Next Generation InfiniBand Clustering and Network Administration Tools

Brady Black
HPC Solutions Architect
QLogic Corporation
Agenda

- Introduction
- What is InfiniBand ‘IB’
- QLogic Simplifying IB networking
  - Deployment
  - Administration
A Global Company

- **Headquarters**
  - Aliso Viejo, California
- **Products**
  - High Performance Networking for Storage & HPC
- **Employees**
  - Approx. 900
- **FY08 Revenue**
  - $597.9M
- **NASDAQ Symbol**
  - QLGC
QLogic portfolio at Dell

Adapters

- QLogic 2500 series
  8Gb FC HBAs

- QLogic 2400 series
  4Gb FC HBAs

- Mezzanine Card
  8Gb FC for Dell PowerEdge Blade Servers

- Mezzanine Card
  4Gb FC for Dell PowerEdge Blade Servers

- 1GbE iSCSI HBA

Switches / Routers

- QLogic SB5802
  Stackable 8Gb FC Switches

- QLogic SB5600
  Stackable 4Gb FC Switches

- QLogic SB9000
  4Gb FC Director Switches

- QLogic 6140/6142
  Intelligent Storage Routers

InfiniBand

- 12-xxx IB Edge Switches

- 12800-xxx IB Director Switches

- QLogic 7000 IB HCAs

- SilverStorm 9240, 9120, 9080, 8040 IB Director Switches

- SilverStorm 902x IB Edge Switches
IB Director Design: Building Blocks

- **Module commonality across switch product line**
  - Spine cards
  - Leaf cards
  - Management card
  - Power Supply
  - Fan Module

- **Interchangeable components**

- **Enclosures**
  - 9240 (24 leaf cards)
  - 9120 (12 leaf cards)
  - 9080 (8 leaf cards)
  - 9040 (4 leaf cards)
  - 9020 (2 leaf cards)
QLogic QDR Switches (12X00)

QLogic 36 Port QDR Switches

- Managed (12300)
  - Redundant hot swappable fan/power supplies
  - Out of Band Management
  - On board SM capabilities

- Unmanaged (12200)
  - Low Cost
  - Single FRU

QLogic QDR Director Class Switches 12800

- Module commonality across switch product line
  - Spine cards
  - Leaf cards
  - Management card

Modularity and Density in 12800 Switches

- Ultra High Performance (UHP) 1:1
  - UHP
    - 12800-360 648 ports 864 ports
    - 12800-180 324 ports 432 ports
    - 12800-120 216 ports 288 ports
    - 12800-060 108 ports 144 ports
    - 12800-040 72 ports 96 ports

- Ultra High Density (UHD) 2:1
  - UHD
IB Management Software
Fabric Verification

- Can you find the loose cable?
- What about the missing cable?
- What about the one which was moved last night?
- Which Server didn’t boot?
- Which Switch has the wrong FW?
Fabric Manager
- 2048 node fabric initialization in <20 sec
- Rapid response to fabric changes (<1sec)
- Full SM/SA Redundancy; IBTA SM Failover
- Sophisticated routing algorithms
- Fabric verification / diagnostics support

FastFabric Toolset
- Centralized Fabric Administration Tools
- Rapid Fabric Installation/Upgrade
- Powerful Verification & Diagnostic tools
- Fabric Congestion Monitoring and Avoidance

Chassis and Element Management
- No user intervention required
- Hot swap FRU(s)
- Optional redundancy
- Common feature set, look and feel across all chassis/switch products
Switch details
Link specific properties
MPI Performance Tool Overview

- Latency/Bandwidth Deviation Test is an analysis and diagnostic tool for performing pair-wise bandwidth and latency testing.
- Tool is available in FastFabric using the “Check MPI Performance” TUI menu option.
- Test will report pairs outside an acceptable tolerance range.
- Will identify specific nodes which have problems and provide a concise summary of results.
- The tool can also be invoked via iba_host mpiperfdeviation or directly by ./run_deviation.
Sequential Mode Example

Running Sequential MPI Latency Tests - Pairs 3 Testing 3
Running Sequential MPI Bandwidth Tests - Pairs 3 Testing 3

Sequential MPI Performance Test Results
Latency Summary:
  Min: 2.51 usec, Max: 3.52 usec, Avg: 3.18 usec
  Range: +40.6% of Min, Worst: +10.7% of Avg
  Cfg: Tolerance: +30% of Avg, Delta: 0.80 usec, Threshold: 4.14 usec
    Message Size: 0, Loops: 4000

Bandwidth Summary:
  Min: 941.6 MB/s, Max: 1304.1 MB/s, Avg: 1178.2 MB/s
  Range: -27.8% of Max, Worst: -20.1% of Avg
  Cfg: Tolerance: -20% of Avg, Delta: 150.0 MB/s, Threshold: 942.5 MB/s
    Message Size: 2097152, Loops: 30

Bandwidth Details:
  Result   BW     Dev     Host (rank) --> Host (rank)
  FAILED   941.6  -20.1%  IBM-3550 (0) --> IBM-3455 (1)

Latency: PASSED
Bandwidth: FAILED
### Verbose Output

#### Latency Details:

<table>
<thead>
<tr>
<th>Result</th>
<th>Lat</th>
<th>Dev</th>
<th>Host (rank)</th>
<th>&lt;-&gt;</th>
<th>Host (rank)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PASSED</td>
<td>3.73</td>
<td>-4.5%</td>
<td>IBM-3550 (10)</td>
<td>&lt;-&gt;</td>
<td>st125 (0)</td>
</tr>
<tr>
<td>PASSED</td>
<td>3.34</td>
<td>-14.4%</td>
<td>IBM-3550 (10)</td>
<td>&lt;-&gt;</td>
<td>st999 (1)</td>
</tr>
<tr>
<td>PASSED</td>
<td>3.81</td>
<td>-2.5%</td>
<td>IBM-3550 (10)</td>
<td>&lt;-&gt;</td>
<td>IBM-3455 (2)</td>
</tr>
<tr>
<td>PASSED</td>
<td>3.79</td>
<td>-3.0%</td>
<td>IBM-3550 (10)</td>
<td>&lt;-&gt;</td>
<td>IBM-3655 (3)</td>
</tr>
<tr>
<td>PASSED</td>
<td>3.98</td>
<td>+1.9%</td>
<td>IBM-3550 (10)</td>
<td>&lt;-&gt;</td>
<td>IBM-3755 (4)</td>
</tr>
</tbody>
</table>

#### Bandwidth Details:

<table>
<thead>
<tr>
<th>Result</th>
<th>BW</th>
<th>Dev</th>
<th>Host (rank)</th>
<th>--&gt;</th>
<th>Host (rank)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PASSED</td>
<td>838.0</td>
<td>-9.9%</td>
<td>IBM-3550 (10)</td>
<td>--&gt;</td>
<td>st125 (0)</td>
</tr>
<tr>
<td>PASSED</td>
<td>947.9</td>
<td>+1.9%</td>
<td>IBM-3550 (10)</td>
<td>--&gt;</td>
<td>st999 (1)</td>
</tr>
<tr>
<td>PASSED</td>
<td>946.7</td>
<td>+1.8%</td>
<td>IBM-3550 (10)</td>
<td>--&gt;</td>
<td>IBM-3455 (2)</td>
</tr>
<tr>
<td>PASSED</td>
<td>873.0</td>
<td>-6.1%</td>
<td>IBM-3550 (10)</td>
<td>--&gt;</td>
<td>IBM-3655 (3)</td>
</tr>
<tr>
<td>PASSED</td>
<td>947.6</td>
<td>+1.9%</td>
<td>IBM-3550 (10)</td>
<td>--&gt;</td>
<td>IBM-3755 (4)</td>
</tr>
</tbody>
</table>
[root@tsg136 ~]$ iba_report
Getting All Node Records...
Done Getting All Node Records
Done Getting All Link Records
Done Getting All SM Info Records
Node Type Brief Summary

36 Connected CAs in Fabric:

<table>
<thead>
<tr>
<th>NodeGUID</th>
<th>Type</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>36 Connected CAs in Fabric:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NodeGUID Port LID PortGUID</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Width Speed</td>
</tr>
<tr>
<td>0x0005ad0000013d94 CA tsg110</td>
<td>1</td>
<td>0x001e 0x0005ad0000013d95 4x 2.5Gb</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>0x00066a00580001a6 CA VEx in Chassis 0x00066a005000010e, Slot 7</td>
<td>2</td>
<td>0x0023 0x00066a02580001a6 4x 2.5Gb</td>
</tr>
</tbody>
</table>

...
iba_report –o errors

[root@tsg136 ~]$ iba_report -o errors

Getting All Node Records...
Done Getting All Node Records
Done Getting All Link Records
Done Getting All SM Info Records
Getting All Port Counters...
Done Getting All Port Counters
Links with errors > threshold Summary

Configured Error Thresholds:
  SymbolErrorCounter             100
  LinkErrorRecoveryCounter       3
  LinkDownedCounter              3
  PortRcvErrors                  100
  PortRcvRemotePhysicalErrors    100
  PortXmitDiscards               100
  PortXmitConstraintErrors       10
  PortRcvConstraintErrors        10
  LocalLinkIntegrityErrors       3
  ExcessiveBufferOverflowErrors   3
  VL15Dropped                    100

Rate  NodeGUID          Port Type Name
  10g  0x00066a0001000108  1 SW i9k156 Leaf 5, Chip A
       LinkDownedCounter: 12 Exceeds Threshold: 3
<->  0x00066a0098005c31  1 CA tsg138

- Rapid analysis of the fabric against user defined threshold.
- Editable threshold for flexibility
- Easy to read output
Fabric Verification – FastFabric Can Find It!

# iba_reports -o errors -o verifylinks
Links with errors > threshold Summary

... Rate MTU NodeGUID Port or PortGUID Type Name
Cable: CableLabel CableLen CableDetails
20g 2048 0x0002c90200217ac0 1 CA n002
<- 0x00066a00d9000169 14 SW iS120
SymbolErrorCounter: 40156 Exceeds Threshold: 100
Cable: SS1145 11m Gore Passive Cu

2532 of 2532 Links Checked, 1 Errors found
-----------------------------------------------------------
Links Topology Verification

Rate MTU NodeGUID Port or PortGUID Type Name
Cable: CableLabel CableLen CableDetails
10g 2048 0x00066a0007000311 10 SW iS150
<- 0x00066a009800413e 1 CA n040
Cable: SS1020 7m Gore Passive Cu

Missing Link

2532 of 2532 Input Links Checked

Total of 1 Incorrect Links found
1 Missing, 0 Unexpected, 0 Misconnected, 0 Duplicate, 0 Different

- Rapid Fabric Wide Error Analysis
- Quickly Pinpoint Bad Links
- Identify Fabric Changes
- Compare fabric against intended design
- Concise Summary of errors
  - Name, port #, Speeds, etc
# iba_reports -o errors -o verifylinks

Links with errors > threshold Summary

... Rate MTU NodeGUID Port Type Name
Cable: CableLabel CableLen CableDetails
20g 2048 0x0002c90200217ac0 1 CA n002
<-> 0x00066a00d9000169 14 SW iS120
SymbolErrorCounter: 40156 Exceeds Threshold: 100
Cable: SS1145 11m Gore Passive Cu

2532 of 2532 Links Checked, 1 Errors found

Links Topology Verification

Rate MTU NodeGUID Port or PortGUID Type Name
Cable: CableLabel CableLen CableDetails
10g 2048 0x00066a007000311 10 SW iS150
<-> 0x00066a009800413e 1 CA n040
Cable: SS1020 7m Gore Passive Cu

Missing Link

2532 of 2532 Input Links Checked

Total of 1 Incorrect Links found
1 Missing, 0 Unexpected, 0 Misconnected, 0 Duplicate, 0 Different

- Rapid Fabric Wide Error Analysis
- Quickly Pinpoint Bad Links
- Identify Fabric Changes
- Compare fabric against intended design
- Concise Summary of errors
  - Name, port #, Speeds, etc

Link found with Excessive symbol errors
Fabric Verification – FastFabric Can Find It!

# iba_reports -o errors -o verifylinks

Links with errors > threshold Summary

... Rate MTU NodeGUID Port or PortGUID Type Name
Cable: CableLabel CableLen CableDetails
20g 2048 0x0002c90200217ac0 1 CA n002
<-> 0x00066a00d9000169 14 SW iS120
SymbolErrorCounter: 40156 Exceeds Threshold: 100
Cable: SS1145 11m Gore Passive Cu

2532 of 2532 Links Checked, 1 Errors found
-----------------------------------------------------------Links Topology Verification

Rate MTU NodeGUID Port or PortGUID Type Name
Cable: CableLabel CableLen CableDetails
10g 2048 0x00066a007000311 10 SW iS150
<-> 0x00066a009800413e 1 CA n040
Cable: SS1020 7m Gore Passive Cu

Missing Link

2532 of 2532 Input Links Checked

Total of 1 Incorrect Links found
1 Missing, 0 Unexpected, 0 Misconnected, 0 Duplicate, 0 Different

- Rapid Fabric Wide Error Analysis
- Quickly Pinpoint Bad Links
- Identify Fabric Changes
- Compare fabric against intended design
- Concise Summary of errors
  - Name, port #, Speeds, etc

Link found with Excessive symbol errors

Missing Cable Found
# iba_reports -o errors -o verifylinks
Links with errors > threshold Summary

...Rate MTU  NodeGUID          Port or PortGUID    Type Name
Cable: CableLabel           CableLen   CableDetails
20g 2048 0x00002c90200217ac0  1 CA n002
<-> 0x00066a00d9000169  14 SW iS120
SymbolErrorCounter: 40156 Exceeds Threshold: 100
Cable: SS1145 11m Gore Passive Cu

2532 of 2532 Links Checked, 1 Errors found

--- Links Topology Verification

Rate MTU  NodeGUID          Port or PortGUID    Type Name
Cable: CableLabel           CableLen   CableDetails
10g 2048 0x00066a007000311  10 SW iS150
<-> 0x00066a009800413e  1 CA n040
Cable: SS1020 7m Gore Passive Cu

Missing Link

2532 of 2532 Input Links Checked

Total of 1 Incorrect Links found
1 Missing, 0 Unexpected, 0 Misconnected, 0 Duplicate, 0 Different

Demonstrated Results: rapidly identified long standing problems in 3rd party fabrics, including problems internal to 3rd party large switches

- Rapid Fabric Wide Error Analysis
- Quickly Pinpoint Bad Links
- Identify Fabric Changes
- Compare fabric against intended design
- Concise Summary of errors
  - Name, port #, Speeds, etc

Link found with Excessive symbol errors
Missing Cable Found
Analysis Tools - Fast Fabric
Usage Model for Monitoring Tools

1. **Perform initial fabric install and verification**
2. **Optionally run tools in “health check only” mode**
   - Performs quick health check
   - Duplicates some of steps already done during verification
3. **Run tools in “baseline” mode**
   - Takes a baseline of present HW/SW/configuration
4. **Periodically run tools in “check” mode**
   - Performs quick health check
   - Compares present HW/SW/configuration to baseline
   - Can be scheduled in hourly cron jobs
5. **As needed rerun “baseline” when expected changes occur**
   - Fabric upgrades
   - Hardware replacements/changes
   - SW Configuration changes
   - Etc.
Fast Fabric Tool Categories

- **Fabric_analysis**
  - Checks for fabric level errors and/or link speeds
  - Checks for fabric level changes
    - Nodes added/removed, links added/removed

- **Chassis_analysis**
  - Checks for chassis configuration changes
  - Checks chassis health

- **SM_analysis**
  - HOST SM and Embedded SM variations
  - Check SM config and health

- **All_analysis**
  - User specified combination of the above
Give your network a kick in the apps!