The Quadrics Network (QsNET): High-Performance Clustering Technology

Fabrizio Petrini, Wu-chun Feng, Adolfy Hoisie,
Salvador Coll and Eitan Frachtenberg

CCS Group

Los Alamos National Laboratory

Resources

- More information can be found at

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

- Quadrics web site

http://www.quadrics.com

- Or sending an e-mail to

fabrizio@lanl.gov

Outline

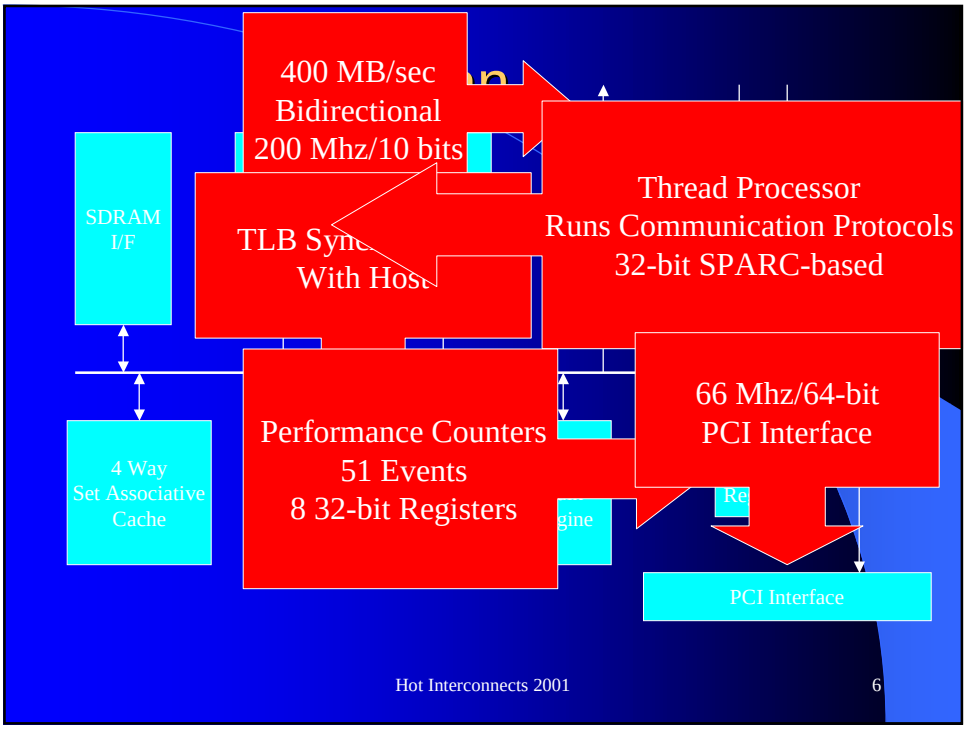
- Introduction
- Quadrics network design
  - Elan
  - Elite
  - Fat-trees, topological properties, routing algorithm, flow control
- Communication/Programming libraries
- Performance Analysis
  - Experiment description
  - Results

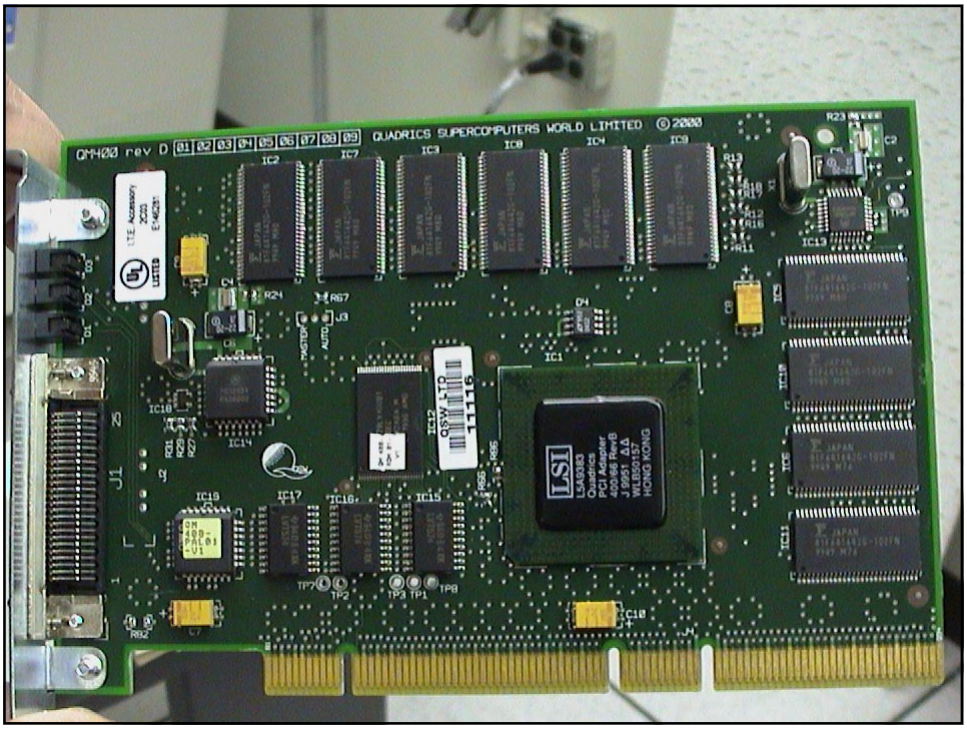
Network Overview

- QsNET provides an abstraction of distributed virtual shared memory
- Each process can map a portion of its address space into the global memory
- These address spaces constitute the virtual shared memory
- This shared memory is fully integrated with the native operating system

Building Blocks

- The QsNET is based on two building blocks
  - A network interface card called Elan
  - A crossbar switch called Elite

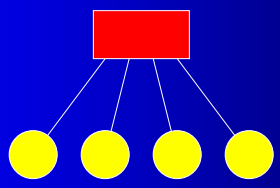




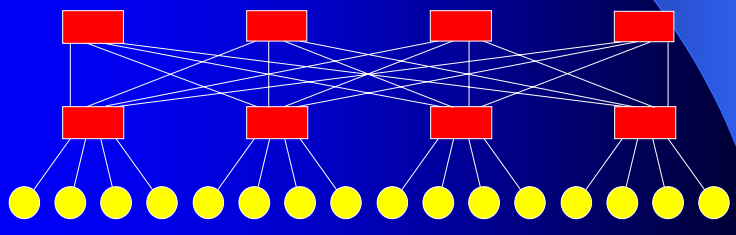
Elite

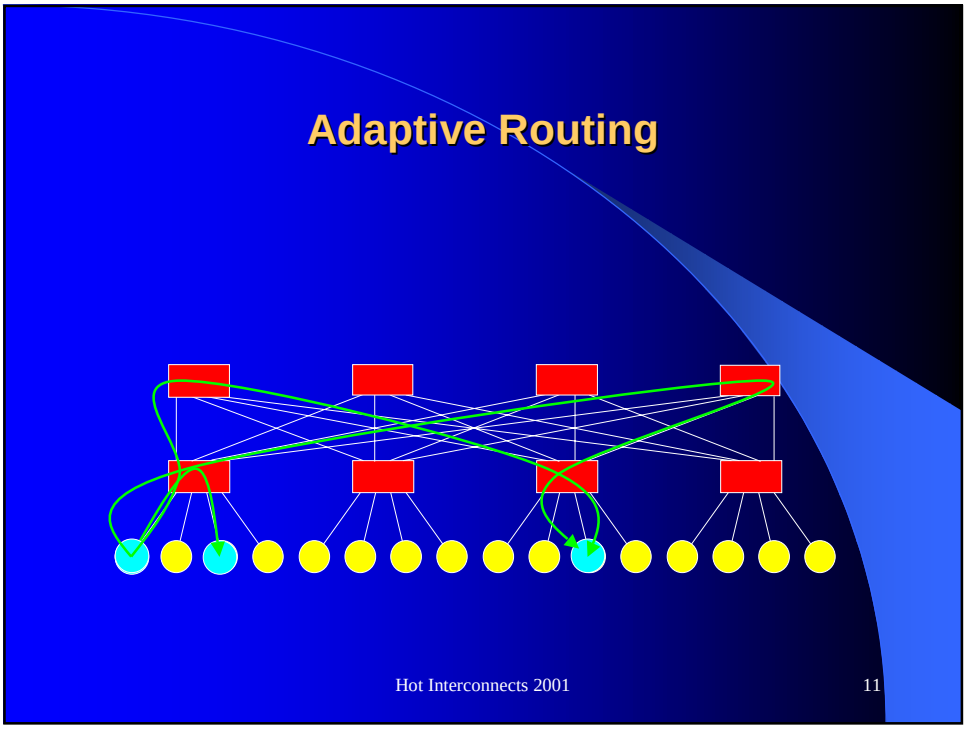
- 8 bidirectional links with 2 virtual channels in each direction
- An internal 16x8 full crossbar switch
- 400 MB/sec on each link direction
- Packet error detection and recovery, with routing and data transactions CRC protected
- 2 priority levels
- Hardware support for broadcast
- Adaptive routing

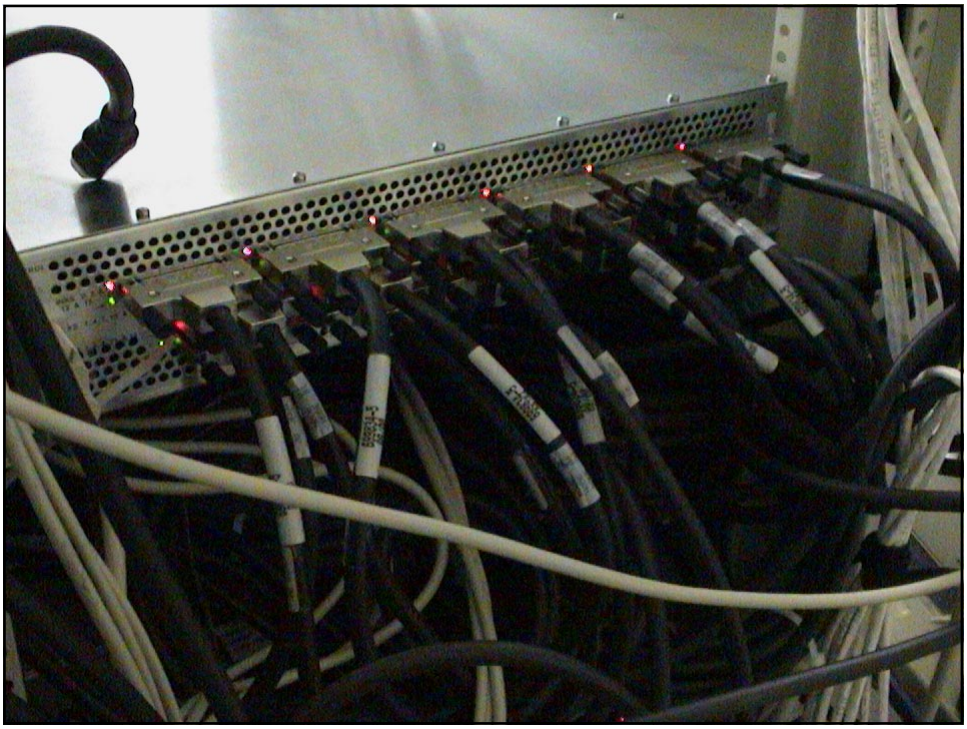
A Quaternary Fat Tree



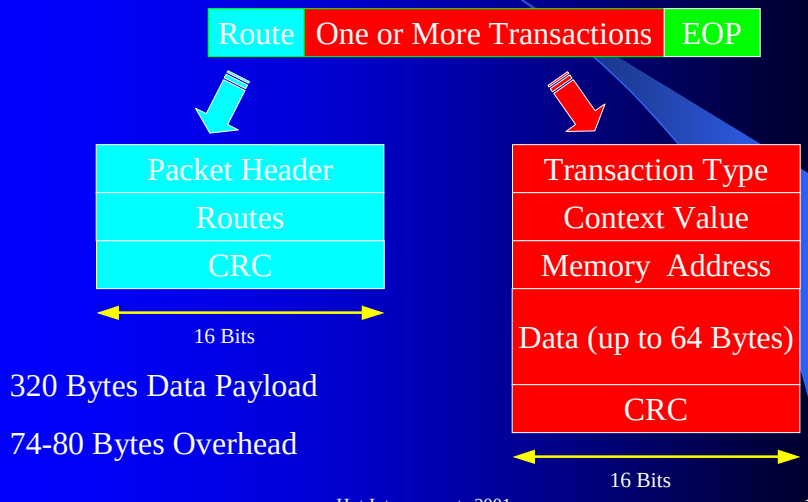
Fat Tree Recursive Topology




Adaptive Routing
Hot Interconnects 2001
11



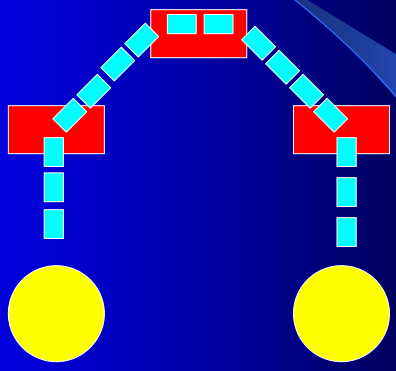
Packet Format



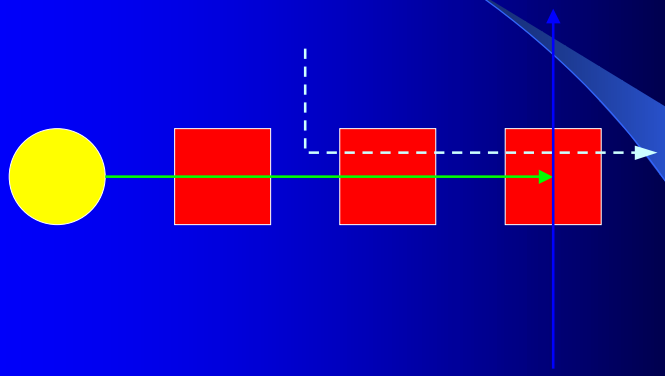
320 Bytes Data Payload

74-80 Bytes Overhead

Wormhole Flow-Control



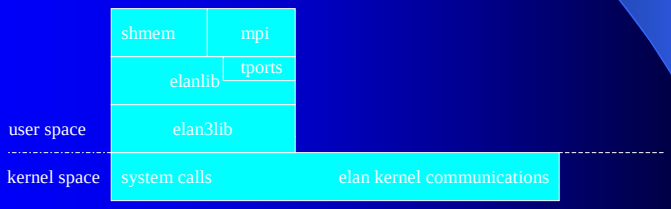
Virtual Channels



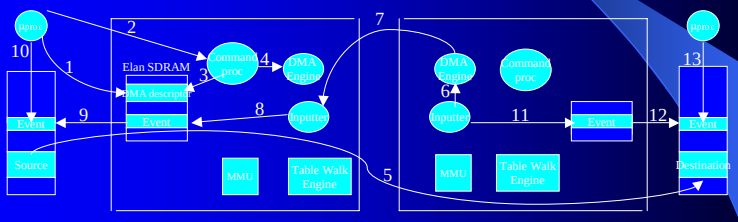
Programming Libraries

- Elan3lib
  - event notification
  - memory mapping and allocation
  - remote DMA
- Elanlib and Tports
- MPI

User Applications



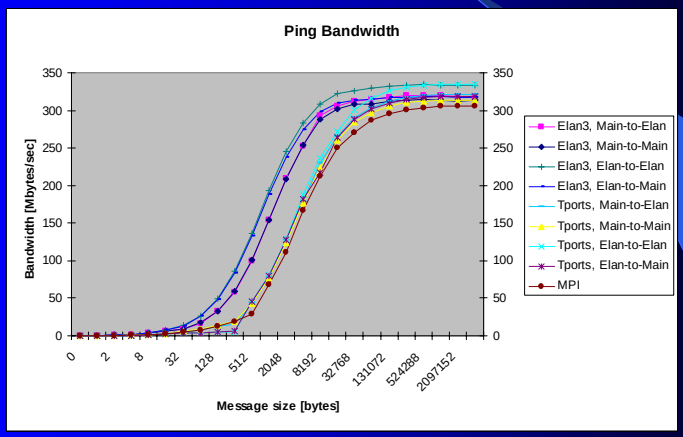
Execution of a Remote DMA



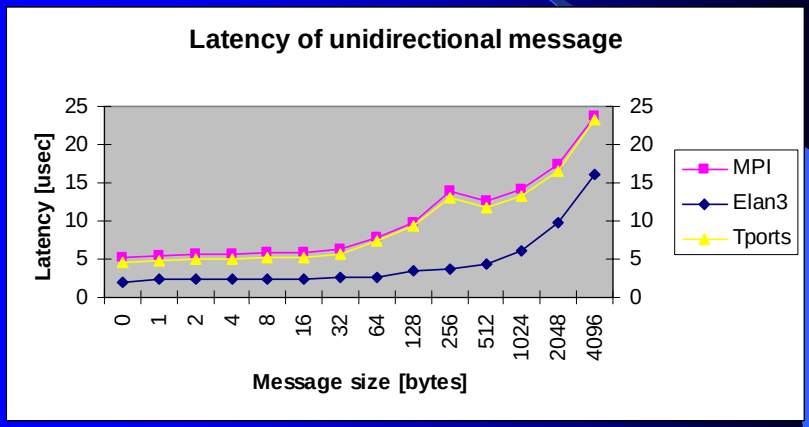
Experimental Framework

- Cluster of 16 dual-processor SMPs
- Intel 733 Mhz Pentium III
- Motherboard: based on ServerWorks HE chipset which provides 2-64 bit 66Mhz PCI slots
- Each SMP is equipped with one Elan
- One 16-port Elite Switch is use for interconnection

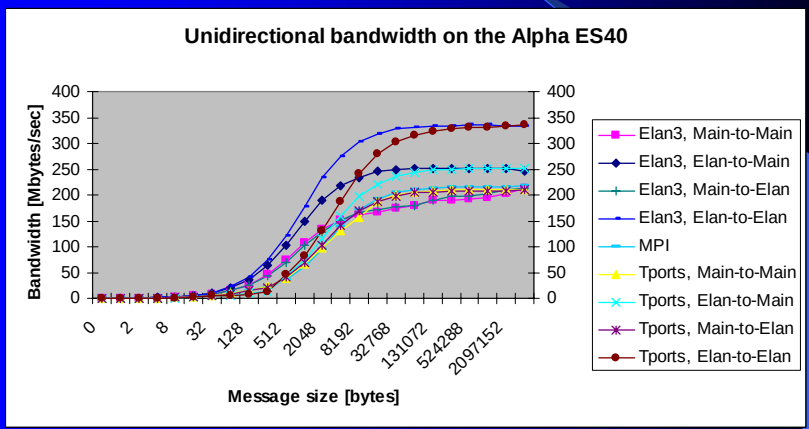
Unidirectional



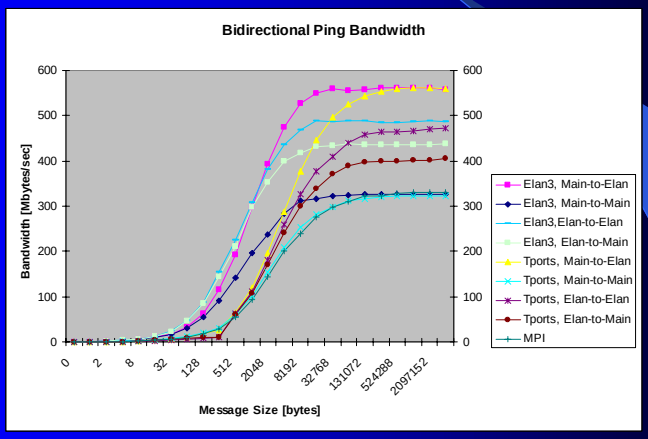
Latency Unidirectional



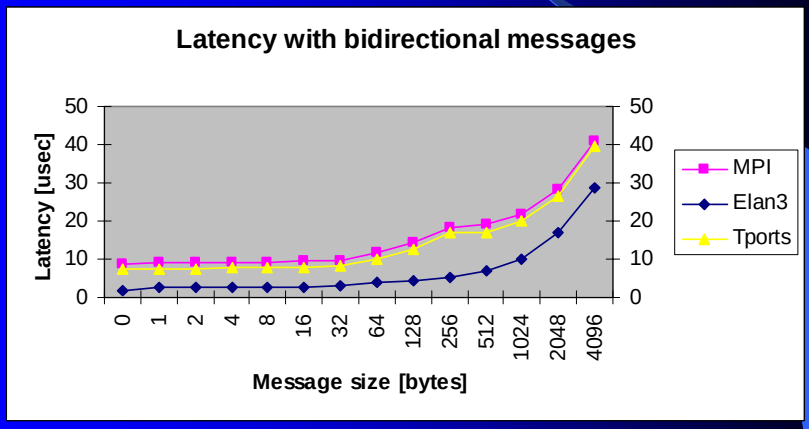
Unidirectional bandwidth on the ES40



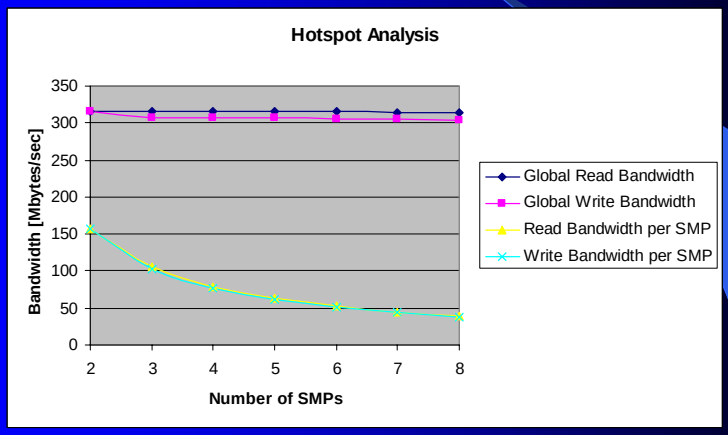
Bidirectional



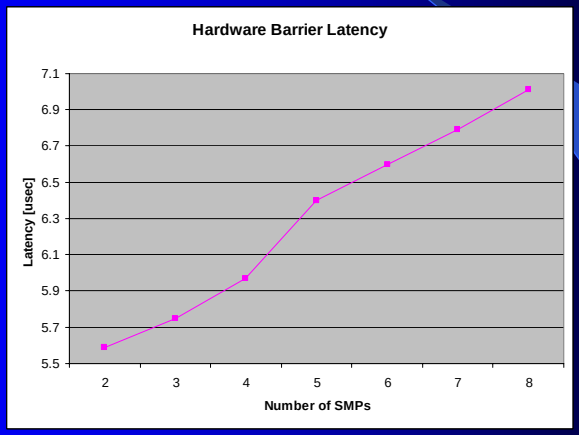
Latency Bidirectional



Hotspot



Hardware Barrier



Conclusion

- QsNet provides the abstraction of a distributed virtual shared memory with protected and fault-tolerant communication
- Latency is as low as 2µs and bandwidth as high as 307 MB/s
- Network Bandwidth between Elan buffers is 335 MB/s
- Performance degradation in the presence of bi-directional traffic, due to the current PCI bus implementations