Intel Cluster Ready
Appro Xtreme-X Computers with Mellanox QDR Infiniband

Steve Lyness
Vice President, HPC Solutions Engineering
slyness@appro.com
Company Overview
:: Corporate Snapshot

- Computer Systems manufacturer founded in 1991
- From OEM to branded products and solutions in 2001
  - Headquarters in Milpitas, CA
  - Regional Office in Houston, TX
  - Global Presence via International Resellers and Partners
  - Manufacturing and Hardware R&D in Asia
  - 100+ employees worldwide

- Leading developer of high performance servers, clusters and supercomputers.
  - Balanced Architecture – Delivering a Competitive Edge
  - Shipped over thousands of clusters and servers since 2001
  - Solid Growth from 69.3M in 2006 to 86.5M in 2007

- Target Markets:
  - Electronic Design Automation
  - Financial Services
  - Government / Defense
  - Manufacturing
  - Oil & Gas

*Source: IDC WW Quarterly PC Tracker
Company Overview

Why Appro

High Performance Computing Expertise – Designing Best in Class Supercomputers

- Leadership in Price/Performance
- Energy Efficient Solution
- Best System Scalability and Manageability
- High Availability Features
Company Overview

:: Our Customers
Company Overview

:: Solid Growth

$33.2M

$69.3M

$86.5M

2005

2006

2007

Compound Annual Revenue Growth

+61%

Empowered by Innovation
HPC Experience

:: Recent Major Installations

September 2006
NOAA Cluster
- 1,424 processor cores
- 2.8TB System Memory
- 15 TFlops

November 2006
LLNL Atlas Cluster
- 9,216 processor cores
- 18.4TB System Memory
- 44 TFlops

June 2007
LLNL Minos Cluster
- 6,912 processor cores
- 13.8TB System Memory
- 33 TFlops

February 2008
DE Shaw Research Cluster
- 4,608 processor cores
- 9.2TB System Memory
- 49 TFlops
April - June 2008
TLCC Clusters
- Total 426TFlops project
- LLNL, LANL, SNL
- 8 Clusters

June 2008
U of Tsukuba Cluster
- Total 95TFlops
- Quad-rail IB

July 2008
Renault F1 CFD Cluster
- Total 38TFlops
- Dual-rail IB
• **Technology is driven by customer’s needs**

• **Needs are driven by desire to reduce pain points**
  – Clusters are still hard to use – as clusters get bigger, the problem grows exponentially
  – Power, cooling and floor space are major issues
  – Weak interconnect performance
  – RAS is a growing issue/problem
Intel Cluster Ready and Intel Multi-Core Processors

APPRO INTERNATIONAL INC

HPC Cluster Solutions
Intel® Cluster Ready

Complete Solution

Certified Solution from a specific Platform Integrator

End-Users

Platform Solutions Provider
Deliver SKU that is a Certified Solution

Intel® Cluster Ready Solution Deployment

Certification

Appro Xtreme-X1 Supercomputer

Certification

Appro

Certification

Mellanox ConnectX

ISV-Applications

Hardware Certification

Software Certification

Appro Xtreme-X1 Supercomputer

Mellanox ConnectX
Intel® Software Tools for Building HPC Applications

Multi-Threading → Message Passing

Performance | Compatibility | Support | Productivity | Cross-Platform

http://www.intel.com/software
Benefits of Choosing an Intel® Cluster Ready Certified System

- **In production faster**
  - Simplified selection of compatible systems and applications
  - Faster and higher confidence initial deployment
- **Less downtime**
  - Ability to detect and correct problems soon after they occur
  - Better support from ISVs and your system vendor
  - Lower initial and ongoing costs
  - Higher ongoing productivity

*Demand an Intel® Cluster Ready certified system for your next cluster*
Quad-Core Intel® Xeon® processor
5400 series

Technology Rich
- Enhanced Intel Core Microarchitecture
- Larger Caches
- New SSE4 Instructions
- 45 nm High-k Process Technology

Compelling IT Benefits
- Greater Performance at Given Frequency and Higher Frequencies
- Greater Energy Efficiency

2nd Generation Intel Quad-Core
also available Dual-Core Intel® Xeon® processor 5200 series
Nehalem Based System Architecture

Benefits

- More application performance
- Improved energy efficiency
- End to end HW assist
  - Improved Virtualization Technology
- Stable IT image
  - Software compatible
  - Live migration compatible with today’s dual and quad-core Intel® Core™

Key Technologies

- New 45nm Intel® Microarchitecture
- New Intel® QuickPath interconnect
- Integrated Memory Controller
- Next Generation Memory (DDR3)
- PCI Express Gen 2

Functional system demonstrated Sept 2007 IDF

Up to 25.6 Gb/sec bandwidth per link

Extending Today’s Leadership

Production Q4’08
Volume Ramp 1H’09

All future products, dates, and figures are preliminary and are subject to change without any notice.
Appro Xtreme-X Supercomputers

APPRO INTERNATIONAL INC

HPC Cluster Solutions
Xtreme-X Supercomputer
:: Reliability by Design

- Redundancy where Required
  - Disk Drives
    - Diskless architecture
  - Fan Trays
    - N+1 Redundancy at the blade
  - Power Supplies
    - Hot Swappable N+1 Power supplies
  - Network components
    - Redundant InfiniBand and Ethernet networks
  - Servers
    - Management Nodes RAID6 Disk storage arrays
    - Management nodes - redundant pair (active/standby)
    - Sub-Management Nodes - redundant pair (active/active)
    - Compute nodes are hot swappable in sub-racks and are easily removed without removing any cabling from other nodes
• A 10°C increase in ambient temperature produces a 50% reduction in the long-term reliability of electronic equipment.
• A 5°C reduction in temperature can triple the life expectancy of electronic equipment.
• MTBF and Availability depend on cooling.

Source: BCC, Inc. Report CB-185R
• As Much as 40% or the cold air flow by passes the equipment

• Since the air is not directly pushed into the equipment, the cold air from the chillers are mixed with the hot air reducing the efficiency to cool the datacenter equipment.
• Delivers Cold Air directly to the equipment for optimum rack cooling efficiency.
• Delivers comfortable temperature to the room for return to Chillers
• Back-to-Back Rack configuration saves floor space in the datacenter and eliminates hot isles
• FRU and maintenance is done in the front side of the blade rack cabinet

• Up to 30% Improvement in Density with Lower Power per Equipment Rack
Xtreme-X Supercomputer
:: Power/Cooling Efficiency & Density

30% IMPROVEMENT IN SPACE BY ELIMINATING HOT ISLES AND HAVING BACK-TO-BACK RACK CONFIGURATION
A new way to approach building Supercomputing Linux clusters

Multiple clusters of various sizes can be built and deployed into production very rapidly.

The SU concept is very flexible by accommodating clusters from very small node counts to thousands of nodes.

This new approach yields many benefits to dramatically lower total cost of ownership.

Awarded Supercomputing Online 2007 Product of the Year
Appro Cluster Engine

APPRO INTERNATIONAL INC

HPC Cluster Solutions
• Issues with today’s cluster environments
  – Change Operating systems EASILY and QUICKLY
  – Run jobs on the cluster efficiently
  – Monitor nodes for health
  – Simple cluster management (Ex. IP Address allocation)

*Cluster Management Software turns a Collection of Servers into a, Functional, Usable, Reliable, Available, computing system*
• Total Management Package
  ✓ Network
  ✓ Servers
  ✓ Host
  ✓ Scheduling
  ✓ RAS

• Two Tier Management Architecture
  ✓ Scalable to large numbers of Servers
  ✓ Minimal Overhead
  ✓ Complete Remote Lights-Out Control

• Redundancy provides both High Availability and High Performance
  ✓ Redundant Management Servers
  ✓ Redundant Networks
  ✓ Active-Active Operation Supported
Appro Cluster Engine

:: ACE: Scalable SW Architecture

---

Global File System

Storage Server

Storage Server

External Network

Firewall Router

Operation Network (10GbE)

Mgmt Node

Operation Network (10GbE)

Operation Network (GbE)

Mgmt Network (GbE)

Operation Network (GbE)

Operation Network (GbE)

Operation Network (GbE)

I/O Node

I/O Node

I/O Node

I/O Node

I/O Node

I/O Node

I/O Node

Compute Node

Compute Node

Compute Node

Compute Node

Compute Node

Compute Node

Parallel File System

Storage Controllers

Servers or Bridge

Storage Controllers

Storage Controllers

InfiniBand Network

---

Networks:

- 10GbE for Computing
- GbE for Operation
- GbE Management

---

Needed Connections:

- 2x GbE per node
- 2x GbE
- 4x IB
- FC or GbE
- N GbE
- 2x 10GbE
- GbE or 10GbE
- 2x GbE per node
- 4X IB
- 4X IB

---

Compute Server Group
Cluster Management

:: Xtreme-X Architecture based on ACE

- Scalable Hierarchical Architecture
  - Uniform boot and response times
- Cached Root File System Support
  - Higher MTBF
  - Higher Bandwidth
  - Diskless Operation without light-weight OS
  - Full Linux OS
- Redundant Networks
  - GbE and IB are both dual-rail
- Run-Through Fault-Tolerance
  - Both Networks
  - Switches and Cables can fail without interrupting programs.
Appro Cluster Engine Management

Switches

<table>
<thead>
<tr>
<th>Name</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>Subnet</th>
<th>Rack</th>
<th>Slot/U#</th>
<th>GUID</th>
<th>State</th>
<th>Ports</th>
</tr>
</thead>
<tbody>
<tr>
<td>switch-0001</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>22</td>
<td>000002c90000003f7</td>
<td>up</td>
<td></td>
</tr>
<tr>
<td>switch-0002</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>22</td>
<td>000002c90000003f6</td>
<td>up</td>
<td></td>
</tr>
</tbody>
</table>

Ports on switch-0001

<table>
<thead>
<tr>
<th>Port</th>
<th>Rate (Gb/s)</th>
<th>NRate (Gb/s)</th>
<th>State</th>
<th>Connected To</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>20</td>
<td>up</td>
<td>server-0005</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>20</td>
<td>up</td>
<td>server-0005</td>
</tr>
</tbody>
</table>

ACE Log Console

2008-05-08 15:54:36.599 INF Host prod-0005 started on server-0011
2008-05-08 15:54:36.599 INF Host prod-0006 started on server-0012
### Appro Cluster Engine Management

#### Cluster Engine Xtreme-X1 (on norbert)

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>N</th>
<th>Grp</th>
<th>Rack</th>
<th>Slot/U#</th>
<th>Host</th>
<th>Net1</th>
<th>Net2</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>server-0001</td>
<td>management</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>active</td>
</tr>
<tr>
<td>server-0002</td>
<td>management</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>unknown</td>
</tr>
<tr>
<td>server-0003</td>
<td>group</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>sysgrp-0001</td>
<td></td>
<td></td>
<td>active</td>
</tr>
<tr>
<td>server-0004</td>
<td>group</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>sysgrp-0002</td>
<td></td>
<td></td>
<td>active</td>
</tr>
<tr>
<td>server-0005</td>
<td>compute</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>prod-0001</td>
<td></td>
<td></td>
<td>active</td>
</tr>
<tr>
<td>server-0006</td>
<td>compute</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>prod-0002</td>
<td></td>
<td></td>
<td>active</td>
</tr>
<tr>
<td>server-0007</td>
<td>compute</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>prod-0003</td>
<td></td>
<td></td>
<td>fail on</td>
</tr>
<tr>
<td>server-0008</td>
<td>group</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>8</td>
<td>sysgrp-0003</td>
<td></td>
<td></td>
<td>active</td>
</tr>
<tr>
<td>server-0009</td>
<td>group</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>2</td>
<td>1</td>
<td>9</td>
<td>sysgrp-0004</td>
<td></td>
<td></td>
<td>active</td>
</tr>
<tr>
<td>server-0010</td>
<td>compute</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td>prod-0004</td>
<td></td>
<td></td>
<td>active</td>
</tr>
<tr>
<td>server-0011</td>
<td>compute</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>2</td>
<td>1</td>
<td>11</td>
<td>prod-0005</td>
<td></td>
<td></td>
<td>active</td>
</tr>
<tr>
<td>server-0012</td>
<td>compute</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>2</td>
<td>1</td>
<td>12</td>
<td>prod-0006</td>
<td></td>
<td></td>
<td>fail on</td>
</tr>
</tbody>
</table>

#### ACE Log Console

- 2008-05-08 15:54:36.599 INF Host prod-0005 started on server-0011
- 2008-05-08 15:54:36.599 INF Host prod-0006 started on server-0012

- connected
Appro’s First Experience with Mellanox QDR Infiniband
Appro, Mellanox, and Intel Demonstration at IS08

- Demonstration of ANSYS Fluent Aircraft Structure Simulation
  - 16 Dual Socket Intel Harpertown Compute Nodes
  - Mellanox QDR Infiniband Fabric
  - See it in Booth 36 here at the show
Mellanox QDR Infiniband with Adaptive Routing

APPRO INTERNATIONAL INC
Mellanox ConnectX Quad Data Rate Infiniband HCA Cards
• CX4 and QSFP Connectors
• 2m reach with 30ga Copper Cables
• GA NOW!

Mellanox 36 Port Quad Data Rate Infiniband Switches
• **Scaleability**
  • 36 4X, 12 8X, or 12 12X ports
• **Bandwidth**
  • IB SDR, DDR, and QDR
  • 4X – up to 40Gb/s
  • 12X – up to 120Gb/s
  • Full QDR Wire Speed
• Dual rail QDR with 4X links to nodes and 12X rails
• 8GB/sec between Compute Nodes
• 48GB/sec between Switch Nodes
Appro Xtreme-X Computer

:: Scalable Unit using Mellanox QDR Infiniband

- Scalable Unit with 64 Compute Nodes
  - 4 Switch Nodes per Cabinet
- 6.1TF Peak Performance
  - 3GHz Nehalem Processors
  - 128 Processors
  - 512 cores
- 3TB Memory per Rack
- Dual-Rail QDR IB Interconnect
  - Redundant Interconnect
  - 8GB/sec BW for 16 node 128 core Groups
  - 48GB/sec BW between Groups
- Requires 32 Racks for 197TF
  - 32 Racks for Compute Nodes
  - 2 Back to Back Rows of 16 Racks
  - 256 36port QDR Switches
  - All Short Cables 2m
- Redundant Cooling and Power
- 28.8KW per Compute Rack