Interconnect-Centric Computing

William J. Dally
Computer Systems Laboratory
Stanford University

HPCA Keynote
February 12, 2007
Outline

• Interconnection Networks (INs) are THE central component of modern computer systems
• Topology driven to high-radix by packaging technology
• Global adaptive routing balances load - and enables efficient topologies
• Case study, the Cray Black Widow
• On-Chip Interconnection Networks (OCINs) face unique challenges
• The road ahead…
Outline

- Interconnection Networks (INs) are THE central component of modern computer systems
- Topology driven to high-radix by packaging technology
- Global adaptive routing balances load - and enables efficient topologies
- Case study, the Cray Black Widow
- On-Chip Interconnection Networks (OCINs) face unique challenges
- The road ahead…
INs: Connect Processors in Clusters

IBM Blue Gene
and on chip
Connect Processors to Memories in Systems

Cray Black Widow
and on chip

Texas TRIPS
provide the fabric for network Switches and Routers

Avici TSR
and connect I/O Devices

Brocade Switch
Group History: Routing Chips & Interconnection Networks

- Mars Router, Torus Routing Chip, Network Design Frame, Reliable Router
- Basis for Intel, Cray/SGI, Mercury, Avici network chips
Group History: Parallel Computer Systems

• J-Machine (MDP) led to Cray T3D/T3E
• M-Machine (MAP)
  - Fast messaging, scalable processing nodes, scalable memory architecture
• Imagine – basis for SPI
Interconnection Networks are THE Central Component of Modern Computer Systems

- Processors are a commodity
  - Performance no longer scaling (ILP mined out)
  - Future growth is through CMPs - connected by INs

- Memory is a commodity
  - Memory system performance determined by interconnect

- I/O systems are largely interconnect

- Embedded systems built using SoCs
  - Standard components
  - Connected by on-chip INs (OCINs)
Outline

• Interconnection Networks (INs) are THE central component of modern computer systems
• Topology driven to high-radix by packaging technology
• Global adaptive routing balances load - and enables efficient topologies
• Case study, the Cray Black Widow
• On-Chip Interconnection Networks (OCINs) face unique challenges
• The road ahead…
Technology Trends...
High-Radix Router
**High-Radix Router**

- **Low-radix (small number of fat ports)**
- **High-radix (large number of skinny ports)**
Low-Radix vs. High-Radix Router

Latency:
- Low-Radix: 4 hops
- High-Radix: 2 hops

Cost:
- Low-Radix: 96 channels
- High-Radix: 32 channels
Latency

Latency  =  \( H t_r + \frac{L}{b} \)
=  \( 2t_r \log_k N + \frac{2kL}{B} \)

where \( k = \text{radix} \)
\( B = \text{total router Bandwidth} \)
\( N = \# \text{ of nodes} \)
\( L = \text{message size} \)
Latency vs. Radix

- Optimal radix ~ 40 (2003 technology)
- Optimal radix ~ 128 (2010 technology)

Header latency decreases
Serialization latency increases
Determining Optimal Radix

Latency = Header Latency + Serialization Latency
= $H t_r + L / b$
= $2t_r \log_k N + 2kL / B$

where $k = \text{radix}$
$B = \text{total router Bandwidth}$
$N = \# \text{ of nodes}$
$L = \text{message size}$

Optimal radix

$\Rightarrow k \log_2 k = (B t_r \log N) / L$
= Aspect Ratio
Higher Aspect Ratio, Higher Optimal Radix

![Graph showing the relationship between Aspect Ratio and Optimal Radix (k) from 1991 to 2010. The graph indicates an increasing trend where higher aspect ratios correspond to higher optimal radices.]
High-Radix Topology

- Use high radix, $k$, to get low hop count
  - $H = \log_k(N)$

- Provide good performance on both benign and adversarial traffic patterns
  - Rules out butterfly networks - no path diversity
  - Clos networks work well
    - $H = 2\log_k(N)$ - with short circuit
  - Cayley graphs have nice properties but are hard to route
Example radix-64 Clos Network
Flattened Butterfly Topology
Packaging the Flattened Butterfly

Dimension 1 connections

Dimension 1

Dimension 2

Dimension 3 connections

R0

R1

R15

P0
P1
P15
P16
P17
P31
P240
P241
P250
Packaging the Flattened Butterfly (2)
Outline

• Interconnection Networks (INs) are THE central component of modern computer systems
• Topology driven to high-radix by packaging technology
• Global adaptive routing balances load - and enables efficient topologies
• Case study, the Cray Black Widow
• On-Chip Interconnection Networks (OCINs) face unique challenges
• The road ahead…
Routing in High-Radix Networks

• Adaptive routing avoids transient load imbalance
• Global adaptive routing balances load for adversarial traffic
  – Cost/perf of a butterfly on benign traffic and at low loads
  – Cost/perf of a clos on adversarial traffic
A Clos can statically load balance traffic using oblivious routing.
Transient Imbalance

Routers in the middle stage of the network
With Adaptive Routing

![Bar Chart with Labels: Maximum buffer size on the y-axis and Routers in the middle stage of the network on the x-axis. The chart shows varying buffer sizes across the routers.]
Latency for UR traffic

- oblivious
- adaptive

Latency (cycles)

Offered load
Flattened Butterfly Topology
Flattened Butterfly Topology

What if node 0 sends all of its traffic to node 1?
What if node 0 sends all of its traffic to node 1?

How much traffic should we route over alternate paths?
Simpler Case - ring of 8 nodes
Send traffic from 2 to 5

- Model: Assume queues to be a network of independent M/D/1 queues

\[
\alpha = x_1 + x_2
\]

Min path delay = \(D_m(x_1)\)
Non-min path delay = \(D_{nm}(x_2)\)

- Routing remains minimal as long as

\[
D_m'(\alpha) \leq D_{nm}'(0)
\]

- Afterwards, route a fraction, \(x_2\), non-minimally such that

\[
D_m'(x_1) = D_{nm}'(x_2)
\]
Traffic divides to balance delay
Load balanced at saturation

![Graph showing Accepted Throughput vs. Offered Load](image)

- **Model Overall**
- **Model Minimal**
- **Model Non-minimal**
Channel-Queue Routing

- Estimate delay per hop by local queue length $Q_i$
- Overall latency estimated by
  - $L_i \sim Q_i H_i$
- Route each packet on route with lowest estimated $L_i$
- Works extremely well in practice
Performance on UR Traffic

![Graph showing performance on UR Traffic with different load levels and latency measurements. The graph compares VAL, MIN, UGAL, UGAL-S, and CLOS AD.]
Performance on WC Traffic

- MIN
- VAL
- UGAL
- UGAL-S
- CLOS AD

Latency (cycles) vs. Offered load
Allocator Design Matters

![Graph showing latency vs. offered load for different allocator designs](image-url)
Outline

• Interconnection Networks (INs) are THE central component of modern computer systems
• Topology driven to high-radix by packaging technology
• Global adaptive routing balances load - and enables efficient topologies
• Case study, the Cray Black Widow
• On-Chip Interconnection Networks (OCINs) face unique challenges
• The road ahead…
Putting it all together
The Cray BlackWidow Network

In collaboration with Steve Scott and Dennis Abts (Cray Inc.)
Cray Black Widow

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

- Up to 32K nodes in a 3-level folded Clos
- Each node has 4 18.75Gb/s channels, one to each of 4 network slices
YARC
Yet Another Router Chip

- 64 Ports
- Each port is 18.75 Gb/s (3 x 6.25Gb/s links)
- Table-driven routing
- Fault tolerance
  - CRC with link-level retry
  - Graceful degradation of links
    - 3 bits -> 2 bits -> 1 bit -> OTS
YARC Microarchitecture

- Regular 8x8 array of tiles
  - Easy to lay out chip
- No global arbitration
  - *All* decisions local
- Simple routing
- Hierarchical organization
  - Input buffers
  - Row buffers
  - Column buffers
A Closer Look at a Tile

- No global arbitration
- Non-blocking with an 8x internal speedup in subswitch
- Simple routing
  - Small 8-entry routing table per tile
  - High routing throughput for small packets
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
Outline

• Interconnection Networks (INs) are THE central component of modern computer systems
• Topology driven to high-radix by packaging technology
• Global adaptive routing balances load - and enables efficient topologies
• Case study, the Cray Black Widow
• On-Chip Interconnection Networks (OCINs) face unique challenges
• The road ahead...
Much of the future is on-chip (CMP, SoC, Operand)
On-Chip Networks are Fundamentally Different

- Different cost model
  - Wires plentiful, no pin constraints
  - Buffers expensive (consume die area)
  - Slow signal propagation

- Different usage patterns
  - Particularly for SoCs
    - Significant isochronous traffic
    - Hard RT constraints

- Different design problems
  - Floorplans
  - Energy-efficient transmission circuits
NSF Workshop Identified 3 Critical Issues

• Power
  – OCINs will have 10x the required power with current approaches
    • Circuit and architecture innovations can close this gap

• Latency
  – OCIN latency currently not competitive with buses and dedicated wiring
    • Novel flow-control strategies required

• Tool Integration
  – OCINs need to be integrated with standard tool flows to enable widespread use
The Road Ahead

- INs become an even more dominant system component
  - Number of processors goes up, cost of processors decreases
  - Communication dominates performance and cost
  - From hand-held media UI devices to huge data centers

- Technology drives topology in new directions
  - On-chip, short reach electrical (10m), optical
  - Expect radix to continue to increase
  - Hybrid topologies to match each packaging level

- Latency will approach that of dedicated wiring
  - Better flow-control and router architecture
  - Optimized circuits

- Adaptivity will optimize performance
  - Balance load, route around defects, tolerate variation, tune power to load
Summary

• Interconnection Networks (INs) are THE central component of modern computing systems
• High-radix topologies have evolved to exploit packaging/signaling technology
  – Including hybrid optical/electrical
  – Flattened Butterfly
• Global adaptive routing balances load and enables advanced topologies
  – Eliminate transient load imbalance
  – Use local queues to estimate global congestion
• Cray Black Widow - an example high-radix network
• On-Chip INs
  – Very different constraints
  – Three “Gaps” identified - power, latency, tools.
• The road ahead
  – Lots of room for improvement, INs are in their infancy
Some very good books
Backup
Virtual Channel Router Architecture

Routing computation

VC Allocator

Switch Allocator

Crossbar switch

Input 1

VC 1
VC 2
VC v

Output 1

Input k

VC 1
VC 2
VC v

Output k
Baseline Performance Evaluation

- low-radix
Baseline Performance Evaluation

The graph shows the latency (cycles) for different offered loads, comparing low-radix and baseline (high-radix) performance. The graph indicates that the low-radix design performs better than the baseline design, especially at higher offered loads.