Microarchitecture of a High-Radix Router

John Kim, William J. Dally, Brian Towles¹, and Amit K. Gupta

Concurrent VLSI Architecture
Stanford University

¹D.E. Shaw
Research & Development
Interconnection Network

Supercomputer networks: Cray X1

On-chip Networks: MIT RAW

I/O Interconnects: Myrinet/Infiniband

Router Fabrics: Avici TSR
Bandwidth Trend

Bandwidth growth result of:
1. Increase in signaling rate
2. Increase in number of signals

Torus Routing Chip
Intel iPSC/2
J-Machine
CM-5
Intel Paragon XP
Cray T3D
MIT Alewife
IBM Vulcan
Cray T3E
SGI Origin 2000
Alpha Server GS320
IBM HPS
SGI Altix 3000

High-Radix Routers
Bandwidth Trend

Approximately an order of magnitude increase in bandwidth every 5 years

How to effectively utilize the increased bandwidth?
Current Interconnection Networks

• Earlier works [Dally ’90, Agarwal ’91] suggested “lower-radix” networks
  – Examples: Torus Routing Chip, Cray T3E, SGI Altix 3000
• Current Routers
  – Myrinet : radix-32
  – Quadrics : radix-8
  – IBM SP2 Switch : radix-8

• Technology trends
  – Increasing Bandwidth
  – Optical signaling

High-Radix Routers can take better advantage of these technology trends.
Outline

- Technology Trends for High-Radix Router
- Motivation for High-Radix Router
- High-Radix Router Architectures
  - Baseline Architecture
  - Fully Buffered Crossbar
  -Hierarchical Crossbar
- Conclusion & Future Work
High-Radix Router
High-Radix Router

Low-radix (small number of fat ports) High-radix (large number of skinny ports)
Latency in a Network

Latency = Header Latency + Serialization Latency

= $H \frac{t_r}{b} + \frac{L}{b}$

= $2t_r \log_k N + 2kL / B$

where $k = \text{radix}$

$B = \text{total Bandwidth}$

$N = \# \text{ of nodes}$

$L = \text{message size}$
Latency vs. Radix

- Header latency decreases
- Serialization latency increases

Optimal radix $\sim 40$

Optimal radix $\sim 128$
Determining Optimal Radix

Latency = Header Latency + Serialization Latency

= $H t_r + L / b$

= $2t_r \log_k N + 2kL / B$

where $k =$ radix
$B =$ total Bandwidth
$N =$ # of nodes
$L =$ message size

Optimal radix

$\Rightarrow k \log_2 k = (B t_r \log N) / L$

= Aspect Ratio
Higher Aspect Ratio, Higher Optimal Radix
Higher Radix, Lower Cost
Outline

- Technology Trend for High-Radix Router
- Motivation for High-Radix Router
- High-Radix Router Architectures
  - Baseline Architecture
  - Fully Buffered Crossbar
  - Hierarchical Crossbar
- Conclusion & Future Work
Virtual Channel Router Architecture
High-Radix Switch Architectures (I)

(a) Baseline design
Simulation Methodology

- Open-loop simulation done with steady-state measurement
- Output switch arbitration required two cycles
- Wire delay included in the pipeline
- Uniform random traffic pattern
- Each packet was assumed to be a single flit
- Radix=64 router evaluated with 4 VCs per input

- Metrics – latency & throughput
- Cost – area
Baseline Performance Evaluation

![Graph showing latency vs. offered load for low-radix routers.](image)
High-Radix Routers

Baseline Performance Evaluation

![Graph showing latency vs offered load for low-radix and baseline (high-radix) routers. The graph indicates that low-radix routers perform better than baseline (high-radix) routers.]

Latency (cycles)

Low radix better
High-Radix Switch Architectures (II)

(a) Baseline design

(b) Fully buffered crossbar

Fully buffered crossbar with shared buffers
Fully Buffered Crossbar Provides High Performance but ....
... Becomes Costly to build

Buffer area dominates at radix 50

Other issues:
1. Credit Information
2. Adaptive Routing
High-Radix Switch Architectures (III)

(a) Baseline design

(b) Fully buffered crossbar

(c) Hierarchical crossbar
Hierarchical Crossbar Performance on Uniform Random Traffic

- baseline
- subswitch 32
- subswitch 16
- subswitch 8
- subswitch 4
- fully-buffered

Graph showing latency (cycles) vs. offered load for different subswitch configurations.
Worst-case Traffic Pattern
Worst Case Performance Comparison

![Graph showing latency vs. offered load for different configurations: baseline, subswitch 32, subswitch 16, subswitch 8, subswitch 4, and fully-buffered. The graph plots latency (cycles) on the y-axis against offered load on the x-axis.]
4k Node Network Performance

High-Radix leads to lower zero-load latency
High-Radix Router Microarchitecture

- In building high radix router, there are scaling issues as the number of ports increase
- Baseline design provides poor performance
- Fully buffered crossbar leads to high performance but a costly design
- Hierarchical crossbar provides a feasible design with minimal loss in performance
Conclusion

- Performance of digital systems is often limited by the interconnection network.
- Increasing on-chip bandwidth can be more efficiently used by a high-radix routers.
- High-radix lead to lower cost and lower latency networks.
- A high-radix router requires a different architecture.
- A hierarchical organization makes high-radix routers feasible.
Future Work

- Optimal topology for high-radix network
  - Clos topology suited for high-radix routers
- Routing - How to adaptively route in a high-radix router?
- Flow control with high-radix router
- Simulating a larger network
- Microarchitecture - Alternative to crossbars: multistage switch organizations, network-on-chip (torus, mesh, etc.)
Thank you

Questions?