Are we there yet?:
10Gb Ethernet for HPC

Name: Dan Tuchler
Title: VP Strategy and Product Management – BLADE Network Technologies
Date: January 28, 2009
This is the year of 10 Gigabit Ethernet in HPC!!

2009?
2010?
2006?
HPC Interconnect Landscape

<table>
<thead>
<tr>
<th>Interconnect</th>
<th>Count</th>
<th>Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gigabit Ethernet</td>
<td>282</td>
<td>56.4%</td>
</tr>
<tr>
<td>10 Gig Ethernet</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>Infiniband</td>
<td>141</td>
<td>28.2%</td>
</tr>
<tr>
<td>Myrinet</td>
<td>10</td>
<td>2.0%</td>
</tr>
<tr>
<td>Others</td>
<td>67</td>
<td>13.4%</td>
</tr>
</tbody>
</table>
The bigger picture for HPC

Mass Market

top500

10Gig

IB

GigE

VIRTUAL COOLER EASIER
## HPC Forecast: Strong Growth Over Next Five Years ($ Millions)

<table>
<thead>
<tr>
<th></th>
<th>2007</th>
<th>2012</th>
<th>CAGR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Supercomputer</strong></td>
<td>$2,682</td>
<td>$3,512</td>
<td>5.5%</td>
</tr>
<tr>
<td><strong>Technical Divisional</strong></td>
<td>$1,610</td>
<td>$3,092</td>
<td>13.9%</td>
</tr>
<tr>
<td><strong>Technical Departmental</strong></td>
<td>$3,384</td>
<td>$5,763</td>
<td>11.2%</td>
</tr>
<tr>
<td><strong>Technical Workgroup</strong></td>
<td>$2,400</td>
<td>$3,193</td>
<td>5.9%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>$10,076</td>
<td>$15,617</td>
<td>9.2%</td>
</tr>
</tbody>
</table>

Source: IDC, 2008
10 Gig E issues

What’s holding up adoption?
- 10 Gig NICs
- Price of Switches
- Switch Scaling
- PHY Confusion
- Proof of Performance
10 Gig NICs

- Prices are dropping fast
- Major server vendors are including 10 Gig Ethernet as standard server feature (LOM)
- Several NIC vendors proving mature and stable for HPC

Chelsio 10GbE adapter for standard rack servers

Chelsio 10GbE adapter for IBM BladeCenter-H servers
Price of 10 Gig Ethernet Switches

- Switch ports used to cost more than servers!
- 10 Gig E switches now list for <$500 / port

RackSwitch G8124
10 Gig SFP+ Switch

BLADE’s Nortel 10G Blade Switch

RackSwitch G8100
10 Gig CX4 Switch
Switch Scaling

Typical CLOS Topology - 144 10GbE Ports
• 2-tier design scales to 288 ports

6 SPINE Switches

12 Leaf Switches
144 Non-blocking ports

VIRTUAL COOLER EASIER
HPC Topology – up to 208 10GbE Servers
IBM BladeCenter Design

10.1.1.0 VLAN 1
10.1.2.0 VLAN 2
10.1.3.0 VLAN 3
10.1.4.0 VLAN 4
10.1.4.0 VLAN 5
10.1.4.0 VLAN 6

Up to 210 Servers:
- 15 Enclosures
- 6 10GbE uplinks per enclosure
- 1 10GbE switch per enclosure

Load distribution across the core using OSPF ECMP:
- Separate IP subnet for each enclosure
- Separate VLAN and IP subnet for each RackSwitch
- Full network path redundancy – not just link level
- No Spanning Tree
- 2.3 to 1 oversubscription
PHY Confusion

- Optical standard interfaces for 10 Gig E:
  - Fixed optics
  - XENPAK
  - X2
  - XFP
  - SFP+
- 10GBase-T (i.e. Cat5, RJ45)
- CX4

Users have been unwilling to bet on a survivor!
New developments:

- **SFP+ Direct Attach Cables**
  - Passive cables with SFP+ ends
  - Low cost - $40 – $50
  - High density – same as RJ45

CX4 – in use today
10GBase-T

- The problem was harder than thought
- Expensive, power-hungry, and 2.6 usec latency
- But – 10GBase-T will eventually become widespread
Performance

10 Gig Ethernet offers:

- Same familiar operating environment
- Ease of use, debug, and management
- Path to 40 and 100 Gig Ethernet
- 10x bandwidth and 8x better latency vs. Gig Ethernet

- **But – do applications run faster!??!?**
  - Vendors talk about micro-benchmarks
  - Most users care about execution time
10GE 32% faster than 1G, equal to IB DDR, for 32 cores
PAM CRASH: Speed Up

10GE 70% faster than 1G, equal to IB DDR, for 64 cores
VASP 4.6.28: 
Elapsed Time (sec)

Vienna Ab-initio Simulation Package 
Molecular Dynamics

10GE 4.25x faster than 1G, equal to IB DDR, for 32 cores
VASP 4.6.28: Speed Up

10GE 6.3x faster than 1G, almost equal to IB DDR, for 64 cores
10GE 30% faster than 1G, equal to IB DDR, for 64 cores
BLADE’s 10GbE outperformed InfiniBand
- Significantly higher updates per second
- 31% Lower latency than InfiniBand

*Voltaire and BLADE tests used similar 3 GHz Xeon 5160 based servers with 4MB L2 cache*
RMDS Performance
BLADE’s 10GbE vs. InfiniBand

• BLADE’s 10GbE outperformed InfiniBand

RDMS Updates Per Second

0 100,000 200,000 300,000 400,000 500,000 600,000 700,000

IB 10 GE

Voltaire InfiniBand BLADE 10GbE

*Voltaire and BLADE tests used similar 3 GHz Xeon 5160 based servers with 4MB L2 cache
Why 10G Ethernet Now?

- Increasing demands of processors like quad core
- Prices are dropping
  - 10 Gig NICs
  - Switches under $500/port
  - Very attractive price performance
- IT Skill sets – easier to move to 10G Ethernet
- Technologies are more proven
  - CX4 and SFP+ are becoming the preferred PHY connections
  - Benchmarks are emerging
  - Early adopters and testing environments are delivering proof points of 10G Switch Scaling
BLADE Network Technologies

BLADE is the market-leading supplier of Gigabit and 10G Ethernet networking infrastructure solutions for blade server based environments

• First blade switch delivered in 2003
  • BLADE was a former division of Nortel and has been fully independent of Nortel since 2006
• Eight embedded Nortel Switch Modules for IBM BladeCenter
  • And growing!
• Over 45% blade networking market share
  • For every Cisco blade switch out there are 2 Nortel switches
  • Over 5 million ports connected to over 1 million blades
  • In over half the Fortune 500
• 6 Million hours of actual MTBF
• Management & Network Virtualization Tools
  • SmartConnect™ (with VMReady™) & BLADEHarmony™
IBM System x iDataPlex

2U Chassis

Web Server

HPC Server

Storage Tray

Storage Drives & Options

I/O Tray

10GbE

3U Chassis

Rack Management Appliance

iDataPlex Rear Door Heat Exchanger

10GbE BNT Switch

PDUs

VIRTUAL COOLER EASIER
IBM BladeCenter

Front View

Rear View

Blade server

10GbE Adapter

Switches

10GbE BNT Switch

VIRTUAL COOLER EASIER
Question & Answer

Thank you!!

IBM

iDataPlex Cluster 1350

RackSwitch G8124 & G8100

BLADE’s Nortel 10G Blade Switch for BladeCenter

VIRTUAL COOLER EASIER