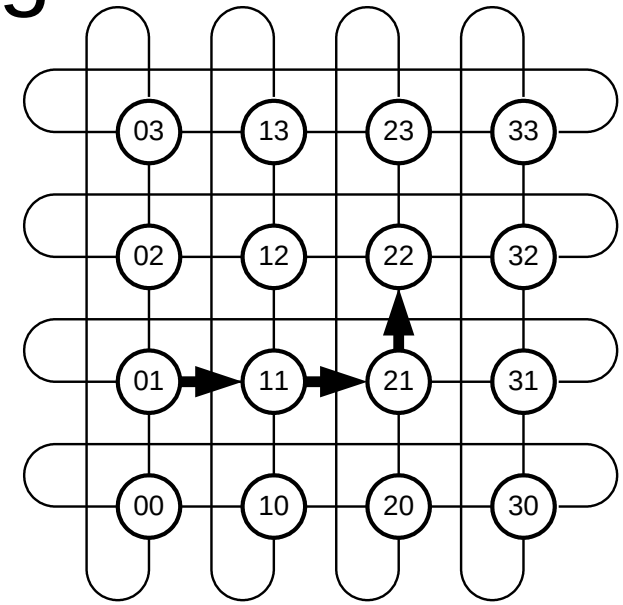
Example

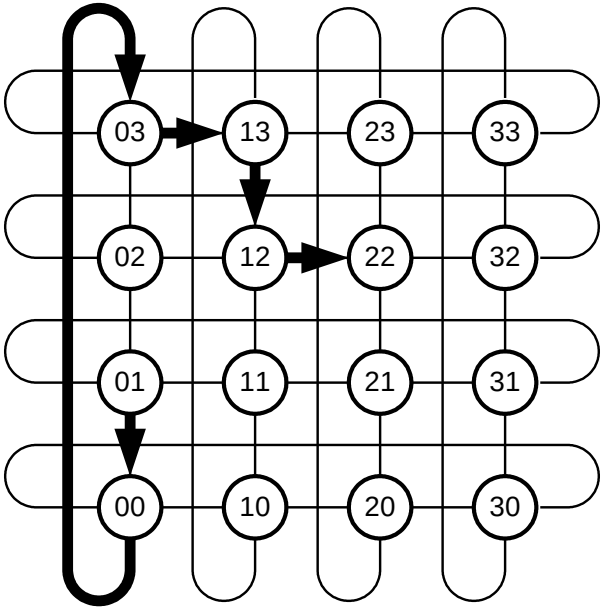

03
13
23
33
02
12
22
32
01
11
21
31
00
10
20
30


0
1
2
3
4
5
6
7
15
14
13
12
11
10
9
8

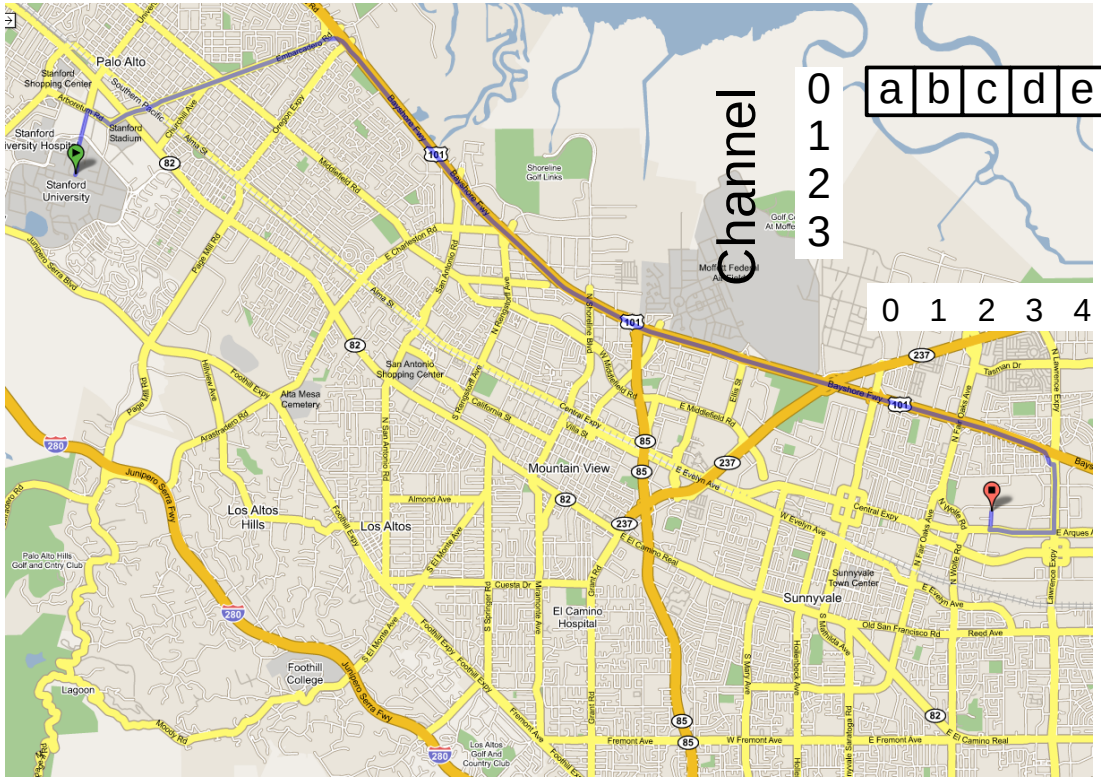
Routing – Selecting a Path

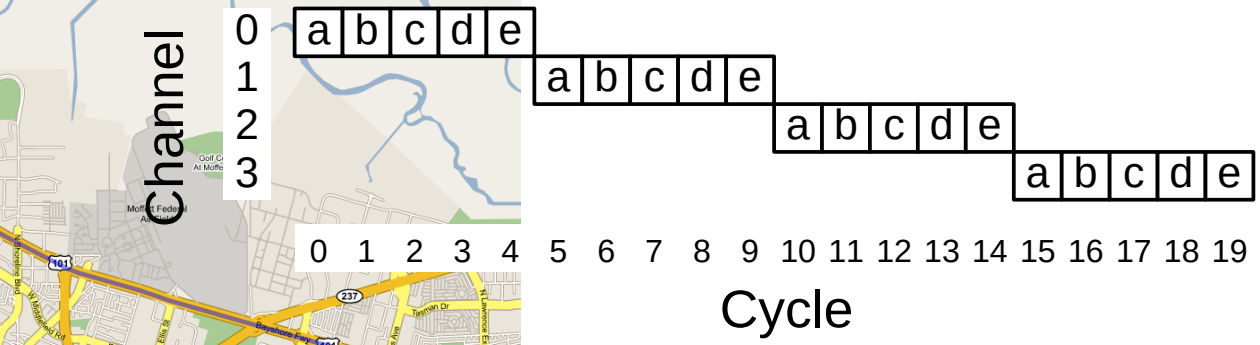


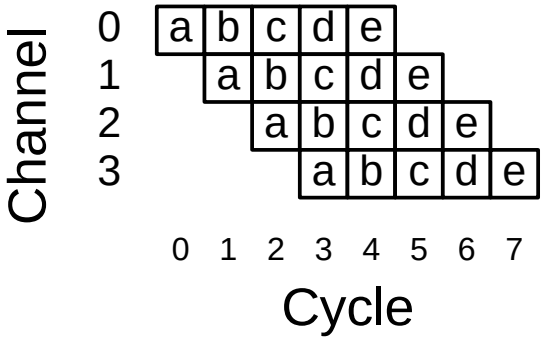




Flow Control – Scheduling Data Motion

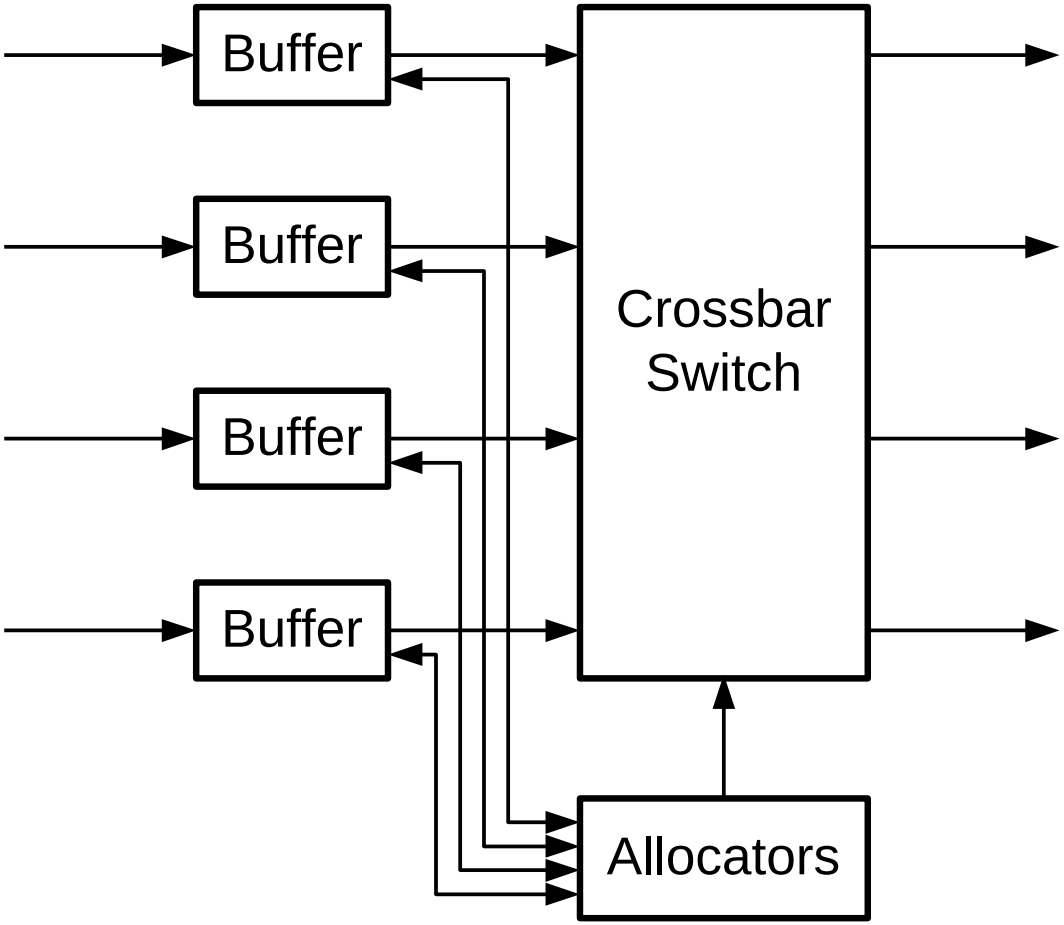




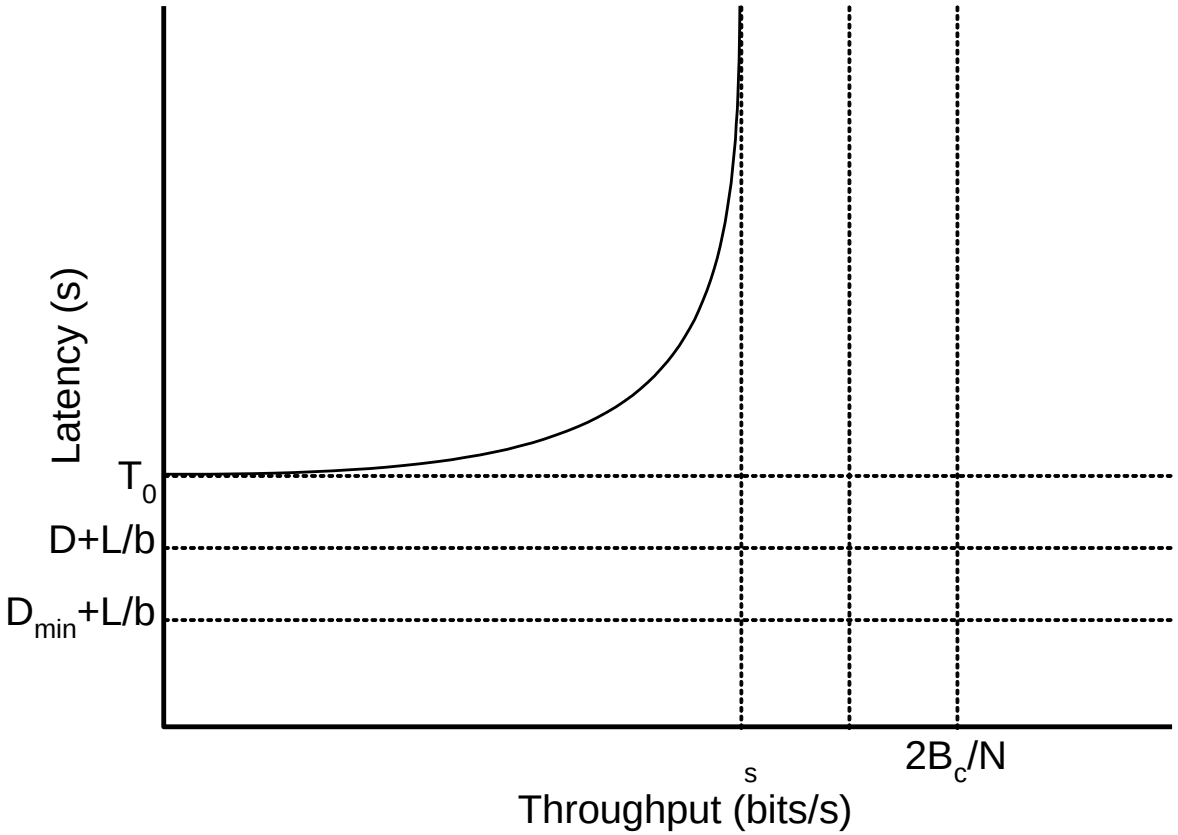


Allocate:
Buffers & Channels

Router Architecture

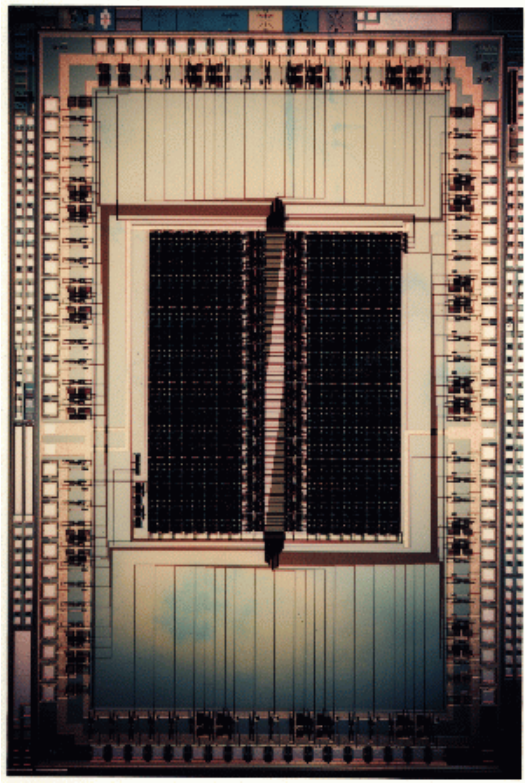


Performance

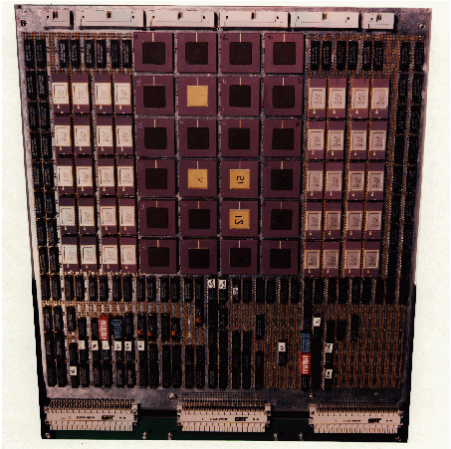


Some Personal Examples

Mars Router

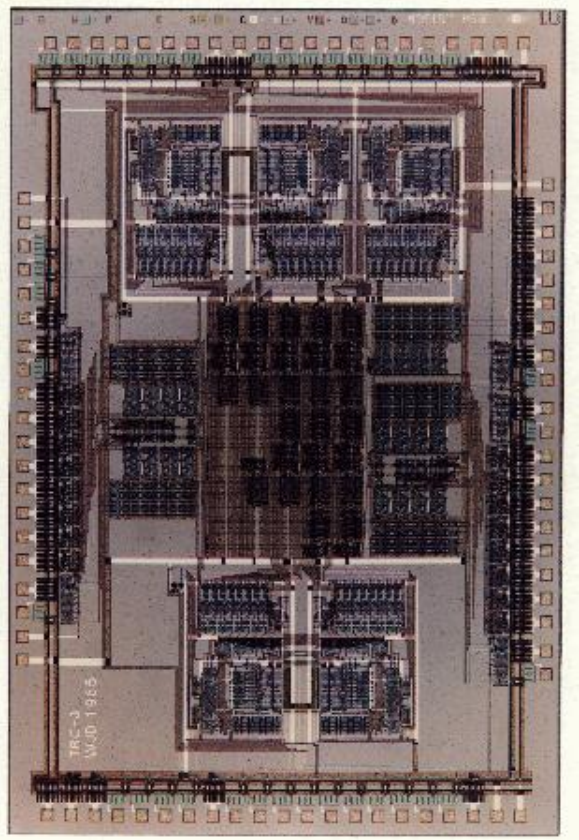


- 1984, 2.5μm CMOS
- 16 x16 crossbar
- Source-routed
- Bit-sliced (2-bits/chip)



Agrawal and Dally, "A Hardware Logic Simulation System", *IEEE TCAD*, 1990

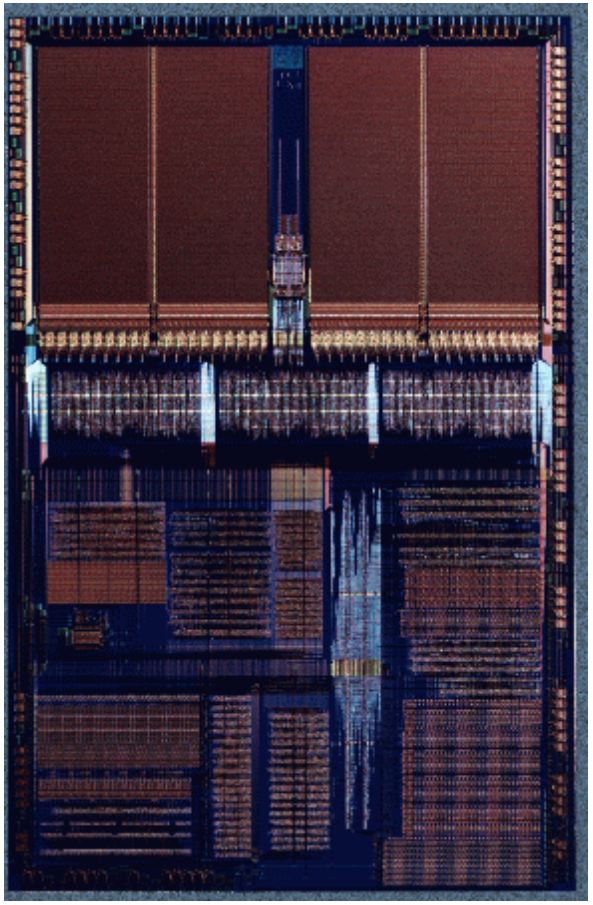
The Torus Routing Chip



- k-ary n-cube topology
  - 2D Torus Network
  - 8bit x 20MHz Channels
- Hardware routing
- Wormhole routing
- Virtual channels
- Fully Self-Timed Design
- Internal Crossbar Architecture

Dally and Seitz, "The Torus Routing Chip", *Distributed Computing*, 1986

Message-Driven Processor



- Integrated Processor, RAM, Router
- Router
  - 3D dimension partitioned
  - 8b x 32MHz multiplexed bidirectional
  - virtual channels
  - synchronous
- Network Interface
  - SEND instruction
  - Create, schedule, and dispatch process on message arrival
- Row buffers

Dally et. al., "Architecture of a Message-Driven Processor", *ISCA*, 1987
Dally et. al., "The Message-Driven Processor: A Multicomputer Processing Node with Efficient Mechanisms," *IEEE Micro*, 1992

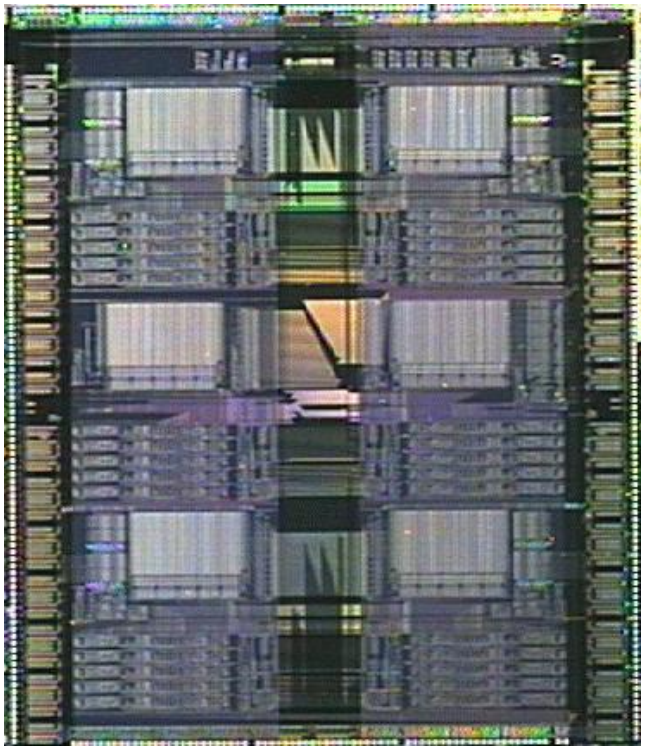
The J-Machine

- 1024 MDPs in a 8 x 8 x 16 grid
- Integrated 80-disk storage array
- Distributed frame buffer

Cray T3D

- 3D Torus network
- Synchronous routers
  - 1 dimension in each of 3 10K-gate ECL gate arrays
  - 150MHz x 16bit channels
- Dimension-order routing
- 4 Virtual channels/physical channel

The Reliable Router



- Fault-tolerant
  - Adaptive routing (adaptation of Duato's algorithm)
  - Link-level retry
  - Unique token protocol
- 32bit x 200MHz channels
  - Simultaneous bidirectional signalling
  - Low latency plesiochronous synchronizers
- Optimisitic routing

Dally, Dennison, Harris, Kan, and Xanthopoulos, "Architecture and Implementation of the Reliable Router", Hot Interconnects II, 1994
Dally, Dennison, and Xanthopoulos, "Low-Latency Plesiochronous Data Retiming, " ARVLSI 1995
Dennison, Lee, and Dally, "High Performance Bidirectional Signalling in VLSI Systems," SIS 1993

Cray T3E

- 3D Torus Network
- 375MHz x 14-bit channels
- Adaptive routing using adaptation of Duato's algorithm.
- Signal switching (for barrier network)

Scott and Thorson, "The Cray T3E Network: Adaptive Routing in a High-Performance 3D Torus

The Avici TSR

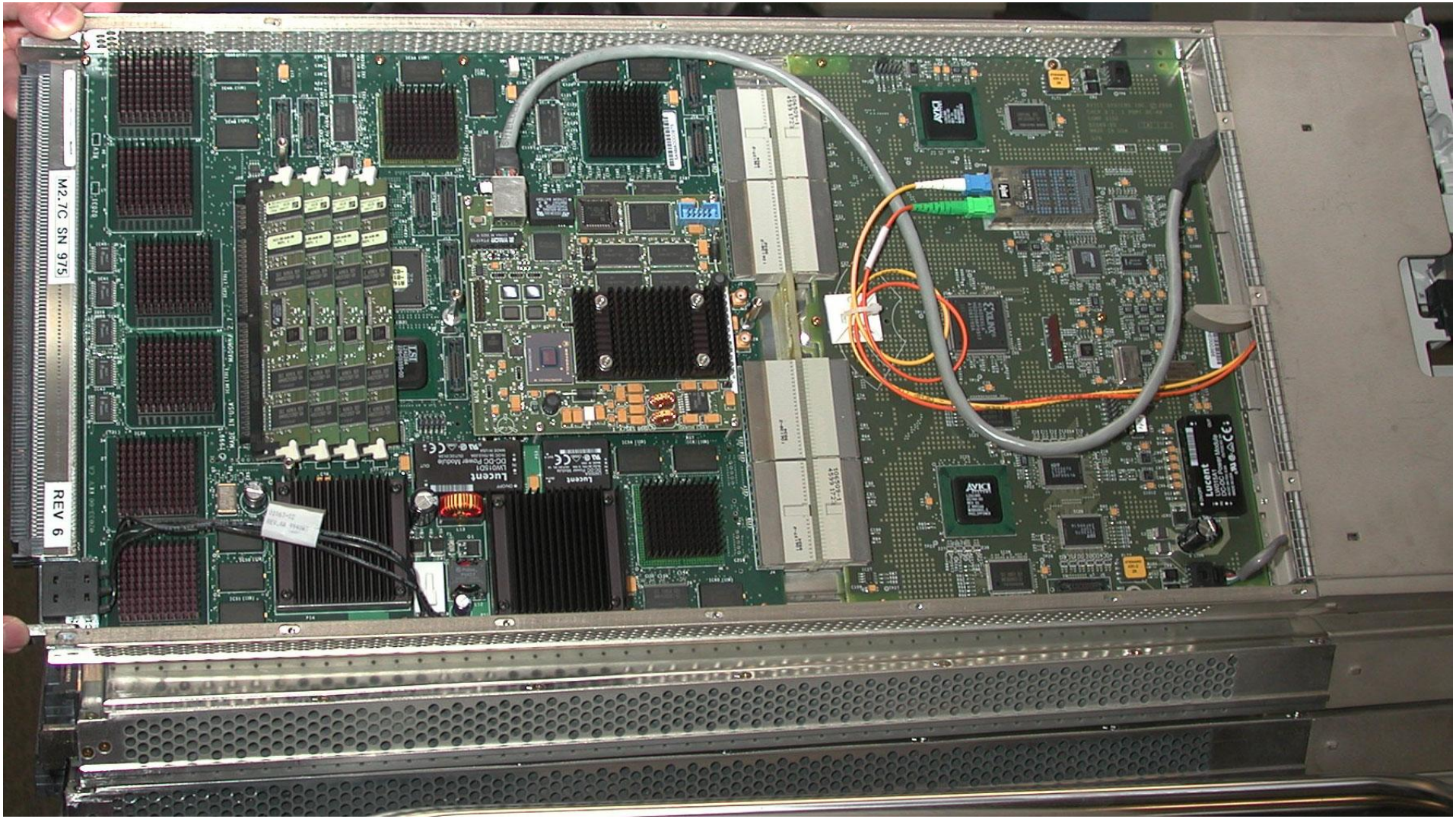
Up to 560 OC-48 Ports

Extensible one port at a time

Fault tolerant

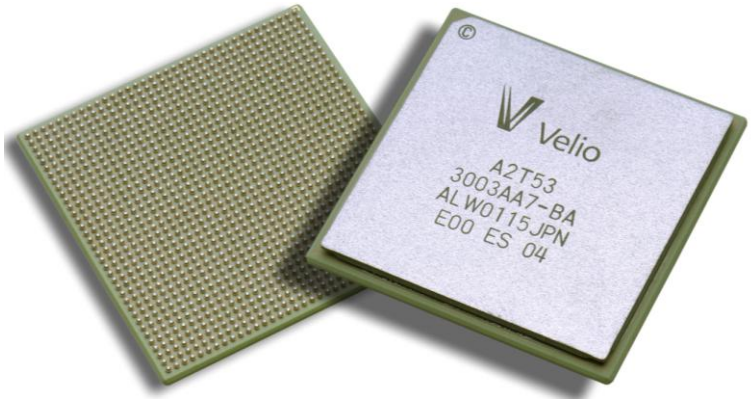
QoS - RED, WFQ, CBR at line speed

Avici TSR Line Module



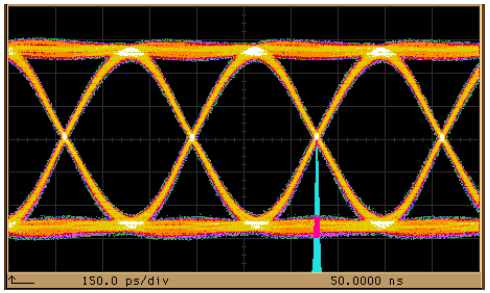
Velio VC2002 and VC2003

VC2002 72 x 72 x 2.488Gb/s
STS-1 Grooming Switch



VC3003 140 x 140 x 3.2Gb/s

Crosspoint Switch

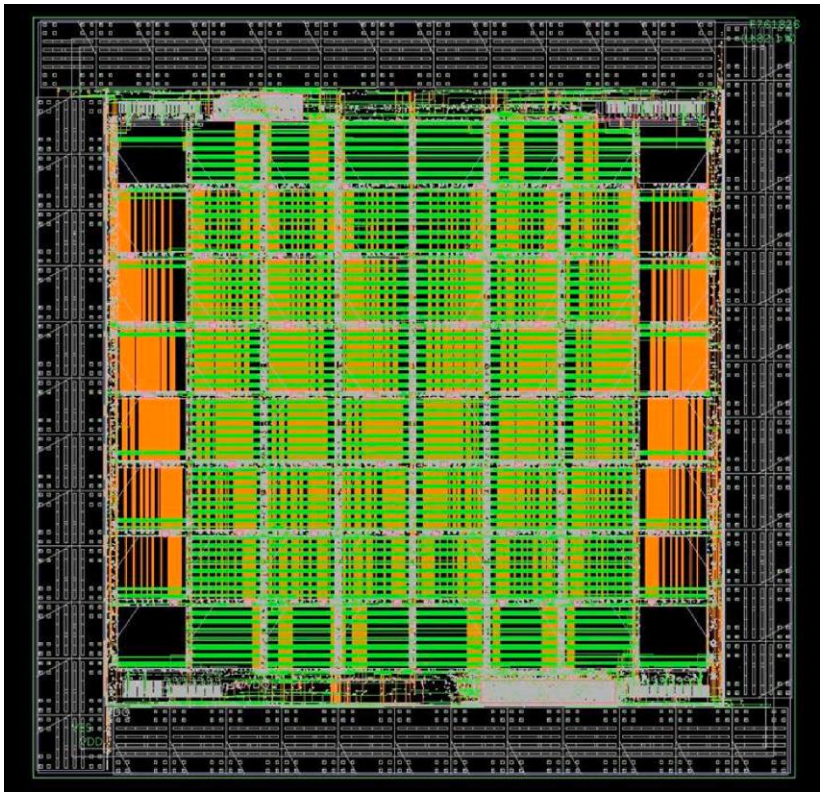


Cray Black Widow

- Shared-memory vector parallel computer
- Up to 32K nodes
- Vector processor per node
- Shared memory across nodes

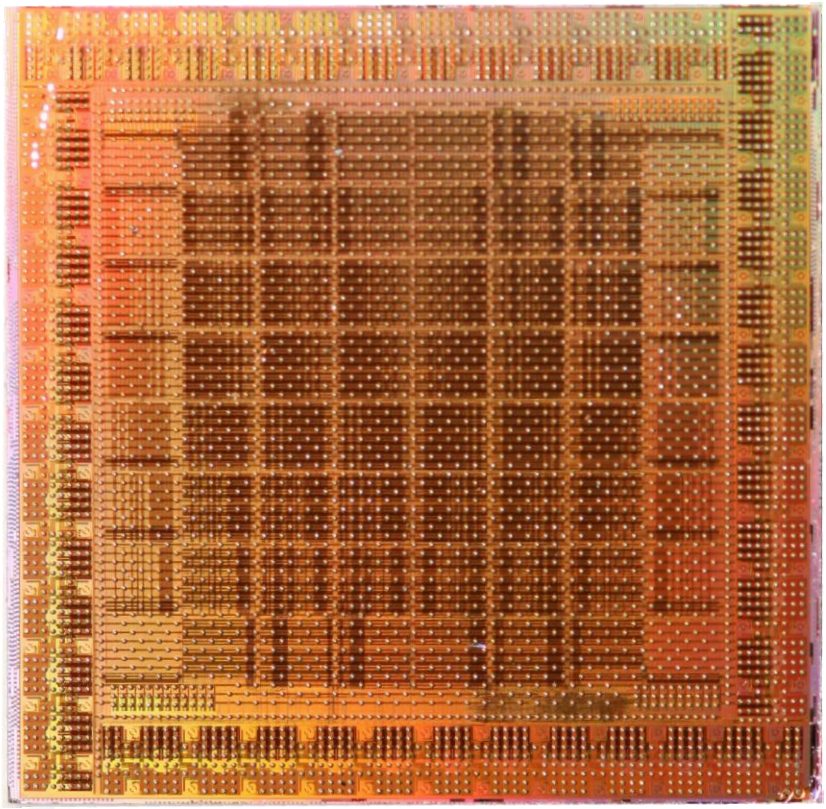
YARC Implementation

- Implemented in a 90nm CMOS standard-cell ASIC technology
- 192 SerDes on the chip
  - (64 ports x 3-bits per port)
- 6.25Gbaud data rate
- Estimated power
  - 80 W (idle)
  - 87 W (peak)
- 17mm x 17mm die



YARC Implementation

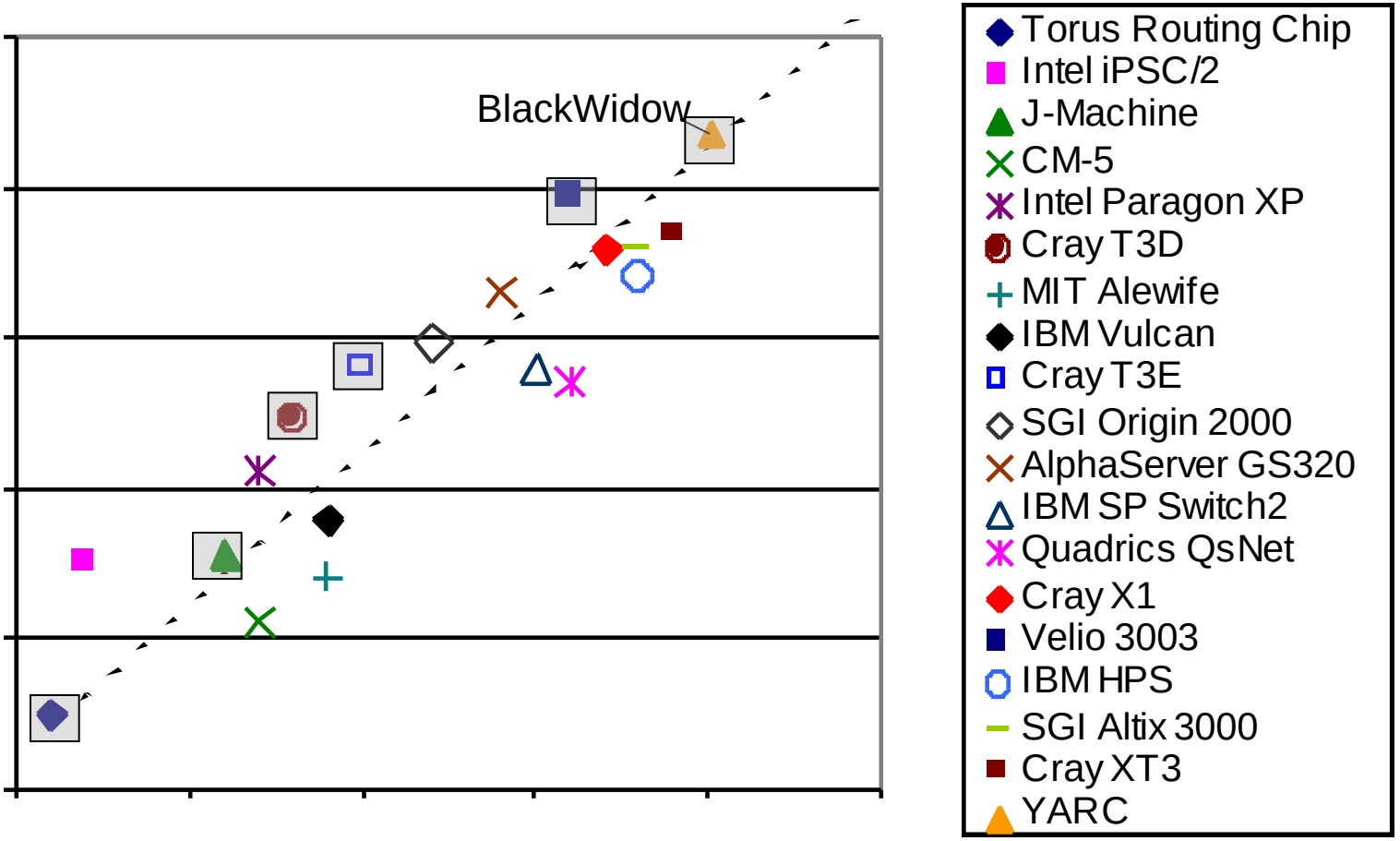
- Implemented in a 90nm CMOS standard-cell ASIC technology
- 192 SerDes on the chip
  - (64 ports x 3-bits per port)
- 6.25Gbaud data rate
- Estimated power
  - 80 W (idle)
  - 87 W (peak)
- 17mm x 17mm die



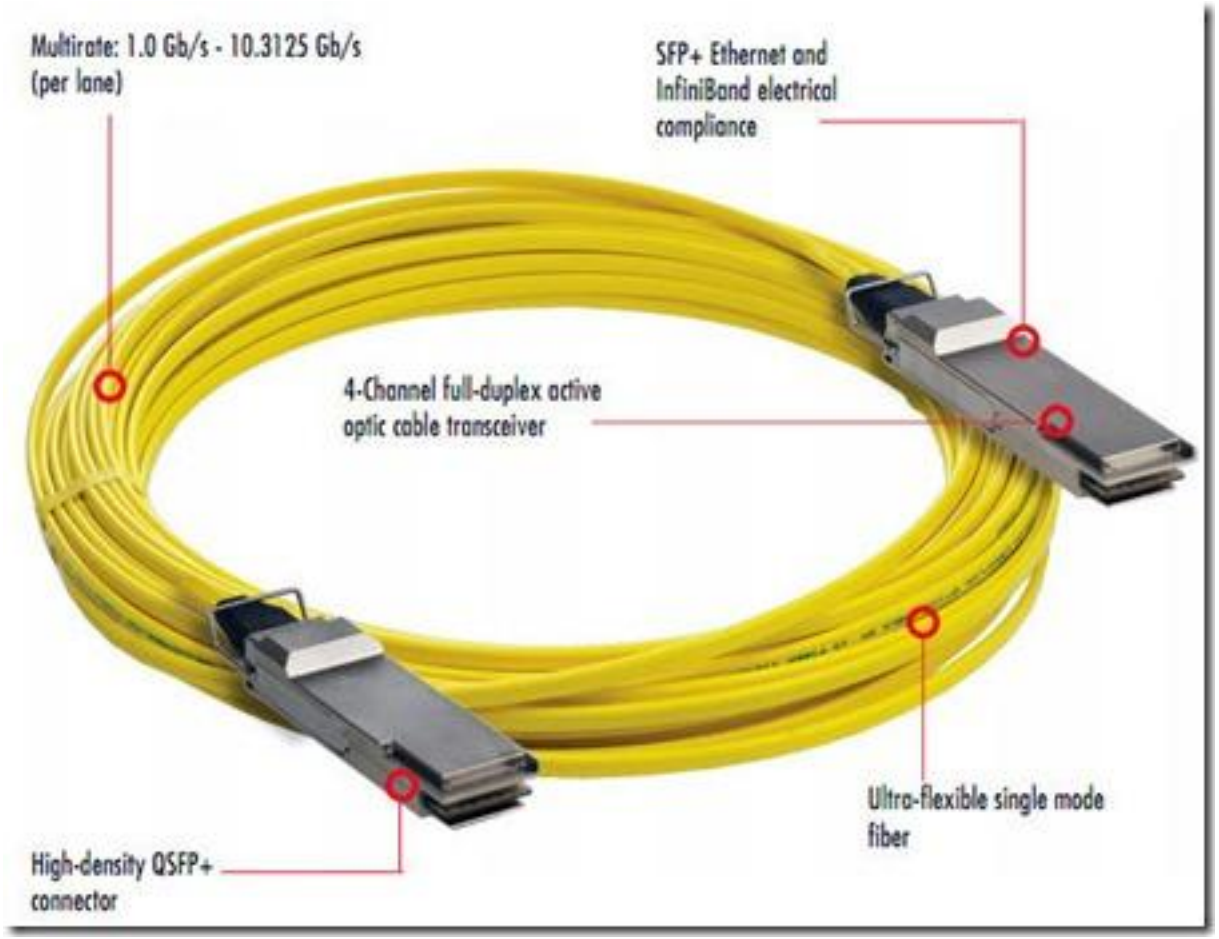
A Constantly Changing Field

- Changing Technology
  - Router pin bandwidth
  - Link bandwidth and energy
  - Techniques for topology, routing, flow control, router architecture
- Changing Applications
  - On-chip networks
  - Data center applications

Router Pin Bandwidth vs Time



Active Optical Cables



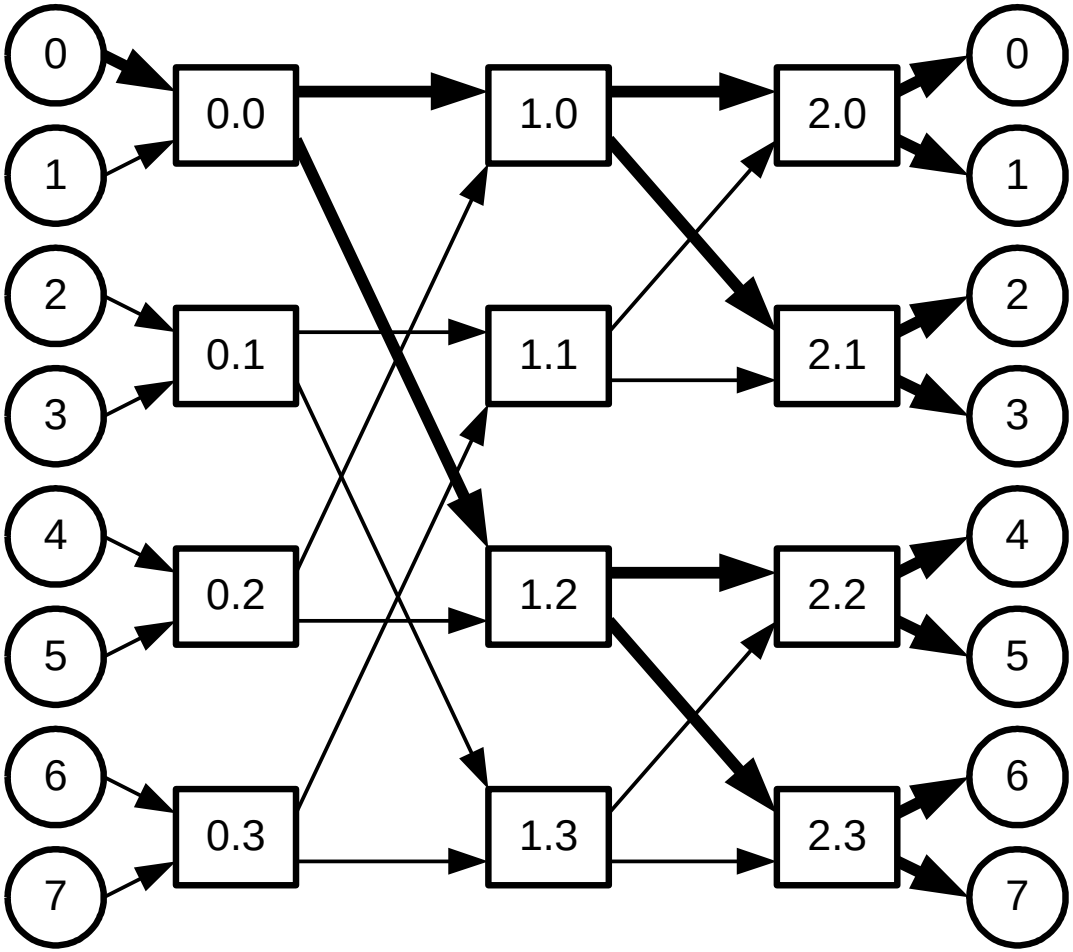
A Simple Interconnection Network

Requirements & Constraints

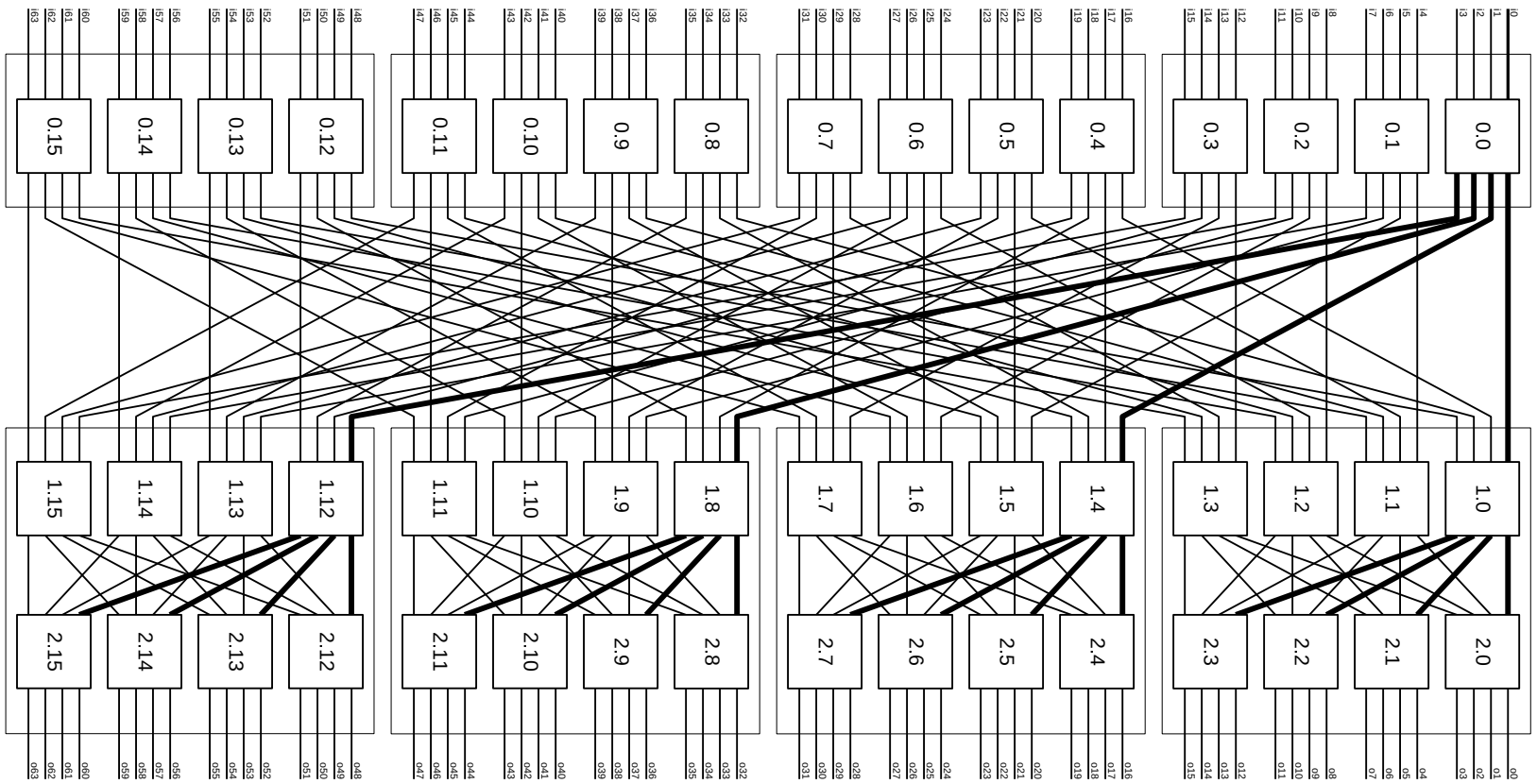
Parameter	Value
Input Ports	64
Output Ports	64
Peak Bandwidth	0.25 GByte/s
Average Bandwidth	0.25 GByte/s
Message Latency	100ns
Message Size	4–64 bytes
Traffic Pattern	random
Quality of Service	dropping acceptable
Reliability	dropping acceptable

Parameter	Value
Port Width	2 bits
Signaling rate	1 GHz
Signals per chip	150
Chip cost	$200
Chip pin bandwidth	1 Gb/s
Signals per circuit board	750
Circuit board cost	$200
Signals per cable	80
Cable cost	$50
Cable length	4m at 1Gb/s

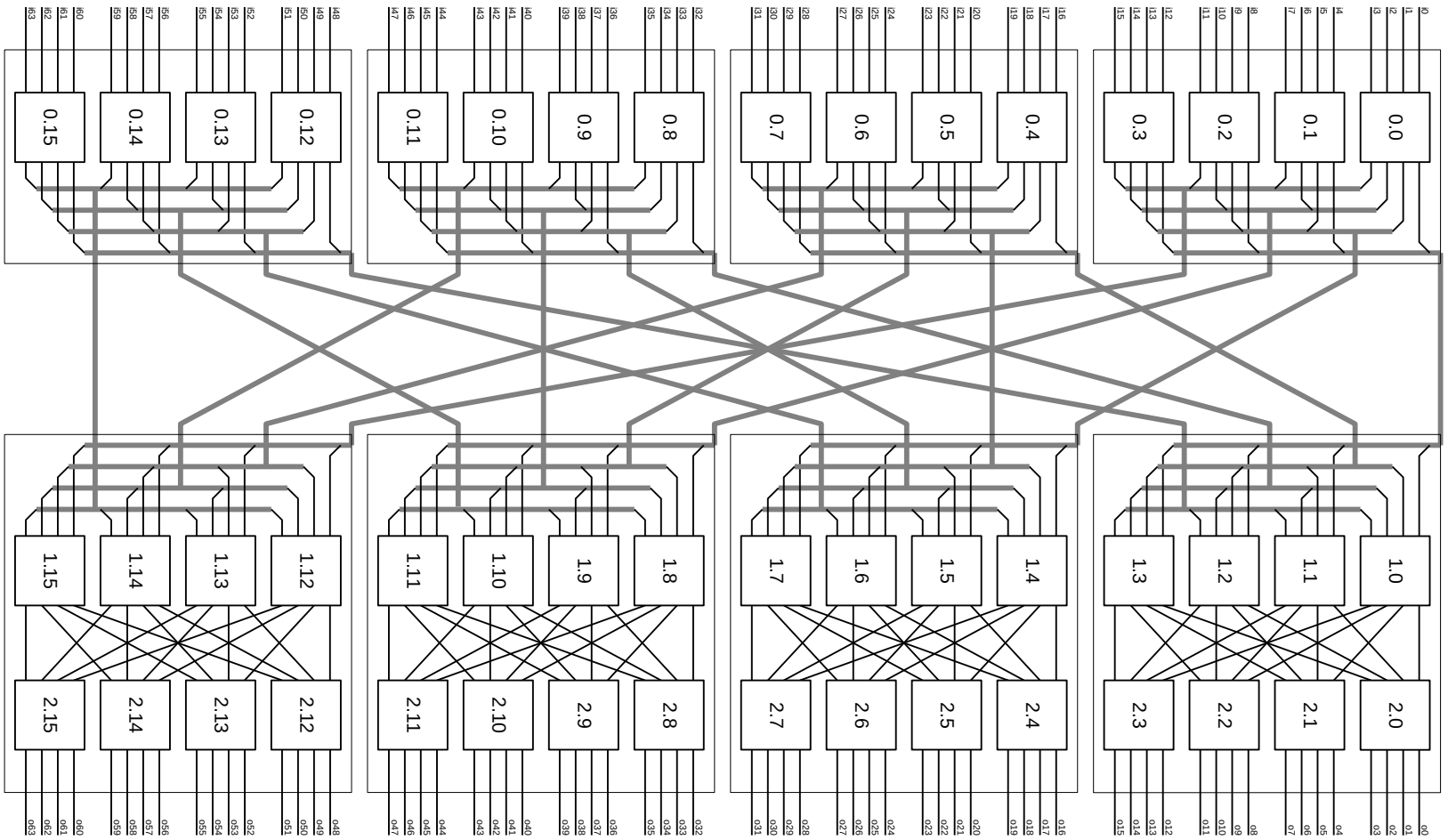
Topology – Butterfly



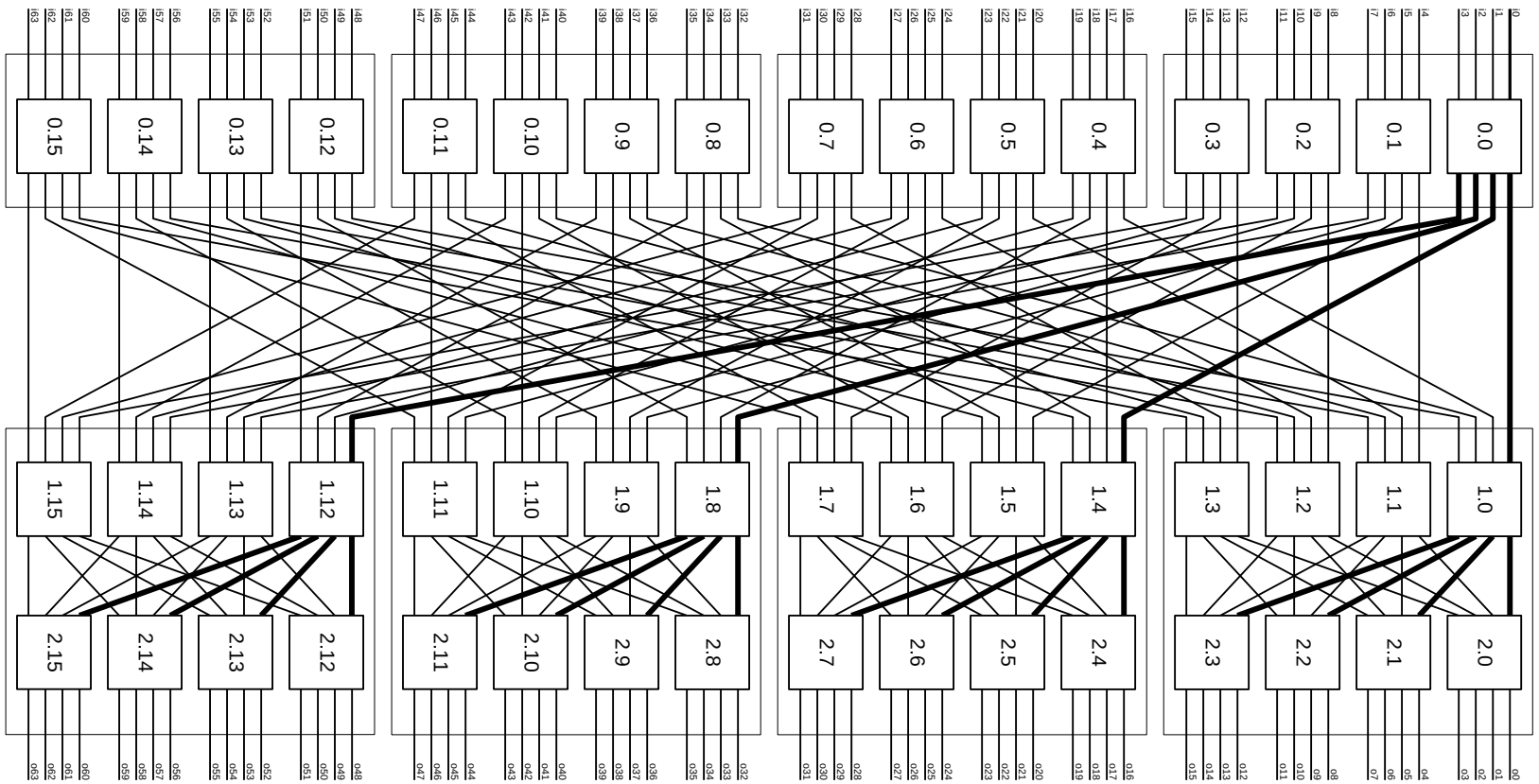
Fit to packaging constraints
4-ary 3-fly



Cables

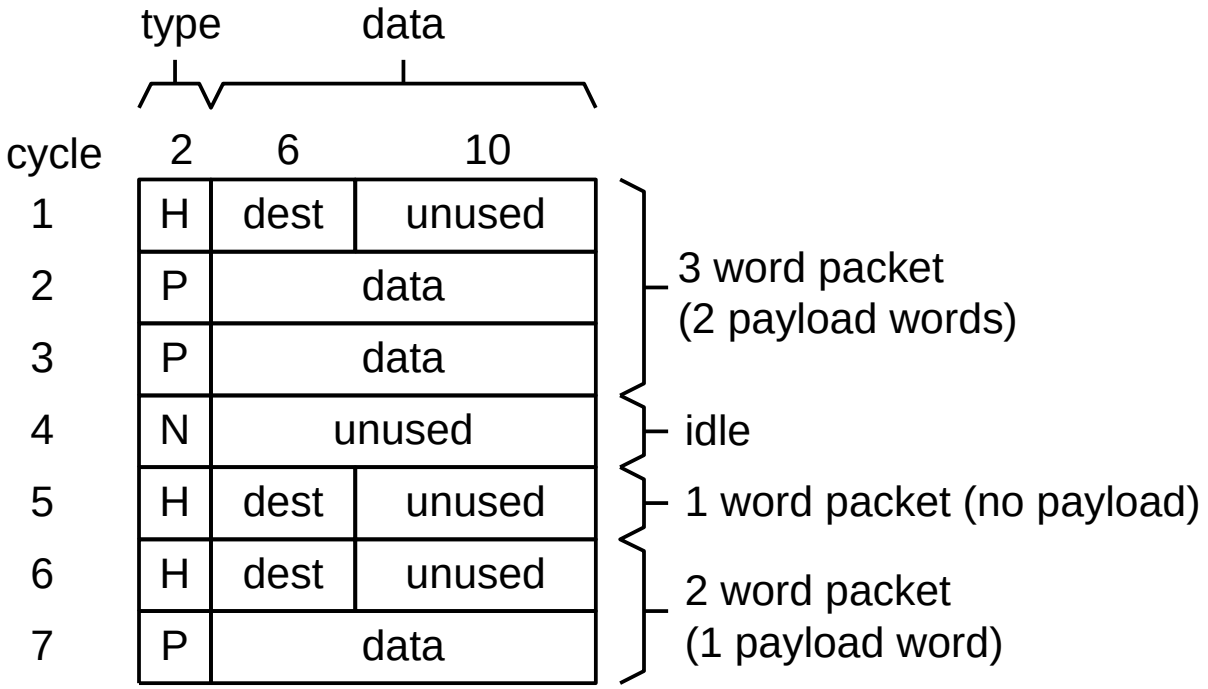

0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.10
0.11
0.12
0.13
0.14
0.15
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
1.10
1.11
1.12
1.13
1.14
1.15
2.0
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
2.10
2.11
2.12
2.13
2.14
2.15

Routing – By Destination Tag

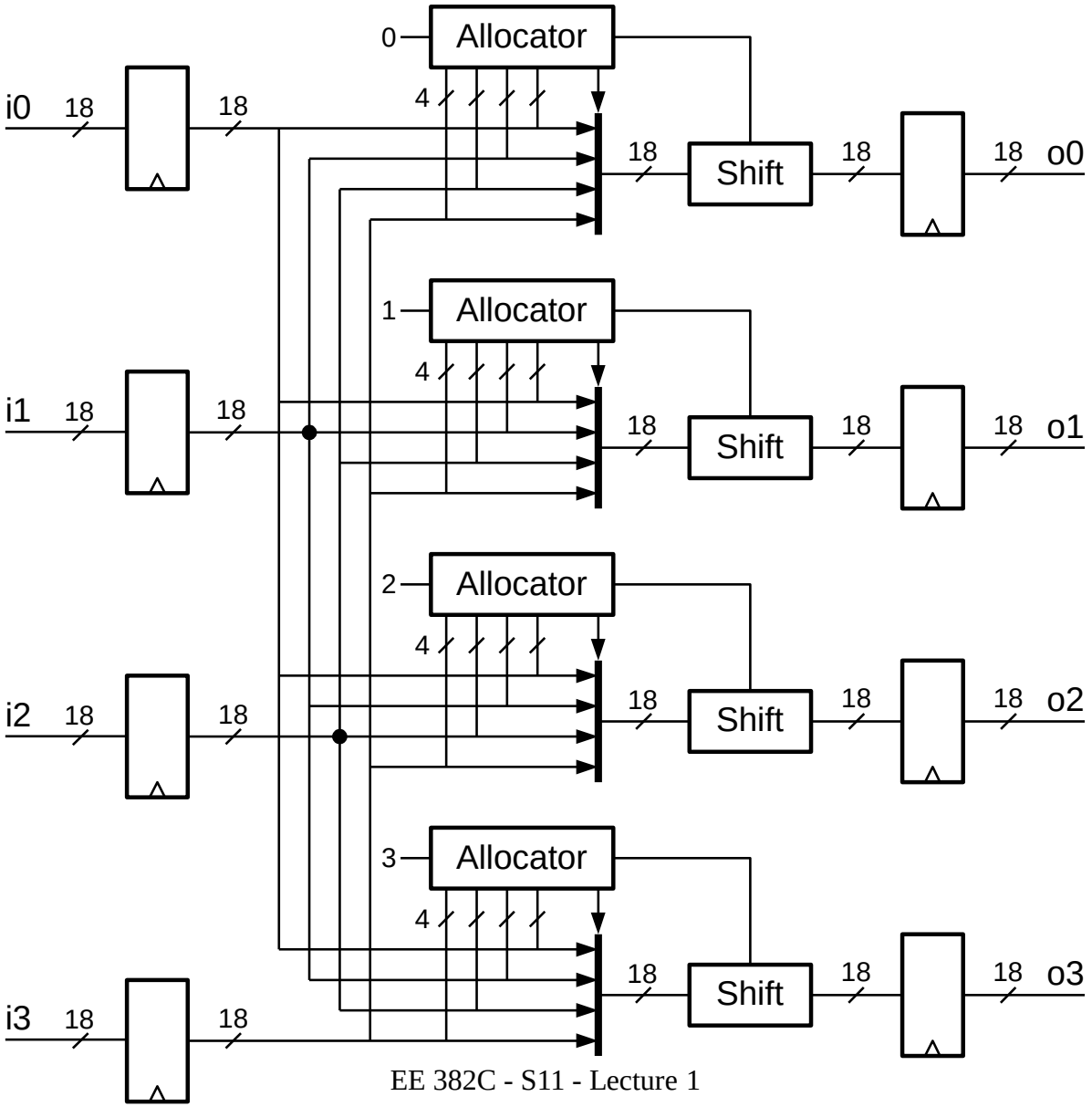


Flow Control

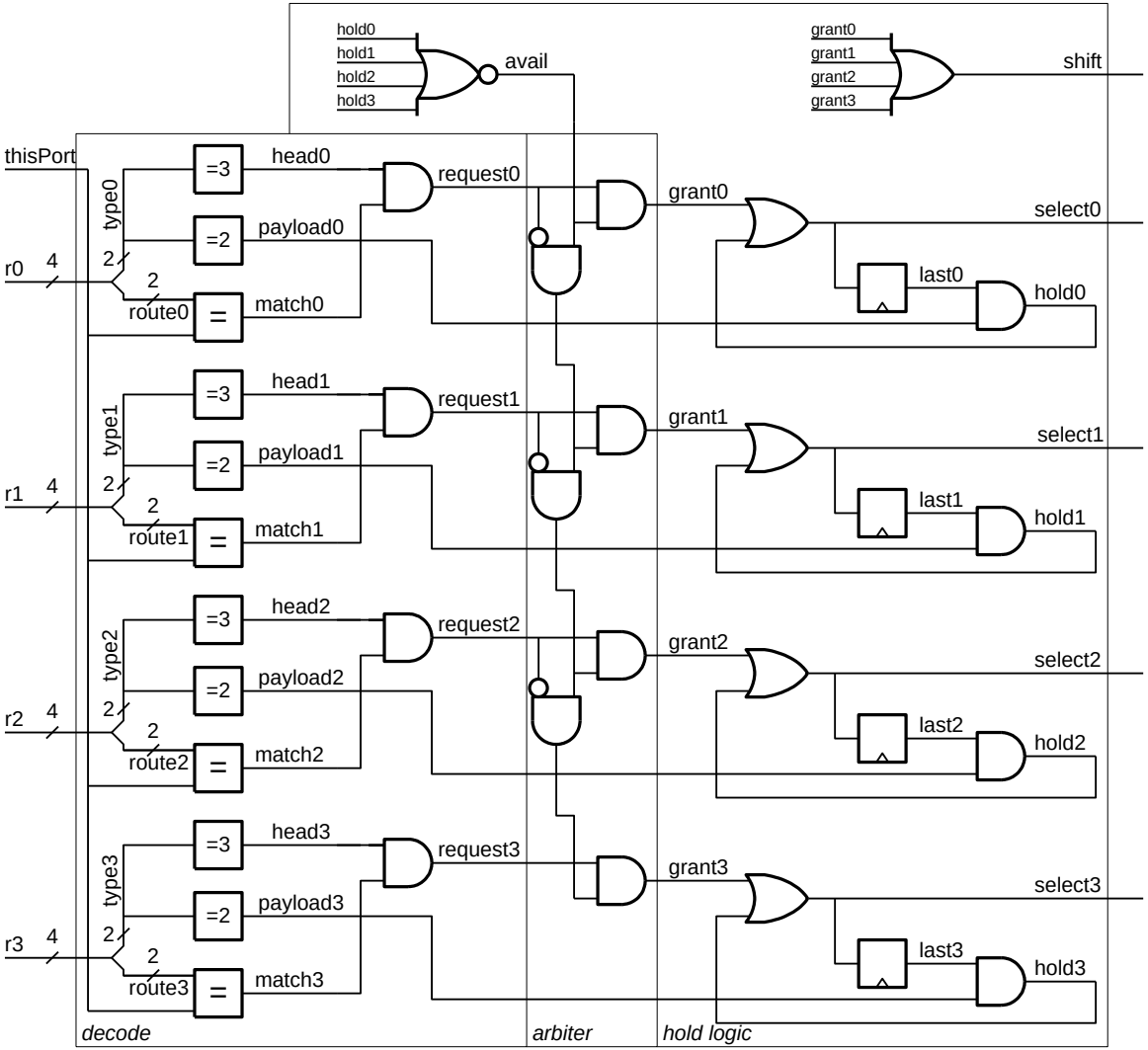
- Drop on conflict
  Assume higher level retransmit protocol



Router Architecture



Allocator


hold0
hold1
hold2
hold3
avail
grant0
grant1
grant2
grant3
shift
thisPort
r0
r1
r2
r3
4
type0
type1
type2
type3
2
=3
=2
=
head0
payload0
match0
route0
request0
grant0
select0
last0
hold0
head1
payload1
match1
route1
request1
grant1
select1
last1
hold1
head2
payload2
match2
route2
request2
grant2
select2
last2
hold2
head3
payload3
match3
route3
request3
grant3
select3
last3
hold3
decode
arbiter
hold logic

Interconnection networks

- Programmable message delivery
- Connect
  - Processor-memory
  - Processor-I/O
  - Router line cards
- Basics
  - Topology
  - Routing
  - Flow control
  - Router Architecture
  - Performance