Take the Expressway

Werner Butscher

http://www.dolphinics.com
Dolphin Interconnect Solutions

- Founded in 1992 as a spin-off of Norsk Data

- Products have since then been chips and hardware based on the SCI interconnect standard (IEEE 1596)
  - First customers Data General, Convex, Siemens-Nixdorf used the chips for their shared-memory servers
  - SCI via the I/O bus (Sbus, PCI, PCI-Express) was introduced later on:
    - Embedded solutions (aircraft, medical imaging)
    - Clustering (HPC, database accelerators)

- Steady and profitable growth in the recent years
  - ~ 30 employees (growing)
    - Main location in Oslo, sales & development worldwide
  - IPO in April 2006, traded at Oslo Stock Exchange
Current Hardware Product Line: SCI-to-PCI-Express

- SCI-to-PCI-Express adapter:
  - Based on LinkController 3
  - D350: Dual-channel SCI (two independent adapters): 8 PCIe lanes
  - D351: Single-channel SCI 1-D, 4 PCIe lanes
  - D352: Single-Channel SCI 2-D, 4 PCIe lanes
Dolphin Software Offerings

- **Driver software for Linux, Windows, Solaris and others**
  - Source code and Windows Installer
  - Available as RPM in next release
    - RedHat & Suse certification (+ Windows)

- **SISCI:** basis for custom *remote shared memory* solutions (develop environment)

- **SCILib: generic low-level messaging library**
  - Basis for SuperSockets

- **SuperSockets: low-latency, high-performance BSD sockets**
  - Transparent socket acceleration
  - Optimal for database cluster acceleration (MySQL, Oracle RAC)
  - Also works great with MPI & PVM!

- **WSD: Microsoft Winsock Direct** for Microsoft CCS with MS MPI via SCI

- Dolphin Interconnect Manager: GUI-driven cluster manager

  All provided as GPL/LGPL'ed Open Source
Dolphin Cluster Interconnect Management: sciadmin

- Cross-Platform Client/Server Operation
  - Clients and servers available for Windows, Linux etc.
  - Support for configuration and automatic re-routing
  - GUI front end for interconnect management and diagnosis
SuperSockets Features

- **Adds new address family AF_SCI to the system**
  - Support for TCP, UDP and RDS

- **Available as user-space and kernel-space version**
  - Even lower latency when running in user-space only
  - **Full(!)** socket semantics with kernel sockets, and for kernel services

**Support unmodified binaries using preload-library**

- **Reliability**
  - Automatic & transparent fall-back and fall forward between Ethernet and SCI
  - Support for redundancy/striping with multiple SCI adapters
  - Multiple routing strategies for dynamic SCI re-routing
SuperSockets Latency

- **Supersockets through Kernel**
  - Linux 2.6 on AMD Opteron via PCI Express

Bandwidth 525MB/s at 64kB
We were looking for a cost-effective solution for our customers that could handle their tough requirements for both performance and fault resilience.”

Roelof Kleinsmiede, senior technical architect at Capgemini, UK.
MPI via SCI – When Latency Matters

• Native MPI Implementations (3rd party):
  • ScaMPI (Scali)
    ➢ MP-MPICH from RWTH Aachen (based on MPICH1)
      • Cross-platform
      • Support for Meta computing (clusters of clusters)
    ➢ NMPI from NiceVT (based on MPICH2)
      • Runs a 500 CPU machine (zBox2 at University Zürich)

• MPI via Dolphin SuperSockets
  ➢ Supports MPICH2, OpenMPI, Scali MPI (GbE) – and also PVM!
  ➢ Combination of full MPI-2 with high-performance communication

<table>
<thead>
<tr>
<th>Ping Pong Latency</th>
<th>SCI</th>
<th>Infiniband</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latency [µs]</td>
<td>3.7</td>
<td>5.6</td>
</tr>
</tbody>
</table>

➢ Node hardware: 2.8GHz Intel Pentium D on Lindenhurst chip set (E7230).
  Benchmark: mpptest (from MPICH)
  SCI: D352 with NMPI
  Infiniband: Mellanox MT25204 [InfiniHost III Lx HCA] with OpenMPI 1.1 and OpenIB
Next Generation Interconnect: Dolphin Express

Direct PCI Express Switching
- PCI Express packets from end to end

Design Targets
- Lowest latency via PIO
  - Shared memory semantics
- High bandwidth w/o CPU usage
  - Advanced RDMA features
- Low-overhead hardware/software interface
  - Efficient conditional interrupts
  - Fast completion check/notification
- Compatibility
  - Existing software can be used w/o modification
  - Use standard PCI Express cables and switch components

Available in volume in 2007
Dolphin Express Main Features

- **Shared Memory**
  - Ultra-short latency for short messages
  - Very low overhead - direct CPU operations on remote memory
  - In order delivery

- **Address Mapping, 64 bits Physical Address Space**
  - Outbound and inbound mapping tables
  - Memory access protection key

- **Efficient Barriers and Error Detection**
  - Simplifies software protocols - improves performance

- **Transaction Controller Atomic Operations**
  - Fetch&Add, Compare&Swap
  - Serialized by protocol engine (no bus locking needed)

- **Reliable Connection**
  - Hardware data integrity (frame check, CRC, optional hardware retry etc.)

- **Programmable Protocol Engine**
  - Peer-to-peer communication
  - Optimized for low overhead software interaction

- **Existing Software will run w/o modification**
**Dolphin Express Topologies**

- **Direct All-to-All Connect**, N-1 channels per node, Low Latency
- **Binary N-Cubes**, $2^N$ nodes, N-1 hops, Low Latency
- **2D Mesh/Torus**, switch-less fine-grained scalability

**Scalable Switches**

**Ultimate flexibility:**
- Adoption to cluster size and application requirements
- Freedom to scale at any rate
- Combination with standard PCI Express switches
Summary

• **Dolphin:** an established player in the interconnect market
  - Strong position in embedded market
  - Dolphin's technology has proven potential for HPC
  - Transparent support of high-performance interconnect with SuperSockets:
    • Database cluster acceleration
    • Legacy HPC applications

• **Dolphin Express** will have a significant impact on HPC market:
  - Excellent performance with lowest latency and high bandwidth
  - Balance between RMA (PIO-based) and RDMA
  - Compatible with generic PCI Express switching hardware
  - Gives existing software solutions a performance boost
  - Will attract new software solutions
More Features

- **Optimized RDMA**
  - High bandwidth, low overhead
  - 4K Entries HW queue
  - 4 virtual engines, programmable priority.
  - Completion control in cacheable main memory
  - Optimized DMA multicast – up to 30 remote targets in one operation.

- **Low Overhead**
  - 156ns host-to-fabric latency
  - < 30ns pass-through routing latency

- **High Sustained Bandwidth**
  - 4 Gbyte/s full-duplex (8 PCI Express lanes end-to-end)

- **MSI - Message Signalled Interrupt**
  - Interrupt coalescing, conditional interrupts controlled by receiver
  - User mailboxes

- **Fast Fail over**
  - Multiple node IDs
  - Pre-configured alternative routing paths