

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

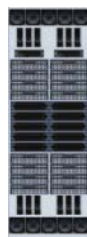


QLogic 6140/6142
Intelligent Storage Routers



InfiniBand

12-xxx IB
Edge Switches



12800-xxx
IB Director
Switches



QLogic 7000
IB HCAs



SilverStorm 9240, 9120, 9080, 8040
IB Director Switches



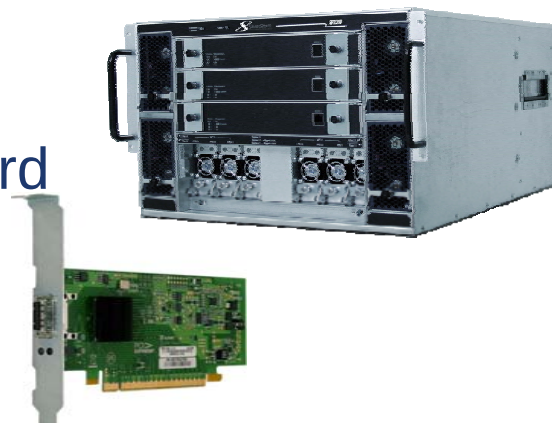
SilverStorm 902x
IB Edge Switches



IB Director Design: Building Blocks

■ Module commonality across switch product line

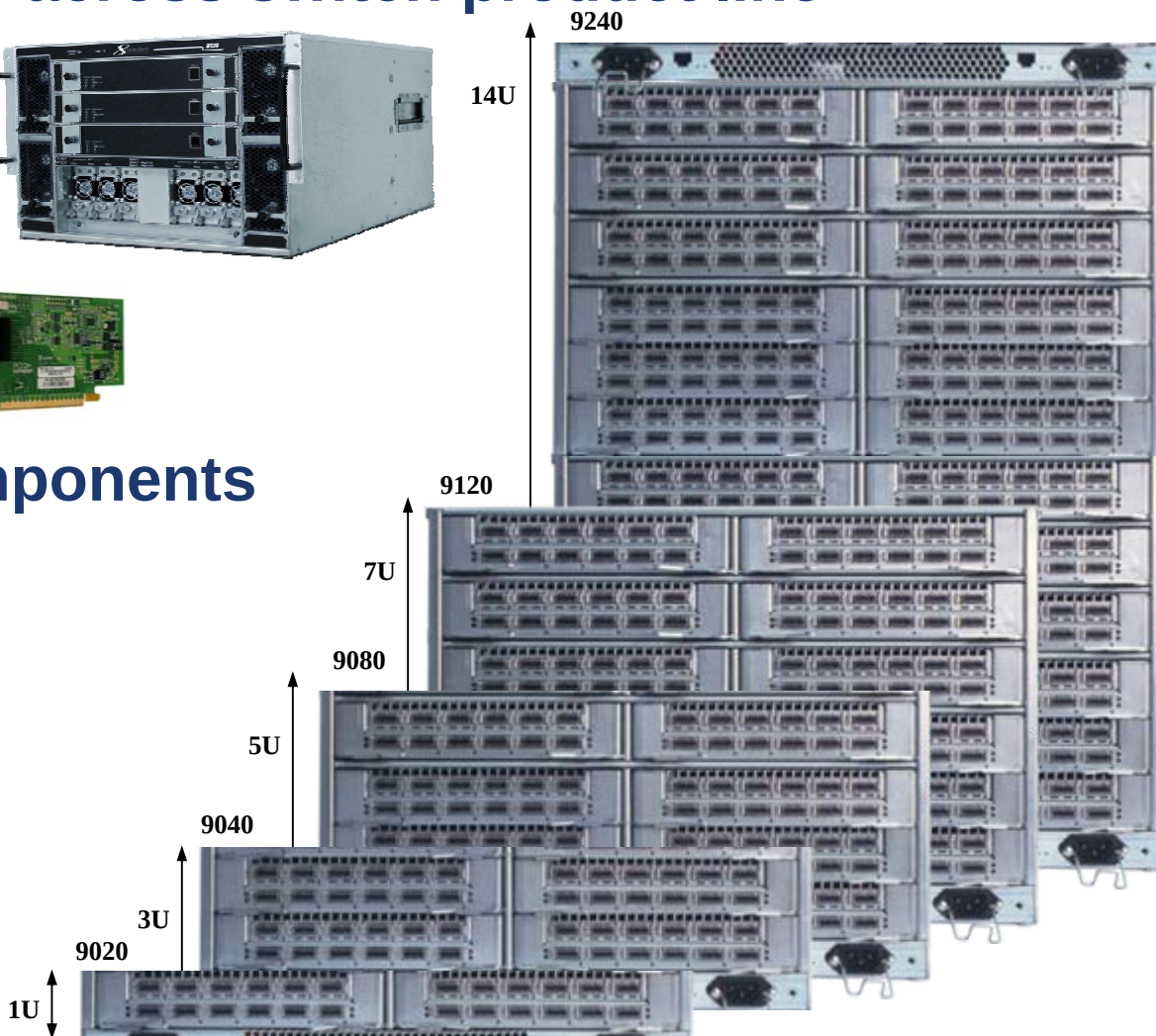
- Spine cards
- Leaf cards
- Management card
- Power Supply
- Fan Module



■ 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



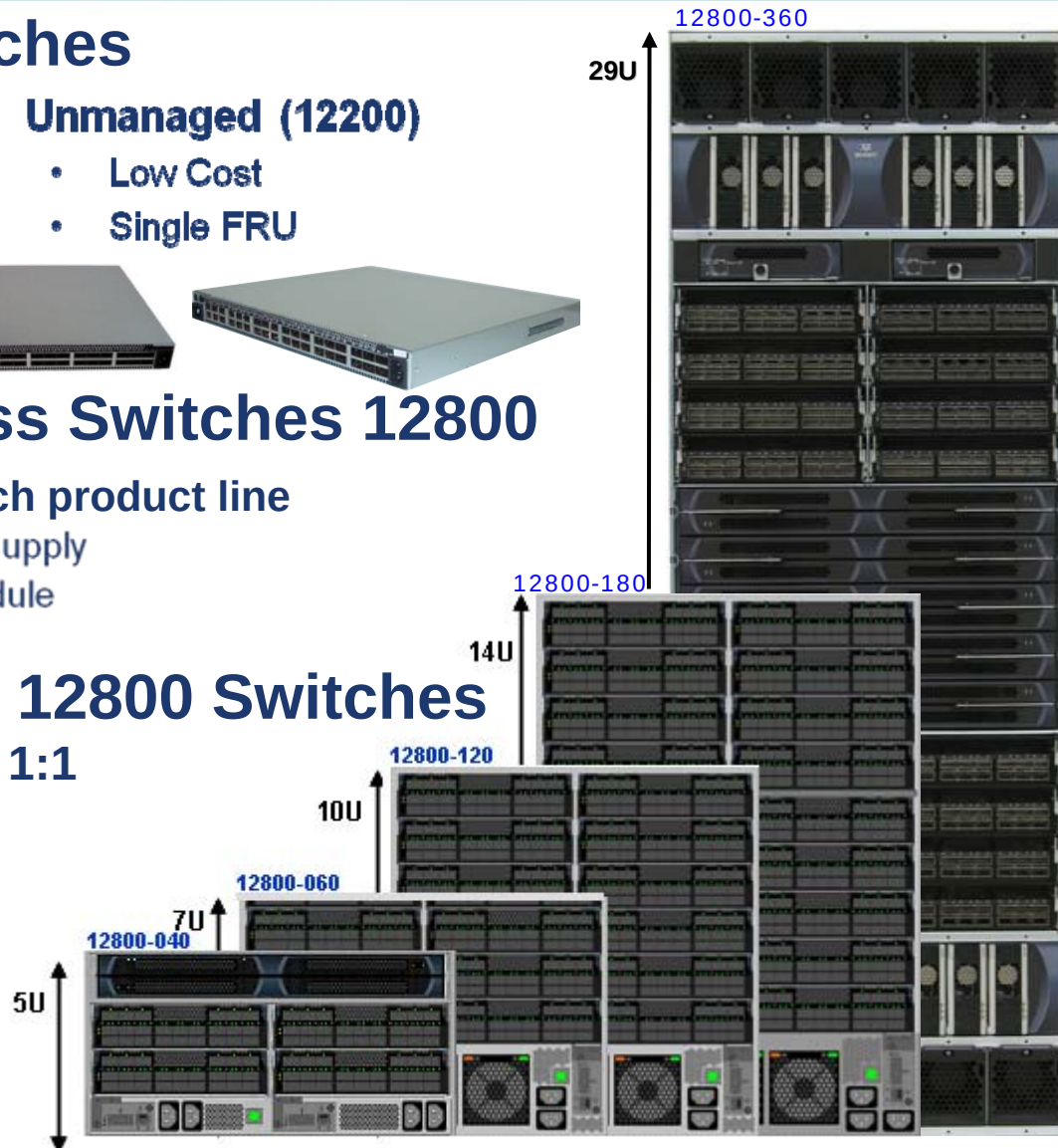
QLogic QDR Director Class Switches 12800

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

Modularity and Density in 12800 Switches

- Ultra High Performance (UHP) 1:1
- Ultra High Density (UHD) 2:1

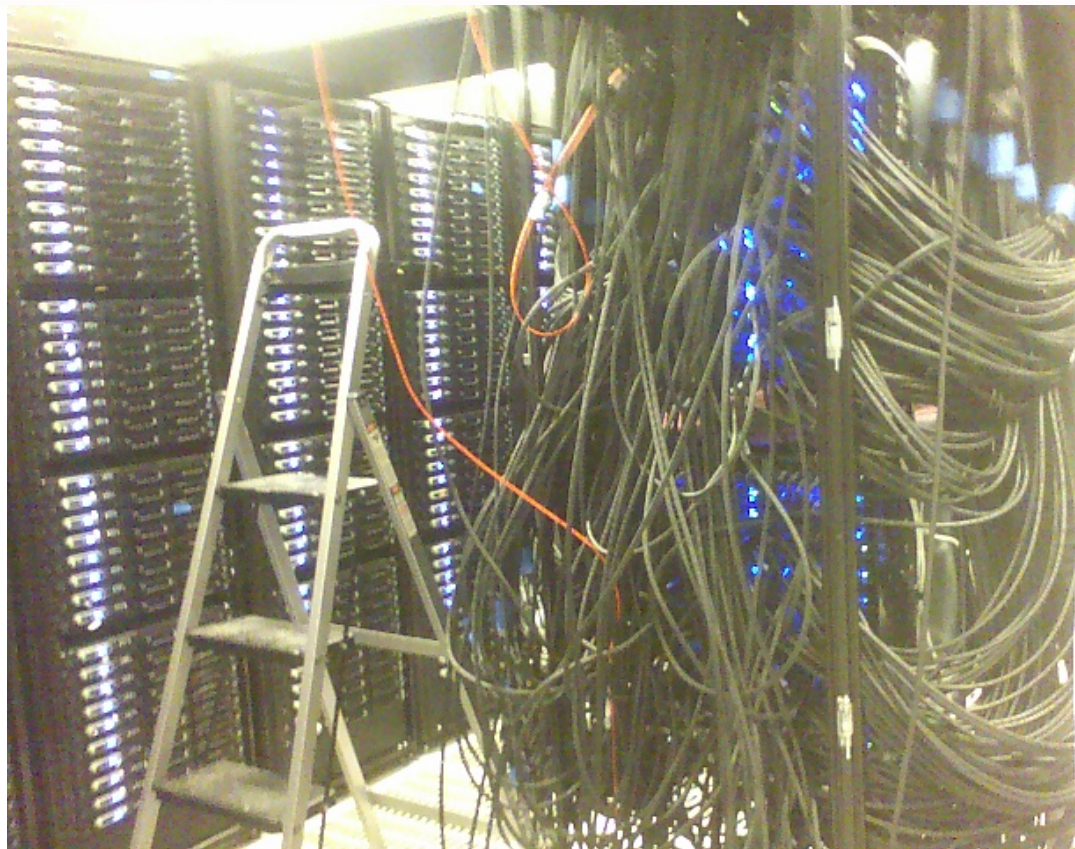
	UHP	UHD
• 12800-360	648 ports	864 ports
• 12800-180	324 ports	432 ports
• 12800-120	216 ports	288 ports
• 12800-060	108 ports	144 ports
• 12800-040	72 ports	96 ports



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?



InfiniBand Fabric Suite 2008

Fabric Manager

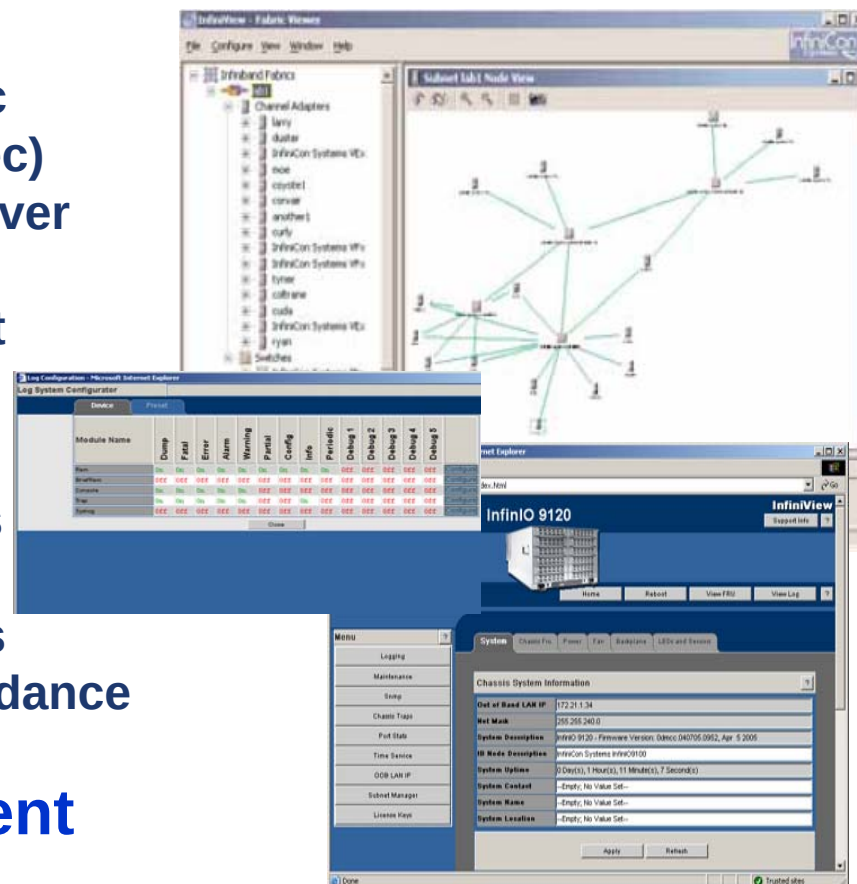
- 2048 node fabric initialization in <20 sec
- Rapid response to fabric changes (<1sec)
- Full SM/SA Redundancy; IBTA SM Failover
- Sophisticated routing algorithms
- Fabric verification / diagnostics support

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



Topology View

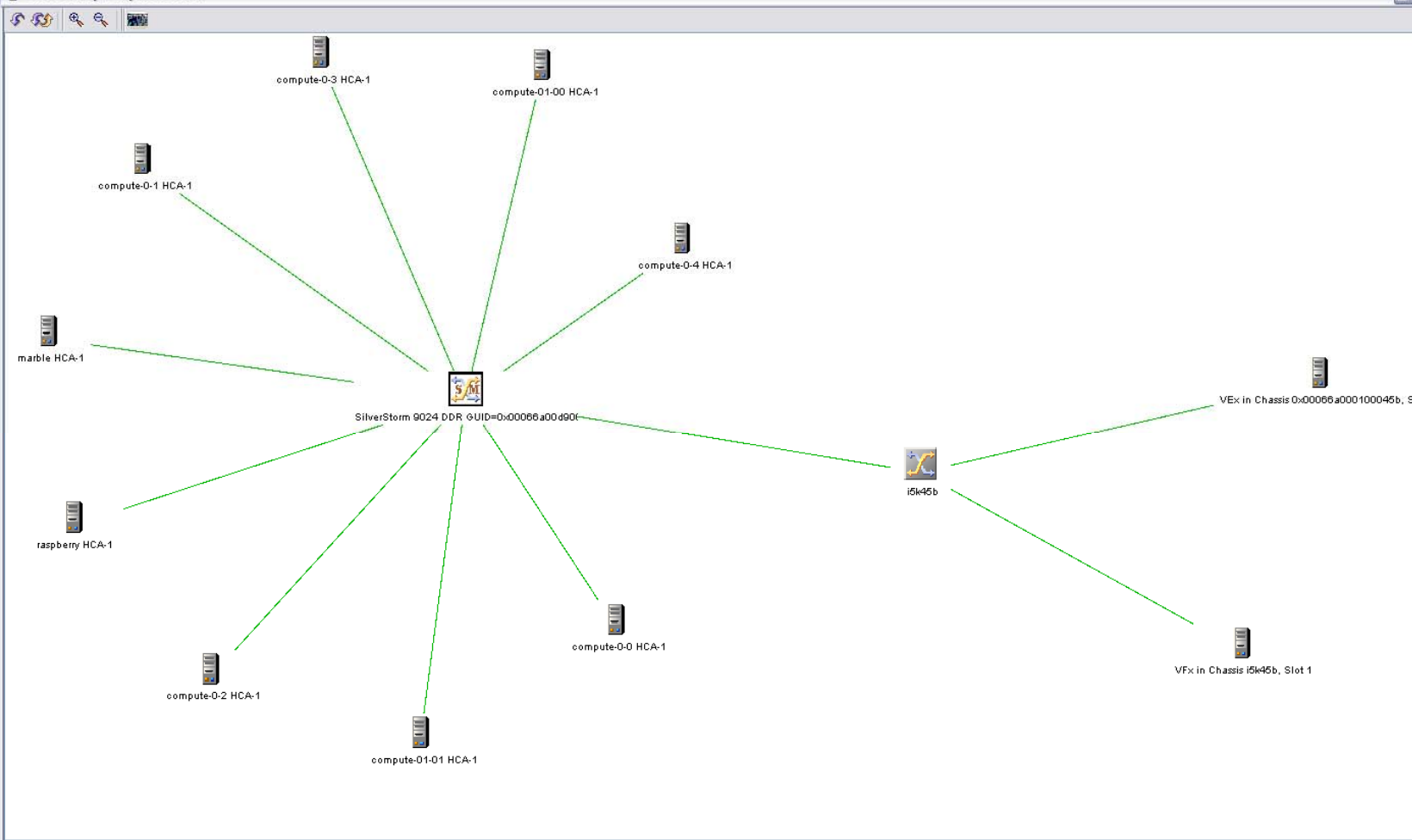
QLogic Fabric Viewer

Configure View Window Help

 **QLOGIC** The Ultimate in Performance

- Infiniband Fabrics
- raspberry
 - Channel Adapters
 - compute-0-4 HCA-1
 - marble HCA-1
 - compute-0-1 HCA-1
 - VFX in Chassis i5k45b, Slot 1
 - compute-0-3 HCA-1
 - raspberry HCA-1
 - compute-01-00 HCA-1
 - compute-01-01 HCA-1
 - compute-0-2 HCA-1
 - VEx in Chassis 0x00066a0001
 - compute-0-0 HCA-1
 - Switches
 - i5k45b
 - SilverStorm 9024 DDR GUID=

Subnet raspberry Node View



onets Groups

Event Name	Severity	Event Source	Description	Time
------------	----------	--------------	-------------	------

Switch details

QLogic Fabric Viewer

File Configure View Window Help

Subnet raspberry Node View

Infiband Fabrics

- raspberry
 - Channel Adapters
 - compute-0-4 HCA-1
 - marble HCA-1
 - compute-0-1 HCA-1
 - VFx in Chassis iSk45b, Slot 1
 - compute-0-3 HCA-1
 - raspberry HCA-1
 - compute-01-00 HCA-1
 - compute-01-01 HCA-1
 - compute-0-2 HCA-1
 - VEx in Chassis 0x00066a0001
 - compute-0-0 HCA-1
 - Switches
 - iSk45b
 - SilverStorm 9024 DDR GUID=

Subnet raspberry Node View

Diagram showing a central switch (SilverStorm 9024 DDR GUID=0x00066a00d900490) connected to multiple compute nodes (compute-0-0 HCA-1, compute-0-1 HCA-1, compute-0-2 HCA-1, compute-0-3 HCA-1, compute-01-00 HCA-1, compute-01-01 HCA-1, compute-0-4 HCA-1, raspberry HCA-1, marble HCA-1, VEx in Chassis 0x00066a0001000490, VFx in Chassis iSk45b, Slot 1).

Node Properties: SilverStorm 9024 DDR GUID=0x00066a00d900490

General Baseboard Mgmt Linear Forwarding Multicast Forwarding

Linear Forwarding

LID	Forwarded To Port
0x0001	0
0x0002	24
0x0003	24
0x0004	24
0x0005	19
0x0006	1
0x0007	2
0x0008	3
0x000c	20
0x000d	22
0x000e	21
0x000f	17
0x0010	18

OK Cancel

Subnets Groups

ID	Event Name	Severity	Event Source	Description	Time
----	------------	----------	--------------	-------------	------

Link specific properties

QLogic Fabric Viewer

Configure View Window Help

Subnet raspberry Node View

Left Panel (Infiniband Fabrics):

- Channel Adapters
 - compute-0-4 HCA-1
 - compute-0-1 HCA-1
 - marble HCA-1
 - 1
 - 2
 - compute-0-1 HCA-1
 - VFX in Chassis iSk45b, Slot 1
 - compute-0-3 HCA-1
 - raspberry HCA-1
 - compute-01-00 HCA-1
 - compute-01-01 HCA-1
 - compute-0-2 HCA-1
 - VEx in Chassis 0x00066a000100045b, Slot 1
 - compute-0-0 HCA-1
- Switches
 - iSk45b
 - SilverStorm 9024 DDR GUID=0x00066a00d9000490

Main View (Subnet raspberry Node View):

Diagram showing connections to a central SilverStorm 9024 DDR GUID=0x00066a00d9000490 node. Connected nodes include:

- compute-0-3 HCA-1
- compute-01-00 HCA-1
- compute-0-4 HCA-1
- compute-0-1 HCA-1
- marble HCA-1
- raspberry HCA-1
- compute-0-2 HCA-1
- compute-01-01 HCA-1
- compute-0-0 HCA-1
- VFX in Chassis iSk45b, Slot 1

Edge Details Window:

Links count: 1 link(s) present between compute-0-4 HCA-1 and SilverStorm 9024 DDR GUID=0x00066a00d9000490

Links details:

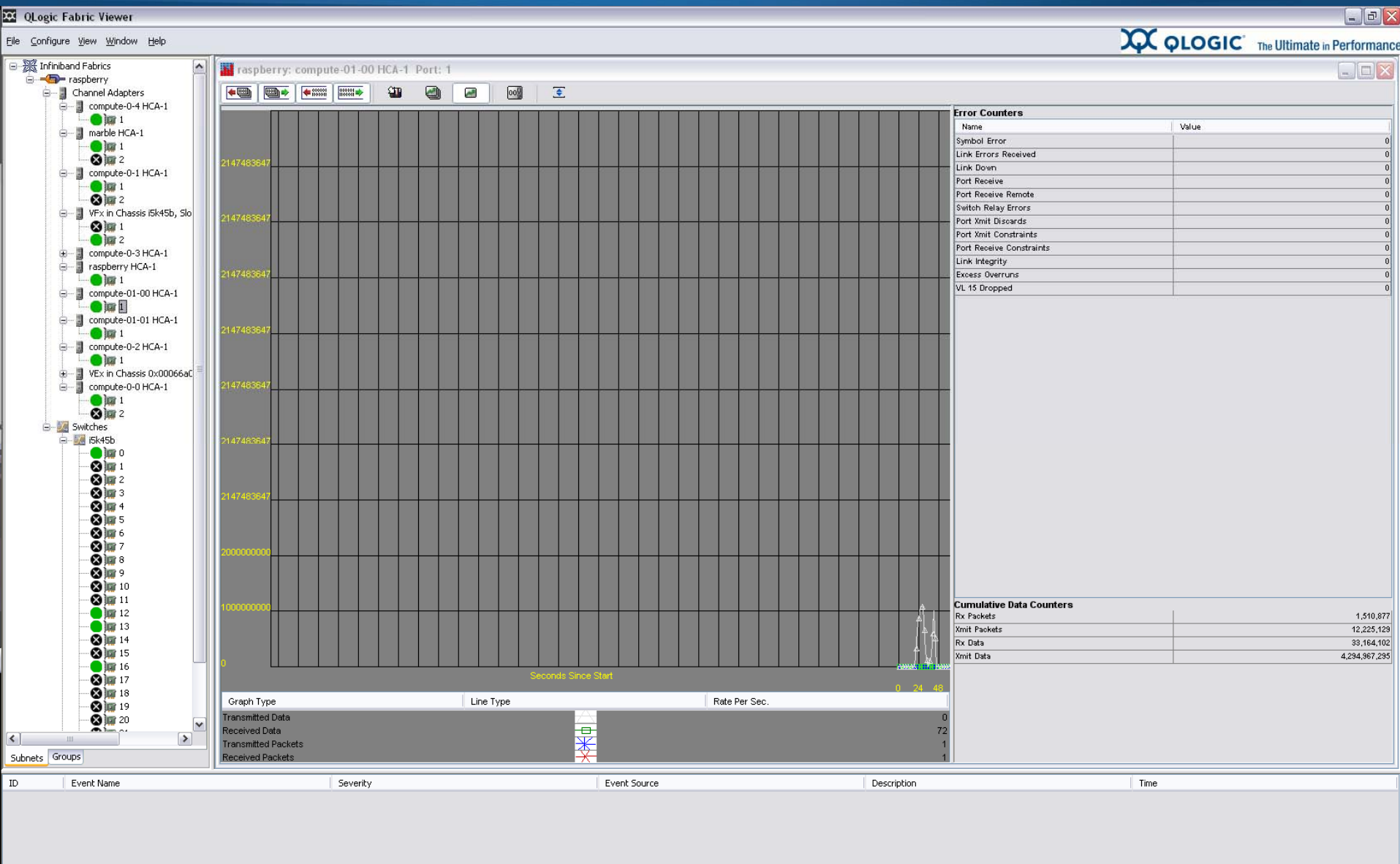
From compute-0-...	To SilverStorm 90...	Width	Speed	Status
1	19	4x	DDR	UP

Close

Bottom Panel (Event Log):

Event Name	Severity	Event Source	Description	Time
ected to raspberry on Port 3245 Active Nodes = 13 Active Links = 12				

HCA Specific Performance Metrics



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 mpirperfdeviation` 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:

Result	Lat	Dev	Host (rank)	<->	Host (rank)
PASSED	3.73	-4.5%	IBM-3550 (10)	<->	st125 (0)
PASSED	3.34	-14.4%	IBM-3550 (10)	<->	st999 (1)
PASSED	3.81	-2.5%	IBM-3550 (10)	<->	IBM-3455 (2)
PASSED	3.79	-3.0%	IBM-3550 (10)	<->	IBM-3655 (3)
PASSED	3.98	+1.9%	IBM-3550 (10)	<->	IBM-3755 (4)

Bandwidth Details:

Result	BW	Dev	Host (rank)	-->	Host (rank)
PASSED	838.0	-9.9%	IBM-3550 (10)	-->	st125 (0)
PASSED	947.9	+1.9%	IBM-3550 (10)	-->	st999 (1)
PASSED	946.7	+1.8%	IBM-3550 (10)	-->	IBM-3455 (2)
PASSED	873.0	-6.1%	IBM-3550 (10)	-->	IBM-3655 (3)
PASSED	947.6	+1.9%	IBM-3550 (10)	-->	IBM-3755 (4)

iba_report

```
[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:

NodeGUID	Type	Name	Port LID	PortGUID	Width	Speed
0x0005ad0000013d94	CA	tsg110				
	1	0x001e	0x0005ad0000013d95	4x	2.5Gb	
	2	0x001f	0x0005ad0000013d96	4x	2.5Gb	
0x00066a00580001a6	CA	VEx in Chassis 0x00066a005000010e, Slot 7				
	2	0x0023	0x00066a02580001a6	4x	2.5Gb	

...

- Generic helpful output about the fabric
- Overview of the fabric, hosts, switches and SM

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
ExcessiveBufferOverrunErrors	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 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

Fabric Verification – FastFabric Can Find It !

```
# 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 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

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	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

Link found with
Excessive symbol errors

Missing Cable Found

Fabric Verification – FastFabric Can Find It !

```
# 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	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

Link found with
Excessive symbol errors

Missing Cable Found

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

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.

■ **Fabric_analysis**

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

■ **Chassis_analysis**

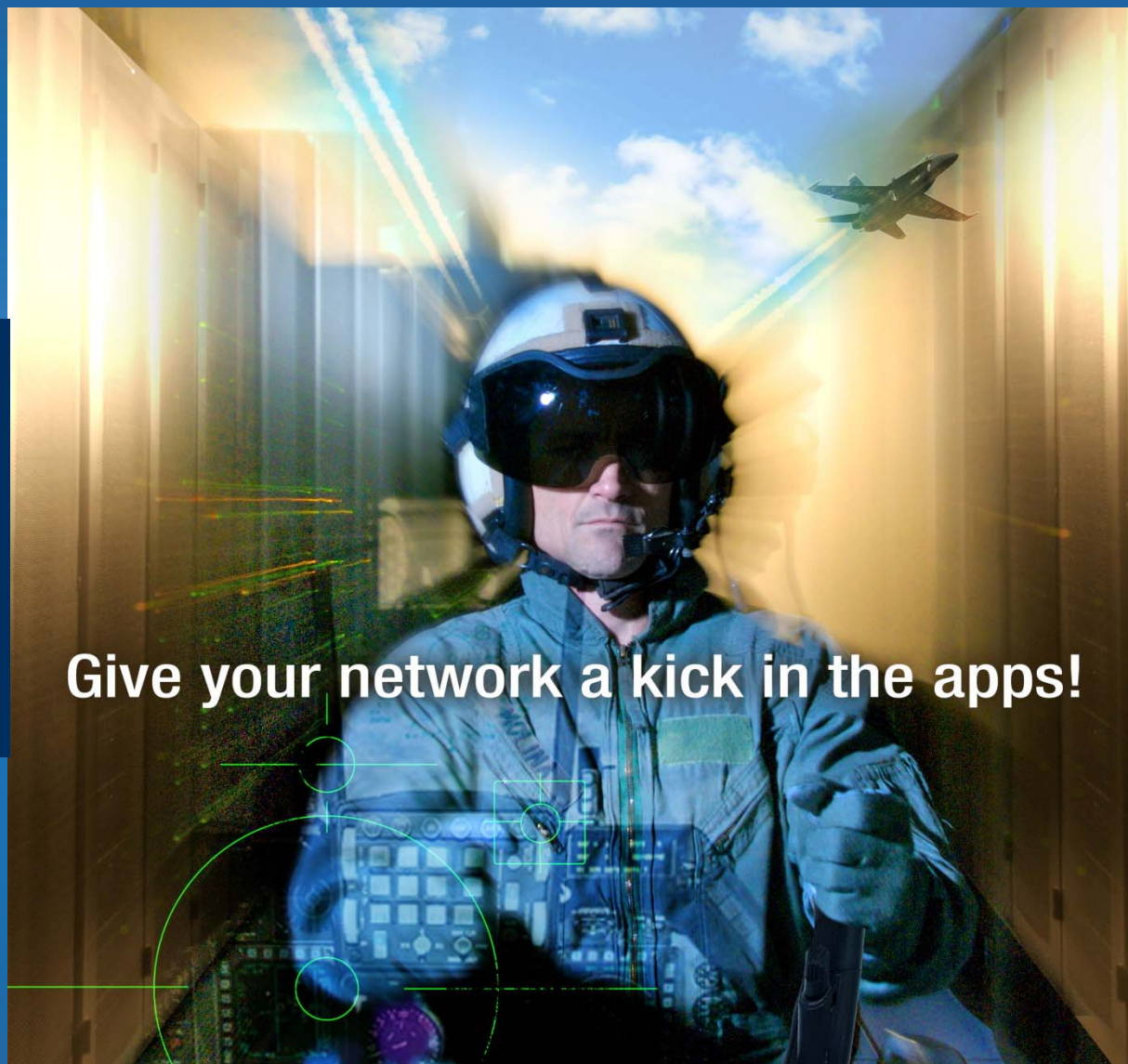
- Checks for chassis configuration changes
- Checks chassis health

■ **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!