Scalable Collective Communication on the ASCI Q Machine

Fabrizio Petrini, Juan Fernandez, Eitan Frachtenberg and Salvador Coll
fabrizio@lanl.gov

http://www.c3.lanl.gov/~fabrizio

Performance and Architecture Laboratory

CCS-3 Modeling, Algorithms, and Informatics Group

Los Alamos National Laboratory

Outline

- Overview of the ASCI Q Machine
- Quadrics network: building blocks and topology
- Network topology of ASCI Q
- Network-based algorithms to perform collective communication
- Hardware support for collective communication
- Performance and scalability results of the most common collective communication operations (barrier, broadcast, allreduce and hot spot) on a 1024-node segment of the Q machine

ASCI Q

ASCI Q

- 2048 4-processor AlphaServer Es45s
- 8192 Alphas EV68
- 2 independent network rails, Quadrics
- 4096 Quadrics Elan NICs
- $>$ 9000 network cables
- 2^{nd} largest machine in the Top500

Contribution

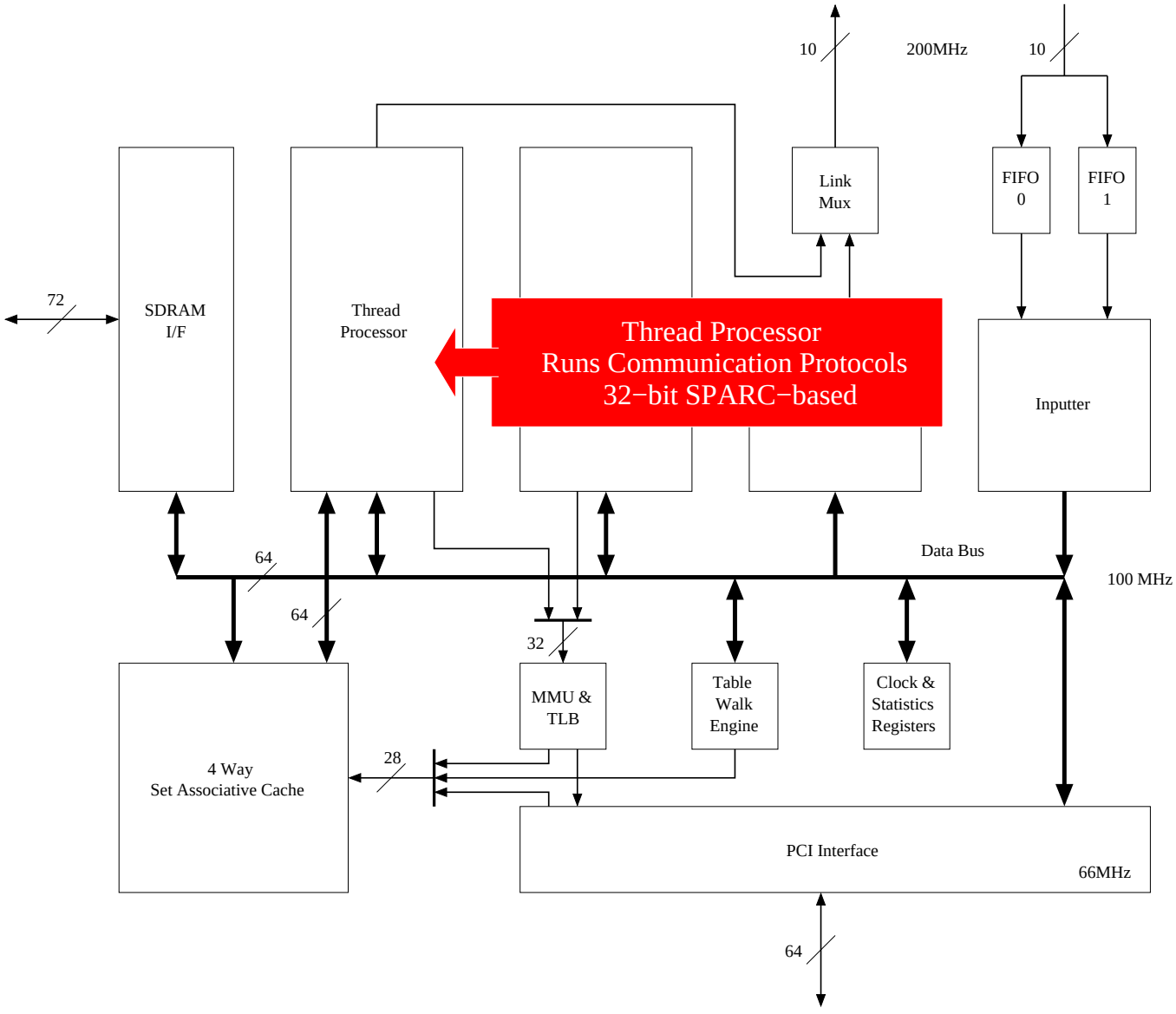
- Describe the network topology and some design choices of a ASCI-class parallel machine
- Provide experimental results on a large configuration (1024 nodes/4096 processors)

Quadrics Network Overview

The Quadrics netwok is based on two building blocks:

- a network interface card called **Elan**
- a crossbar switch called **Elite**

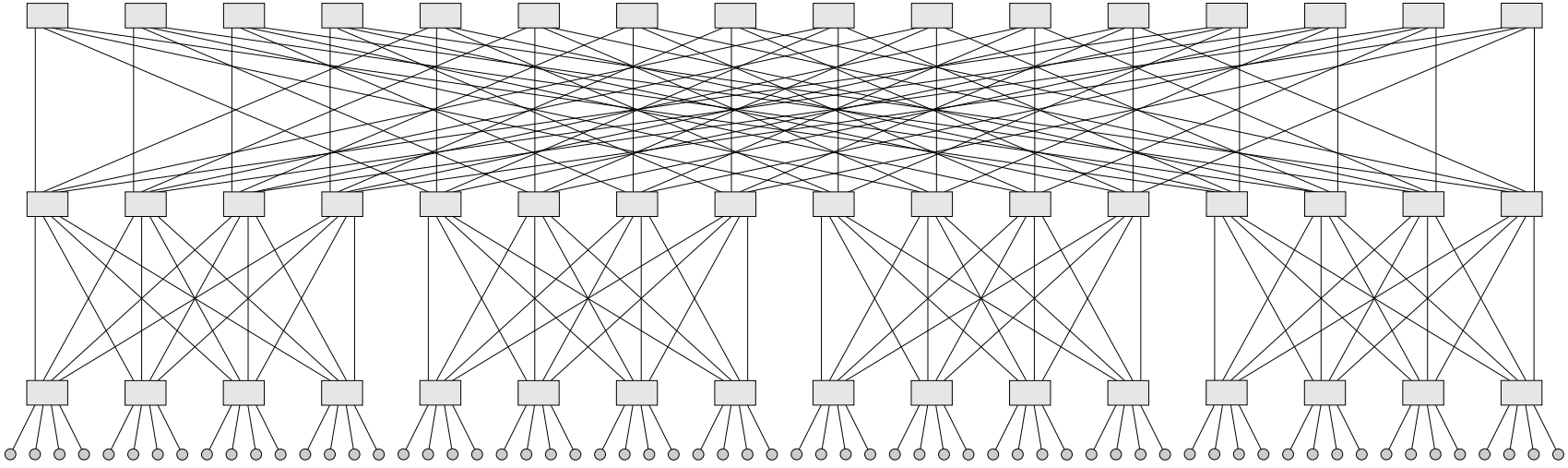
Quadrics Network: Elan



Quadrics Network: Elite

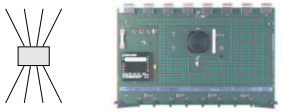
- 8 bidirectional links with 2 virtual channels in each direction
- An internal 16x8 full crossbar switch
- 400 MB/s on each link direction
- 2 priority levels plus an aging mechanism
- Adaptive routing
- Hardware support for broadcast

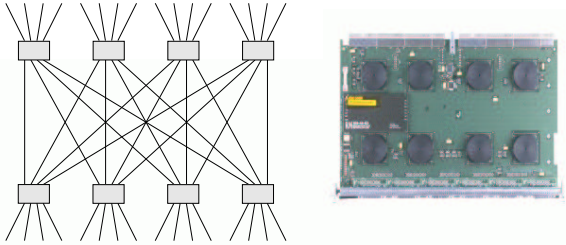
Logical Topology: Quaternary fat-tree

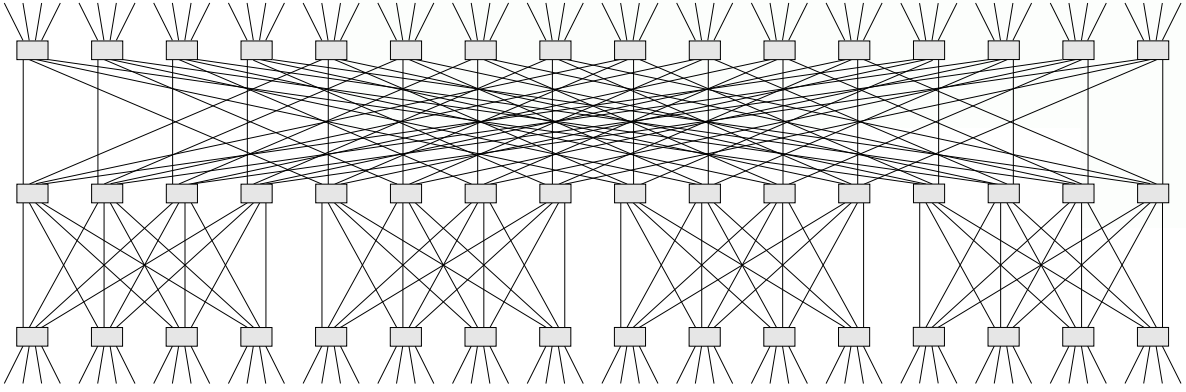


- Elans and Elites are connected in a fat-tree topology

Network Building Blocks



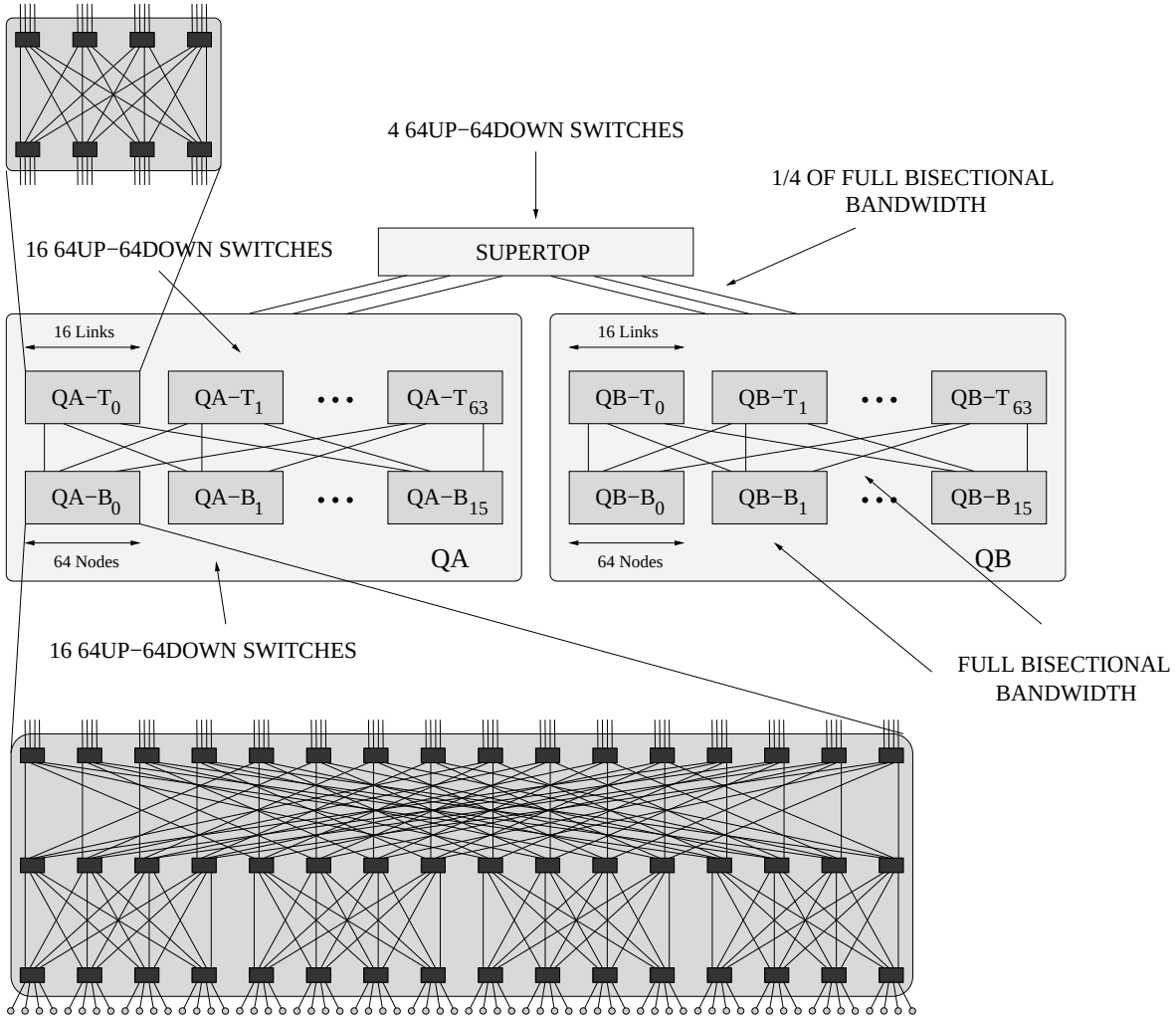




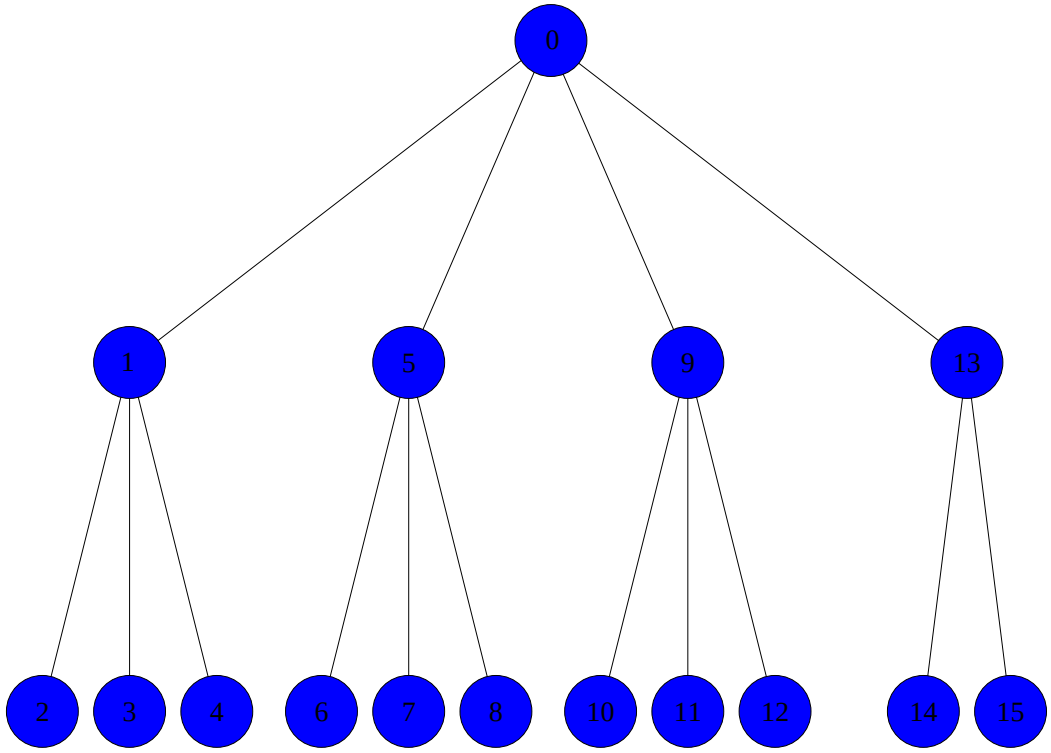
Three main building blocks:

- Single Elite (backplane)
- 16 up/16 down (level 2 fat-tree)
- 64 up/64 down (level 3 fat-tree)

Network Topology of ASCI Q

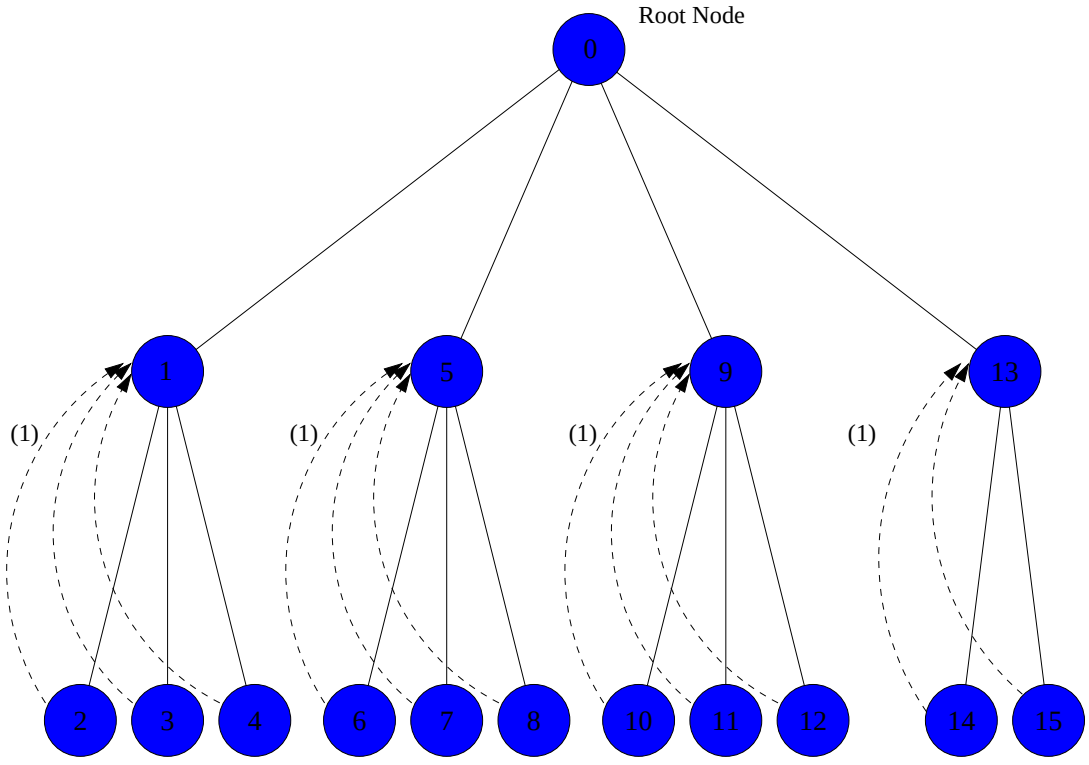


Software-Based Barrier



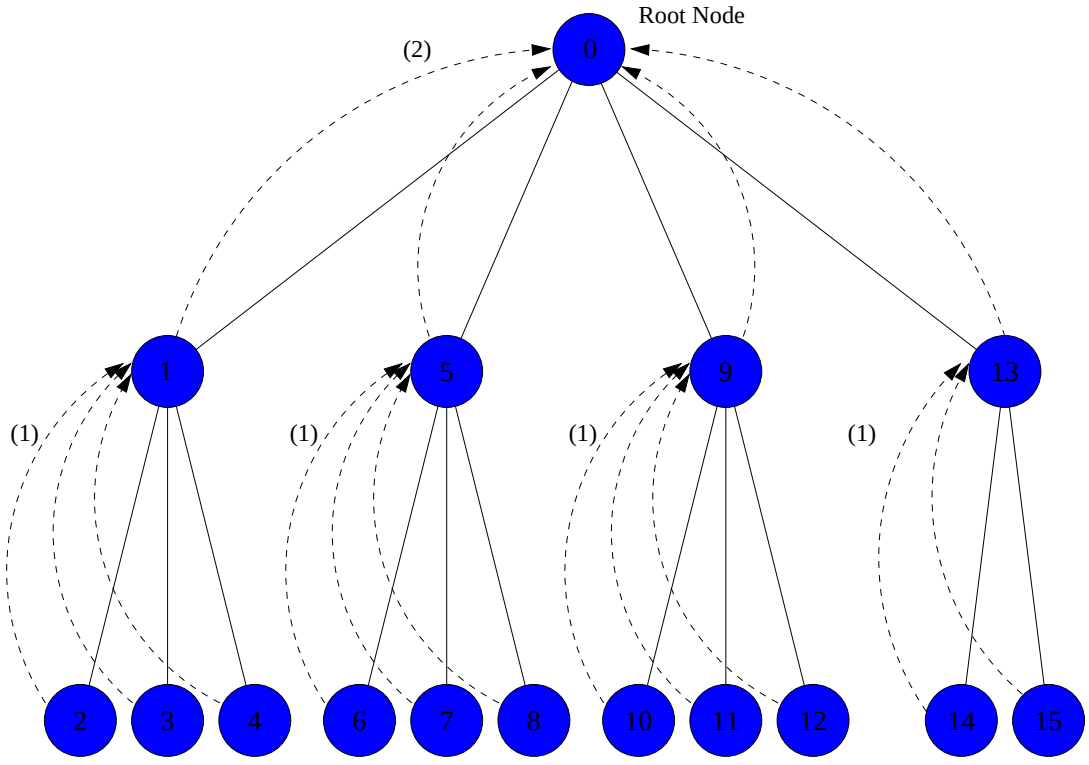
- The software-based barrier is executed is using point to point messages
- These messages are sent from Elan to Elan, without interrupting the processing node

Software-Based Barrier



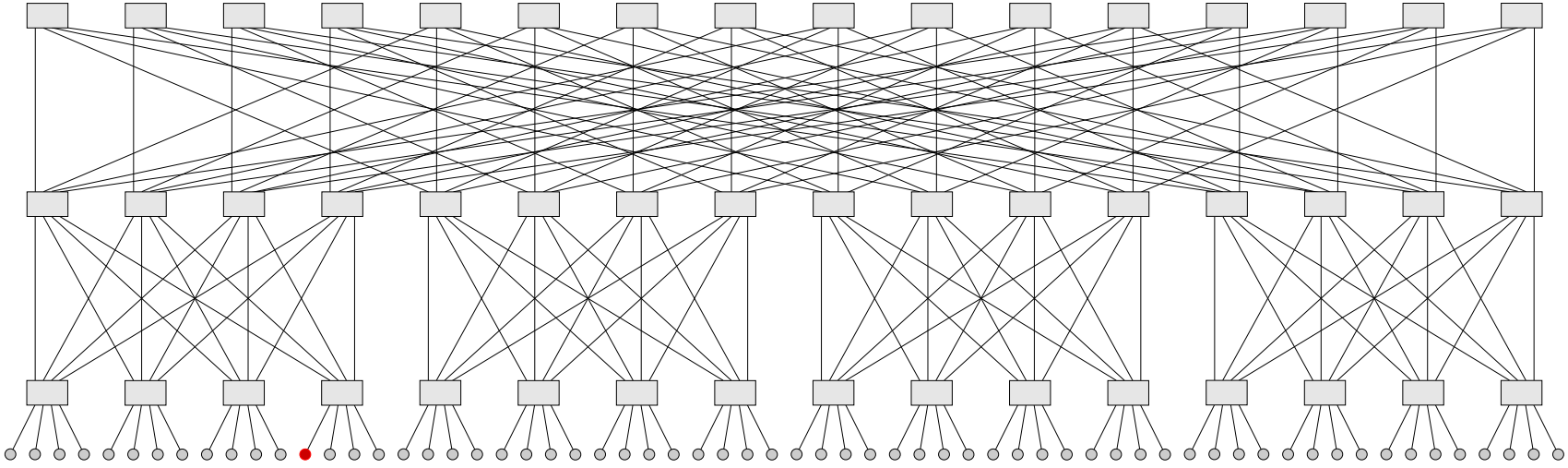
- Each Elan Network Interface waits for ’ready’ signals from its children (1) ...

Software-Based Barrier



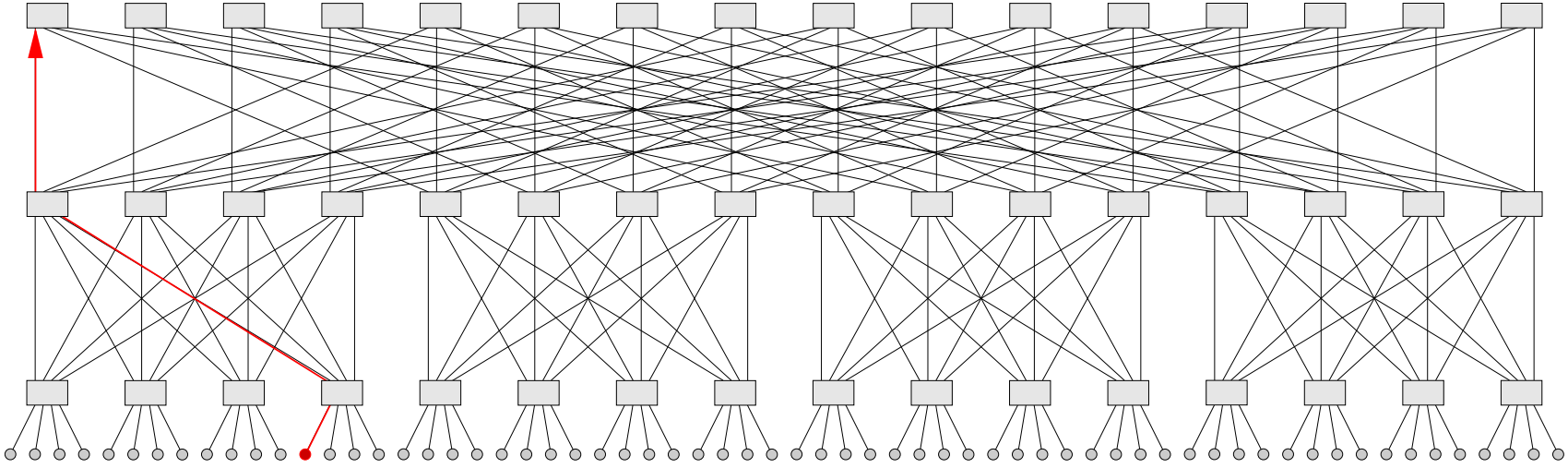
- ... and sends its own signal up to the parent process (2)

Hardware-Based Barrier



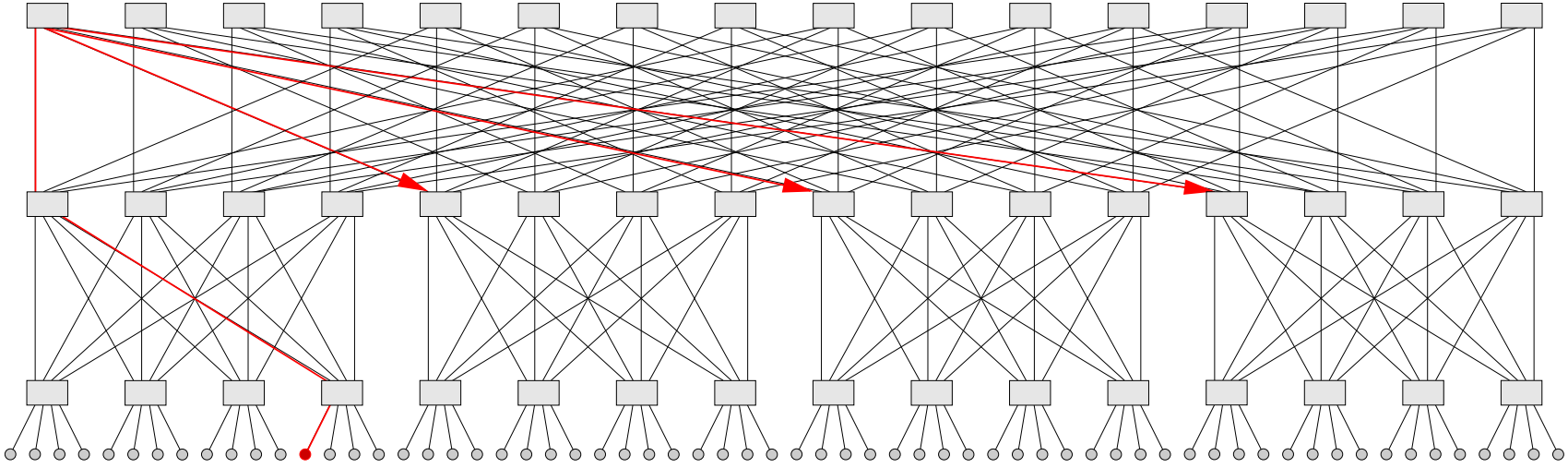
- The root node sends a multicast packet

Hardware-Based Barrier



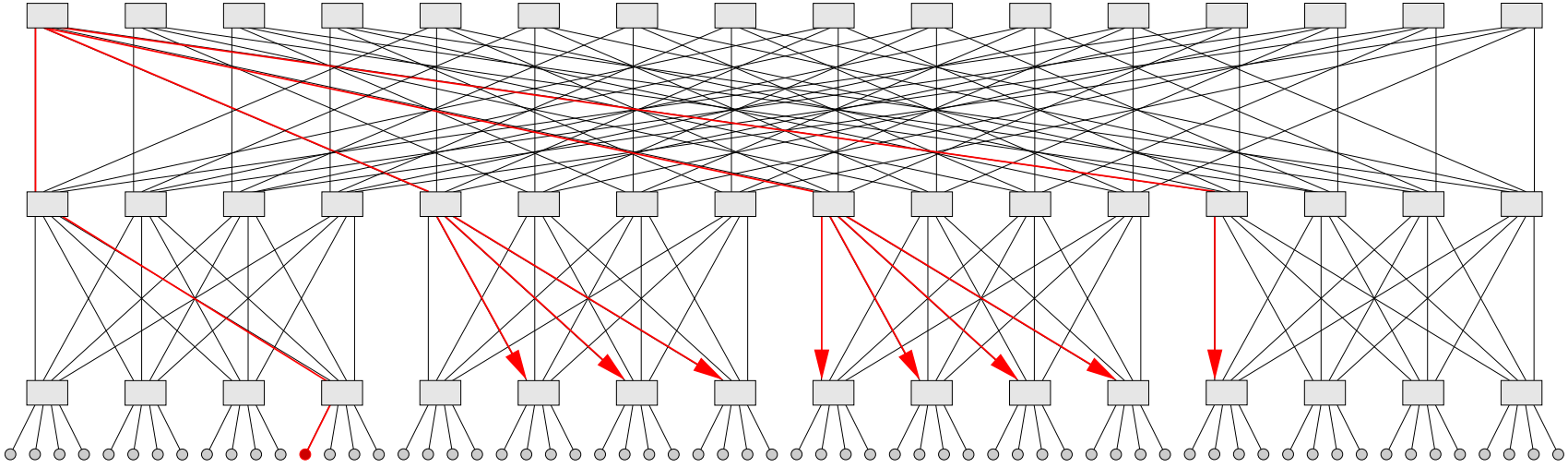
- The packet reaches the top of the tree

Hardware-Based Barrier



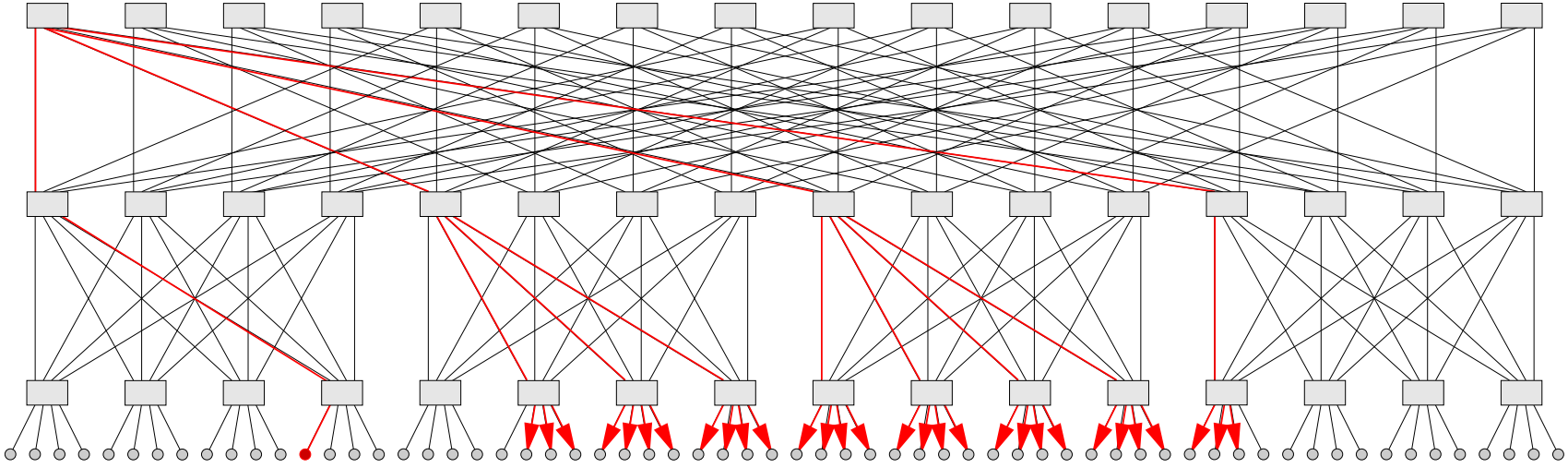
- The packet is multicast down the logical tree

Hardware-Based Barrier



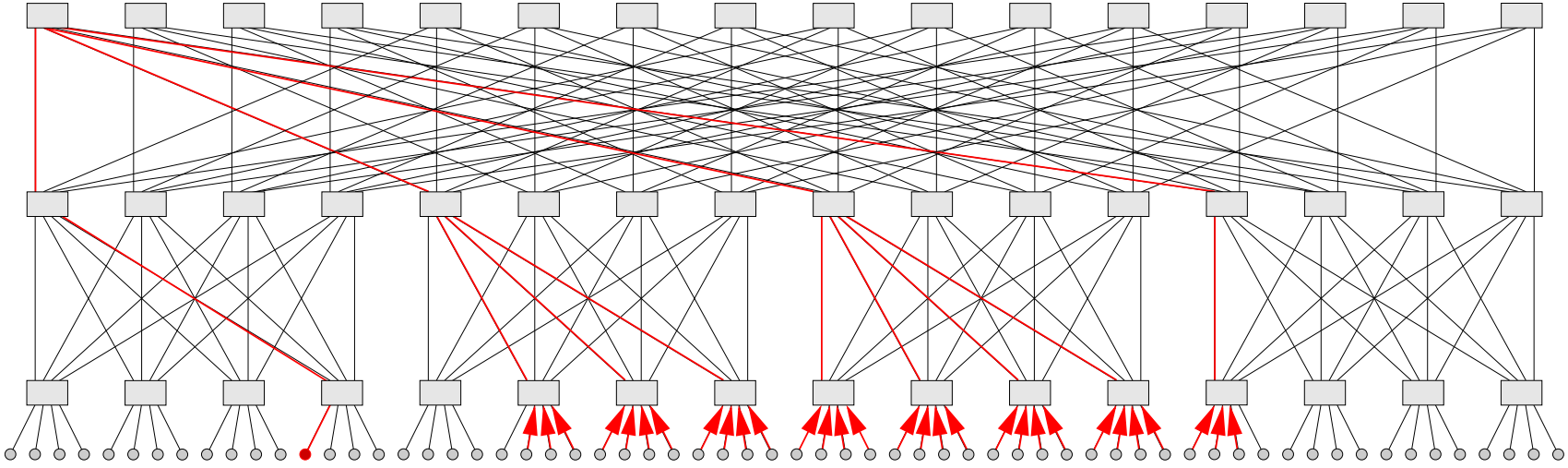
- The packet is multicast down the logical tree

Hardware-Based Barrier



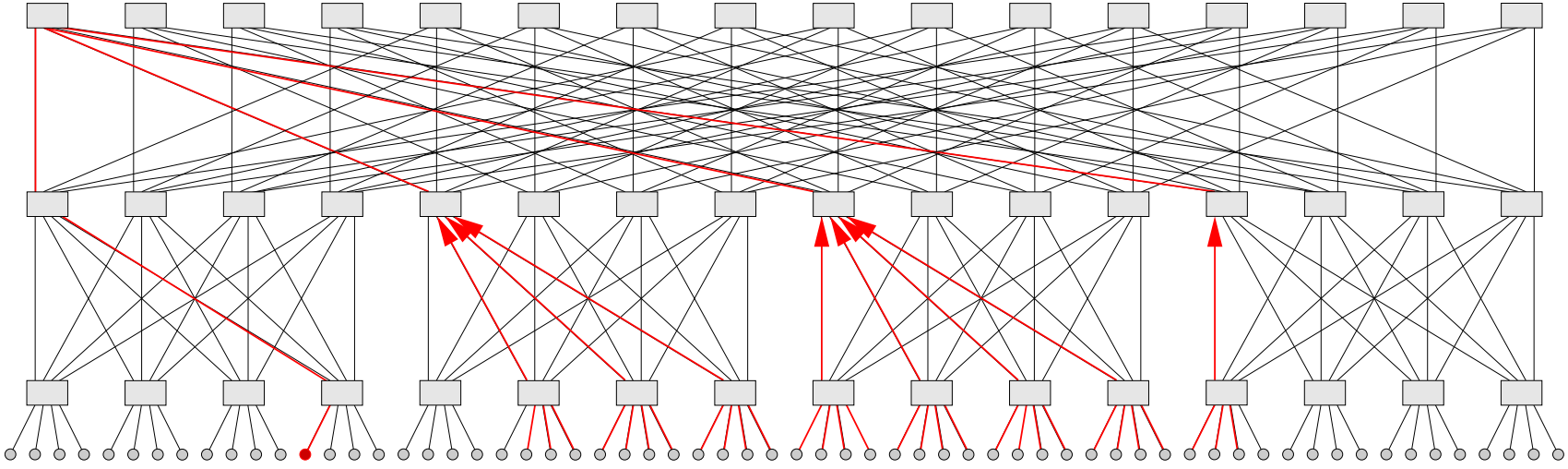
- The packet is multicast down the logical tree

Hardware-Based Barrier

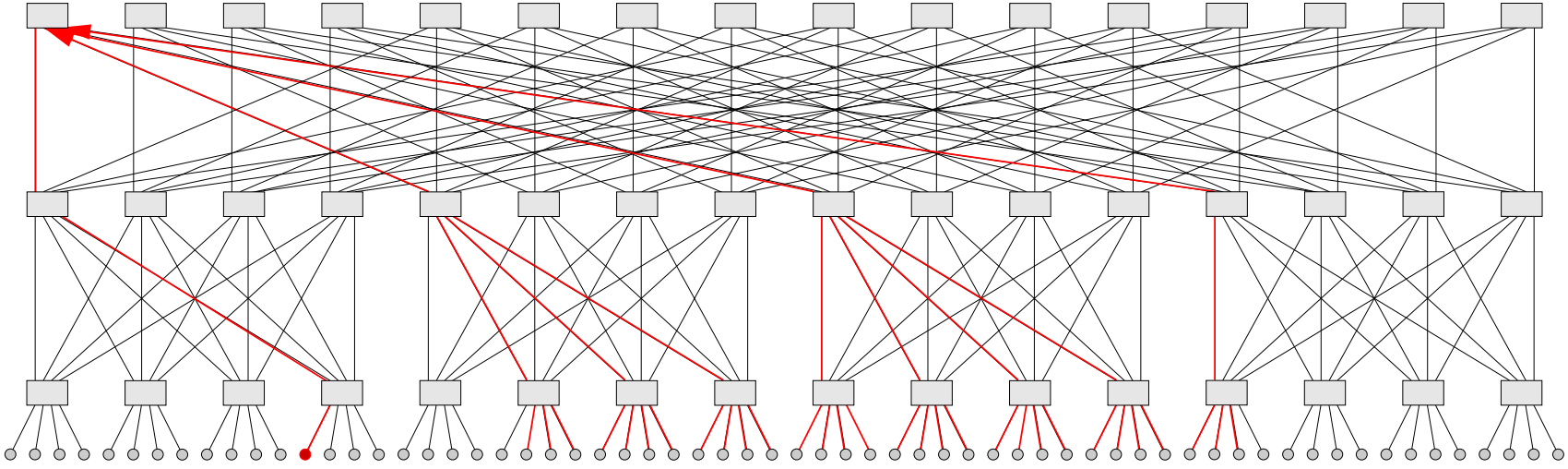


- The results of the collective operation are combined and sent back to the root.
- The tree of circuits is active during the whole collective communication.

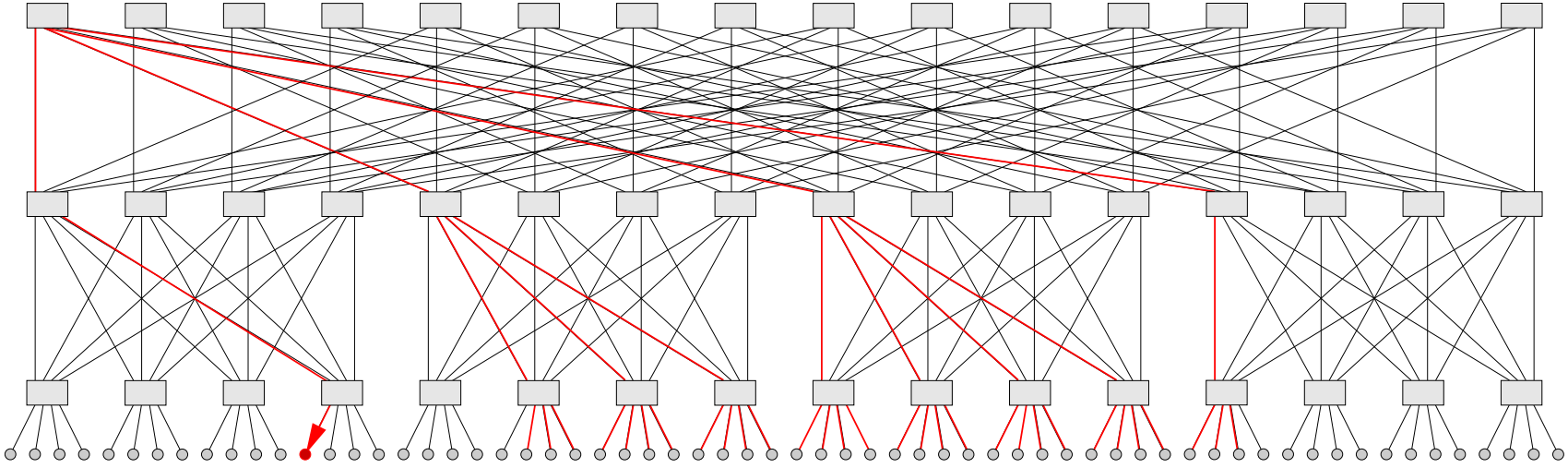
Hardware-Based Barrier



Hardware-Based Barrier



Hardware-Based Barrier

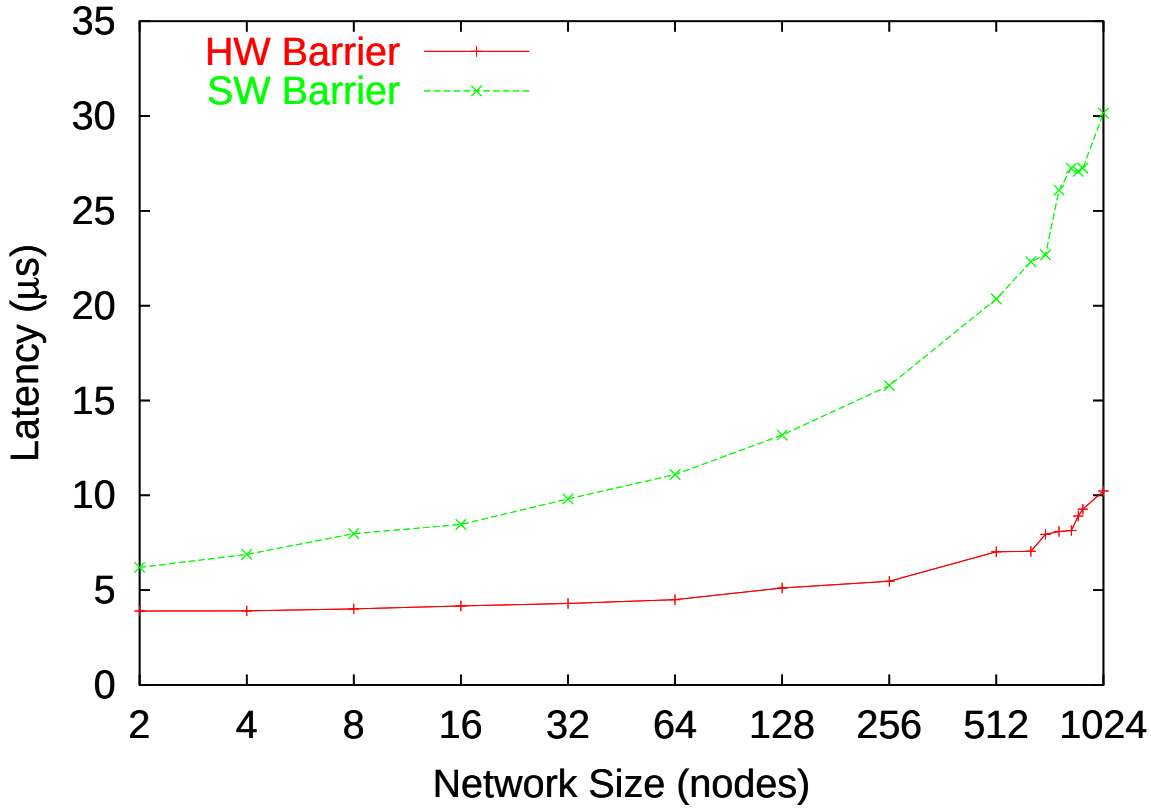


- The final result reaches the root
- The whole collective communication is atomic

Performance and Scalability

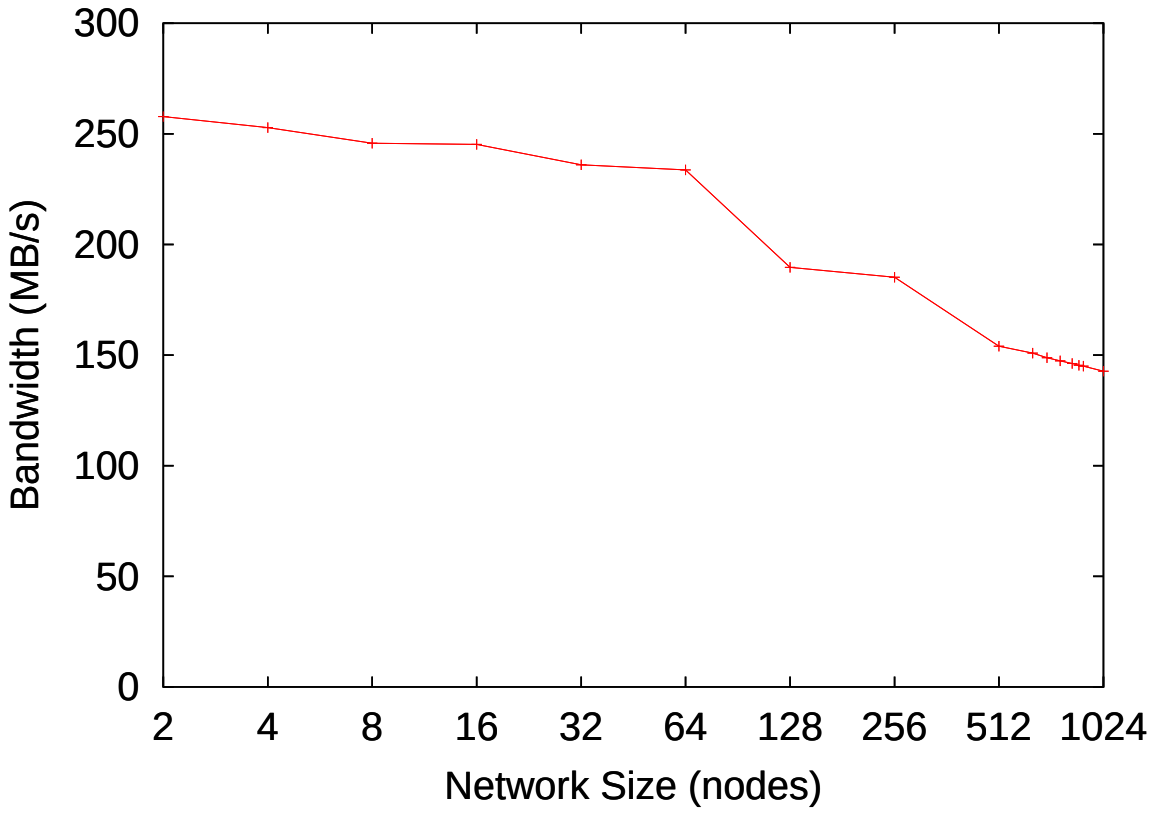
- We report performance and scalability results of four common collective communication patterns on a 1024-node segment of the Q machine
  - Barrier Synchronization
  - Broadcast (one to all)
  - Hot-spot (all to one)
  - Allreduce (many to one)

Barrier Synchronization



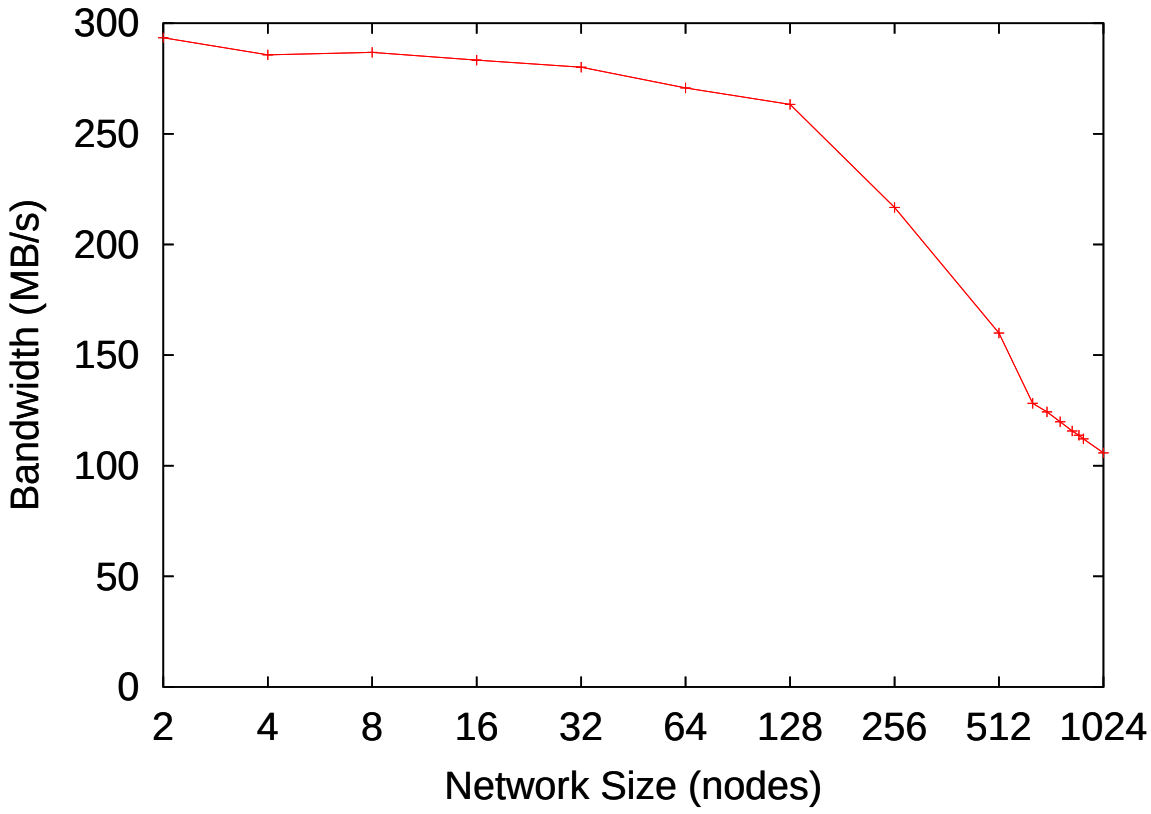
The hardware-based barrier can synchronize 4096 processors in less than 10 μs.

Broadcast



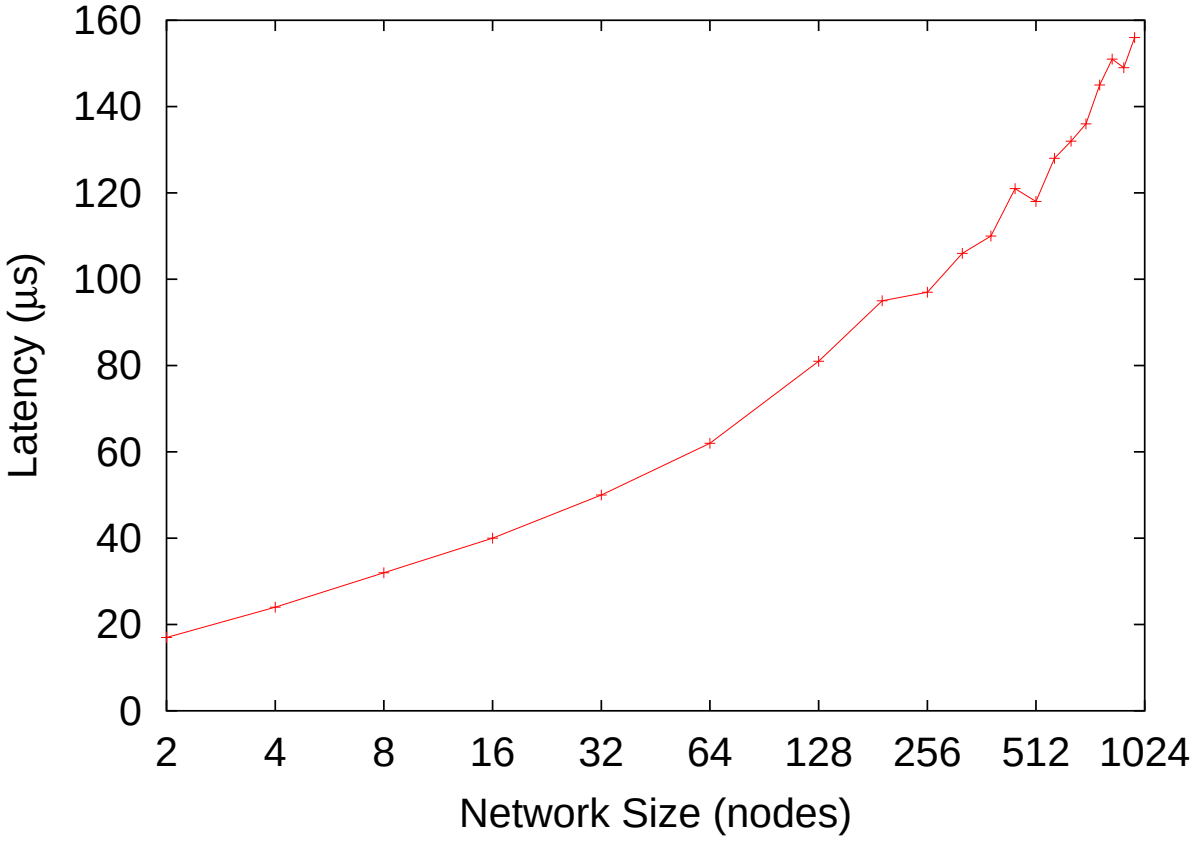
The aggregate bandwidth delivered by the broadcast is > 140 GB/sec

Hot Spot



The performance degradation for large processor counts is caused by the end-to-end flow control (circuit switched, maximum packet size 320 bytes)

Allreduce


Latency (µs)
0
20
40
60
80
100
120
140
160
2
4
8
16
32
64
128
256
512
1024
Network Size (nodes)

Conclusions

- We described the network topology of the ASCI Q machine.
- We presented an overview of both software- and hardware-based collective communication algorithms on the Quadrics network
- We also presented some scalability and performance results of four collective primitives, barrier, broadcast, hot spot and allreduce on a 1024-node segment of the Q machine

Resources

More information can be found at the following URLs:

Quadrics network

http://www.quadrics.com

http://www.c3.lanl.gov/~fabrizio/publications.html