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

Fabrizio Petrini, Wu-chan 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 website 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
The QsNET is based on two building blocks:

- A network interface card called Elan
- A crossbar switch called Elite

**Building Blocks**

- Elan
  - SDRAM I/F
  - Thread Processor
  - Processor
  - µCode
  - DMA Buffers
  - Inputter
  - FIFO 0
  - Link Mux
  - Clock & Statistics Registers
  - Table Walk
  - Engine
  - MMU & TLB
  - FIFO 1
  - PCI Interface (4 Way Set Associative Cache)
  - 400 MB/sec
  - Bidirectional
  - 200 Mhz/10 bits
  - 2 Virtual Channels
  - TLB Synchronized With Host
  - Thread Processor Runs Communication Protocols
  - 32-bit SPARC-based
  - Performance Counters
  - 51 Events
  - 8 32-bit Registers
  - 66 Mhz/64-bit PCI Interface
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
Packet Format

- Route
- One or More Transactions
- EOP
- Packet Header
- Routes
- CRC
- Transaction Type
- Context Value
- Memory Address
- Data (up to 64 Bytes)
- CRC

74-80 Bytes Overhead
320 Bytes Data Payload

Wormhole Flow-Control

[Diagram showing wormhole flow-control]
**Virtual Channels**

**Programming Libraries**

- ElanSiit
  - 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 used for interconnection

Unidirectional

Unidirectional Ping Bandwidth

<table>
<thead>
<tr>
<th>Message size [bytes]</th>
<th>Bandwidth [Mbytes/sec]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Elan3, Main-to-Elan</td>
</tr>
<tr>
<td></td>
<td>Elan3, Main-to-Main</td>
</tr>
<tr>
<td></td>
<td>Tports, Main-to-Elan</td>
</tr>
<tr>
<td></td>
<td>Tports, Main-to-Main</td>
</tr>
<tr>
<td></td>
<td>Tports, Elan-to-Elan</td>
</tr>
<tr>
<td></td>
<td>Tports, Elan-to-Main</td>
</tr>
<tr>
<td></td>
<td>MPI</td>
</tr>
</tbody>
</table>

Main-to-Elan
Main-to-Main
Elan-to-Elan
Elan-to-Main
Tports, Main-to-Elan
Tports, Main-to-Main
Tports, Elan-to-Elan
Tports, Elan-to-Main
MPI
Latency Unidirectional

Latency of unidirectional message

Unidirectional bandwidth on the ES40

Unidirectional bandwidth on the Alpha ES40
Bidirectional Ping Bandwidth

Latency with bidirectional messages
Hotspot Analysis

- Global Read Bandwidth
- Global Write Bandwidth
- Read Bandwidth per SMP
- Write Bandwidth per SMP

Number of SMPs

Hardware Barrier Latency

- Global Read Bandwidth
- Global Write Bandwidth
- Read Bandwidth per SMP
- Write Bandwidth per SMP

Number of SMPs
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.