The Quadrics Network

High Performance Clustering Technology

Fabrizio Petrini
Wu-chun Feng
Adolfy Hoisie
Salvador Coll
Eitan Frachtenburg
Los Alamos National Laboratory

What is the Quadrics Network?

It's a distributed virtual memory system

Featuring:

- Single global virtual address space
- Fault tolerance

Presentation Outline

- Quadrics network anatomy
- Implementation of global virtual memory
- Libraries
- Experimental results

Virtual Memory

- Every process has virtual address space
- On each memory access:
  - Processor reads process's page table
  - Page table converts virtual address to physical address
  - Memory access performed with physical address

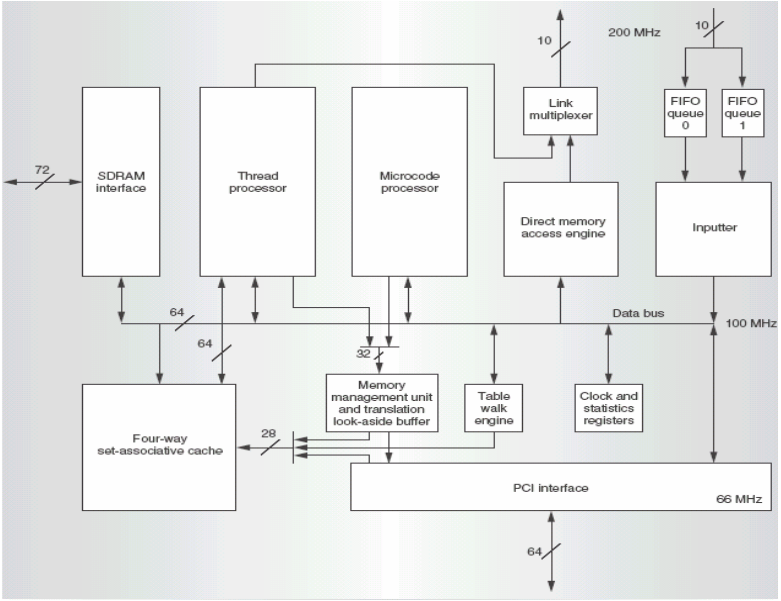
Global Virtual Memory

- Virtual memory space spans all processors
- Process can access page located on remote memory transparently

The Quadrics Network

- “Elan” network interface - PCI card in each node
- “Elite” communication switches
- Multiple communication libraries
  - allow custom protocols
  - library trade-off: performance vs. ease-of-use

Elan Network Interface



Elan's microcode processor

Handles memory requests.

- 4 threads:
  - inputter
  - DMA engine
  - processor scheduling
  - command processing
- 2 stage pipeline / thread => 8 outstanding memory requests

Elan components

Thread processor:

- Implements messaging libraries
- 32 bit RISC + extra specialized instructions

MMU:

- Converts 32 bit virtual address -->
  - 28 bit SDRAM physical address, or
  - 48 bit PCI address
- 16 entry TLB

Elan components

Routing table

- Virtual process # --> tags to determine route

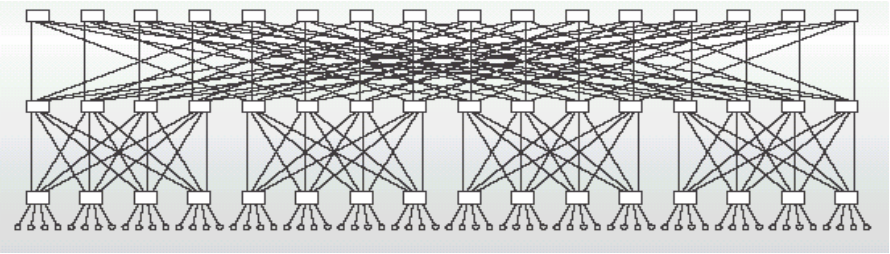
64MB SDRAM

8k Cache for SDRAM

Link logic

- 2 virtual channels
- 128 entry FIFO buffer

Quaternary Fat-Tree Network



- 4-ary n-tree (n=3 above) - each switch connects to 4 switches
- comprised of “Elite” crossbar switches

Elite switch

- 8 bi-directional links:
  - 2 virtual channels in each direction
- 400 MB/s bandwidth
- 35 ns latency
- CRC error detection between links
- 2 priority levels

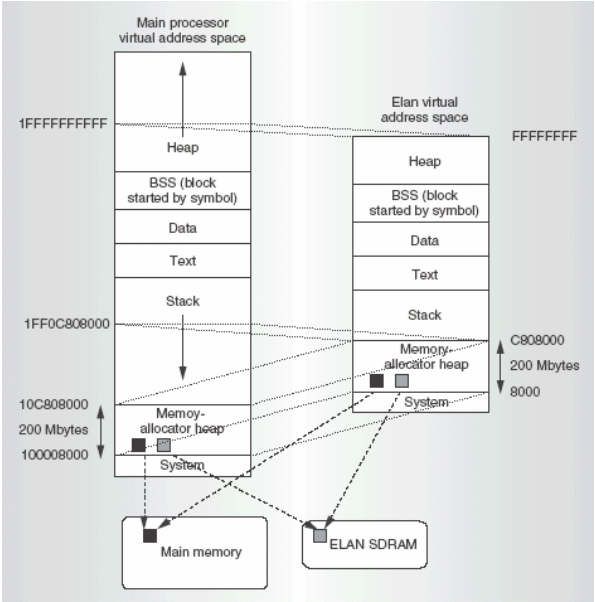
Routing

- Elan router puts tag sequence in header
- Elite switch removes first tag, routes to next switch
- At data link level:
  - Elite partitions packet into “flits”
  - flits sent independently
  - after last flit in packet, receiver sends ACK

Elan virtual memory

- MMU converts virtual-->physical
  - can translate between architectures
- Physical data can be on Elan SDRAM or on local memory
- Location of physical data not normally visible to users

Virtual address translation



Virtual Memory Extension

Extension to virtual memory: *virtual operation*

- Cooperating processes can transfer data between address spaces
- Protection still maintained

Context

- Virtual process id replaced with *context*
  - context + virtual address identify page
- Multiple processes (on multiple machines) can have same context
  - allows for distributed shared memory

Fault tolerance

- Fault tolerance steps:
  - Packet consists of route info + transactions
  - Last transaction contains ACK Now flag
  - Packet not successful until receiver sends ACK
  - Link reused only after ACK received
- After fixed # of errors, new route negotiated

Programming Libraries

- Allows programmer to write intelligent protocols
- Elan3lib
  - Low-level, high efficiency
  - Allows user to program Elan, move data manually between Elan memory & local memory (w/o operating system knowing)
- Elanlib
  - Higher-level, lower efficiency
  - Allows MPI-like message passing

Experimental Methodology

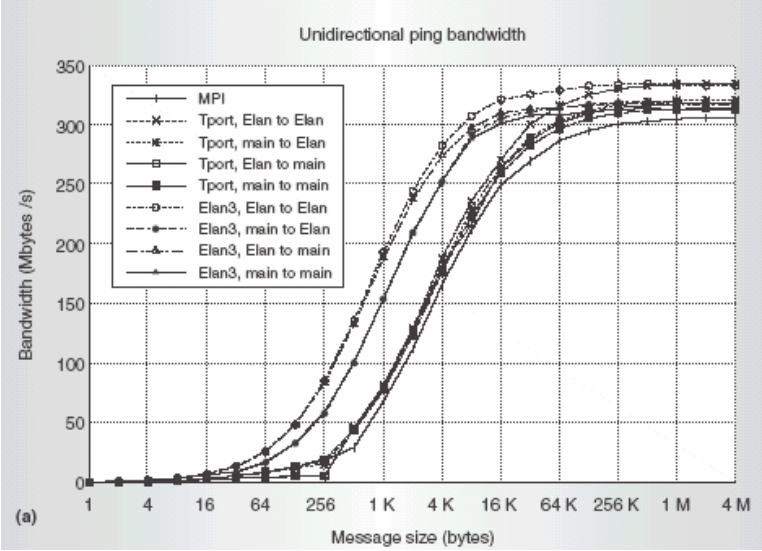
Setup:

- 16 dual processor 733 MHz Pentium III's
  - 1 GB RAM
  - 64 bit, 66 MHz PCI slot for Elan card
- Quaternary 2-dimensional fat tree network
- Linux 2.4.0-test7 operating system

Benchmarks:

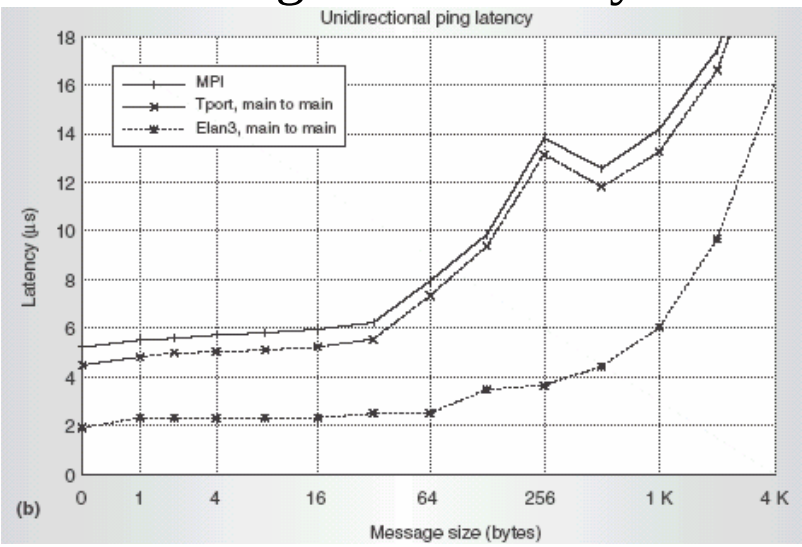
- Elan3lib benchmark to show best performance
- Elanlib benchmark to simulate MPI-2

Ping test - Bandwidth



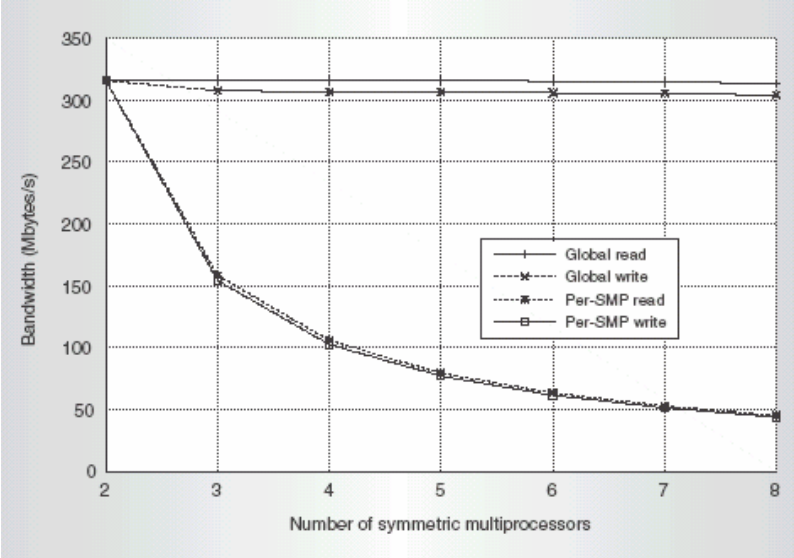
Bandwidth lies between 307 MB/s for MPI to 335 MB/s for Elan3lib

Ping test - Latency



Latency lies between 5.0 us for MPI to 2.4 us for Elan3lib *up to 64 bytes*

Scalability - Hot spot vulnerability



Virtually no bandwidth decrease when 8 processors access same address

Authors' conclusions

- Analysis demonstrates that "the network and its libraries deliver excellent performance to users"
- Future work:
  - analyzing scalability with larger numbers of nodes
  - testing actual scientific applications
  - testing more elaborate communication patterns