QLogic InfiniBand Best Practices Guide

D000052-000

Preliminary

Information furnished in this manual is believed to be accurate and reliable. However, QLogic Corporation assumes no responsibility for its use, nor for any infringements of patents or other rights of third parties which may result from its use. QLogic Corporation reserves the right to change product specifications at any time without notice. Applications described in this document for any of these products are for illustrative purposes only. QLogic Corporation makes no representation nor warranty that such applications are suitable for the specified use without further testing or modification. QLogic Corporation assumes no responsibility for any errors that may appear in this document.

Document Revision History	
Revision A, March 7, 2008	
Changes	**Sections Affected**

First Published: March 2008

QLogic Corporation, 26650 Aliso Viejo Parkway, Aliso Viejo, CA 92656, (800) 662-4471 or (949) 389-6000

Table of Contents

1 Introduction

This manual describes best practices for a QLogic InfiniBand cluster.

This manual is organized as follows:

Section 1 describes the intended audience and technical support.

Section 2 describes the best practices for switch installation.

Section 3 describes the best practices for switch setup and operations.

Section 4 describes the best practices for fabric management.

Section 5 describes the best practices for Fast Fabric usage and fabric diagnosis.

Section 6 describes the best practices for emergency power off (EPO) procedures.

Section 7 describes the best practices for obtaining technical support information.

Intended Audience

This manual is intended to provide network administrators and other qualified personnel a reference for following best practices for a QLogic InfiniBand cluster.

Related Materials

- SilverStorm 9000 Hardware Installation Guide
- SilverStorm 9000 Users Guide
- SilverStorm 9000 CLI Reference Guide
- Fast Fabric Users Guide
- QuickSilver Fabric Manager and Fabric Viewer Users Guide
- QLogic InfiniBand Cluster Planning Guide
- QLogic InfiniBand Cluster Troubleshooting Guide
- InfiniBand Architecture Specification Volume 1

License Agreements

Refer to the *QLogic Software End User License Agreement* for a complete listing of all license agreements affecting this product.

Technical Support

Customers should contact their authorized maintenance provider for technical support of their QLogic switch products. QLogic-direct customers may contact QLogic Technical Support; others will be redirected to their authorized maintenance provider.

Visit the QLogic support Web site listed in Contact Information for the latest firmware and software updates.

Availability

QLogic Technical Support for products under warranty is available during local standard working hours excluding QLogic Observed Holidays.

Contact Information

Support Headquarters	QLogic Corporation 12984 Valley View Road Eden Prairie, MN 55344-3657 USA
QLogic Web Site	www.qlogic.com
Technical Support Web Site	support.qlogic.com
Technical Support Email	support@qlogic.com
Technical Training Email	tech.training@qlogic.com
North American Region	
Email	support@qlogic.com
Phone	+1-952-932-4040
Fax	+1 952-974-4910
All other regions of the world	
QLogic Web Site	www.qlogic.com

2 Switch Installation

Switch Hardware Installation

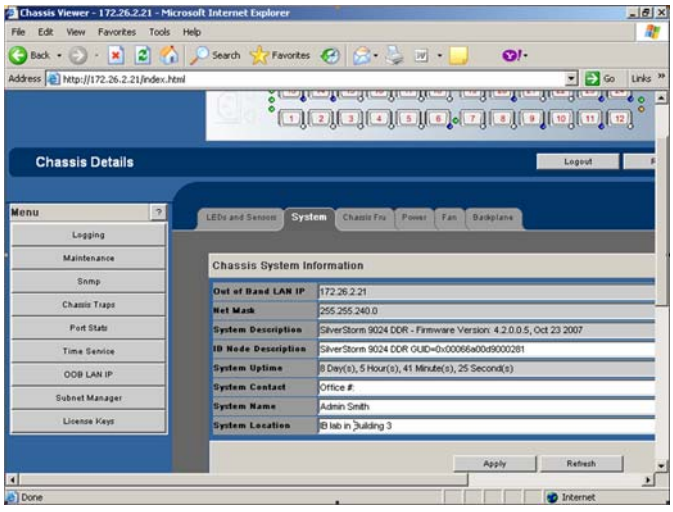
- Read the applicable SilverStorm InfiniBand Switch Quick Start Guide(s) included with the switch(es) for a summary of tasks ranging from a pre-installation checklist to installing document from a CD.
- Read the *SilverStorm 9000 Hardware Installation Guide* for a description of hardware installation and initial configuration tasks for SilverStorm 9000 series switches.

Switch & Cable Labeling

In any cluster (e.g., a one or two tier cluster), labeling the switches and cables is important for maintenance and troubleshooting purposes. All switches should be labeled. In a two-tier cluster, switches should be labeled as **core switch #** or **edge switch #** along with additional details (i.e., switch type, location, etc.), as needed.

In the Chassis Viewer GUI under the **System** tab specify the fields “System Contact", "System Name", "System Location". See example below:

Figure 2-1 Editing System Tab Fields



All cables should be labeled to identify the connected devices. In a two-tier cluster, it is either an "ISL" cable or a "Host" cable. For an "ISL" cable, additional information such as "From core switch # to edge switch #" should be added. For "Host" cable, information such as "Edge switch # to Host #" should be added.

Cable routing & handling

This section describes how to properly handle cables. This document also discusses cable routing to prevent symbol errors or damage to the cables. The following examples assume installtion of a fully populated SilverStorm 9120 (144 ports), however the same cable management principles apply regardless of switch chassis size.

Assumptions

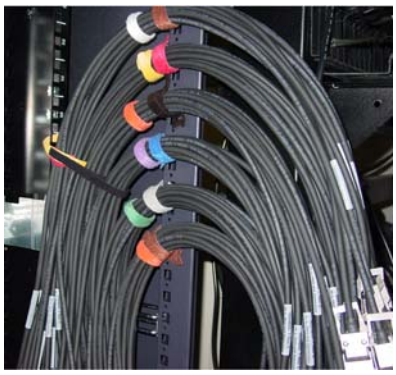
- Use a four-post server cabinet that is 36-inches deep.
- All InfiniBand cables have been pre-labeled before installation.

Instructions

1. Using Velcro straps, bundle InfiniBand cables in groups of twelve (12) to correspond with the number of InfiniBand ports for each SilverStorm 9000 series leaf card. Twelve (12) bundles are required for a fully populated 9120 (144 ports).
2. Attach the bundles to the rack in a vertical pattern. There will be six (6) bundles attached to the left post, and six (6) bundles attached to the right post.

NOTE:It is recommended to bundle cables at the swtich side. Depending upon the specific cluster configuration, it may not be practical to bundle for the full run of the cables.

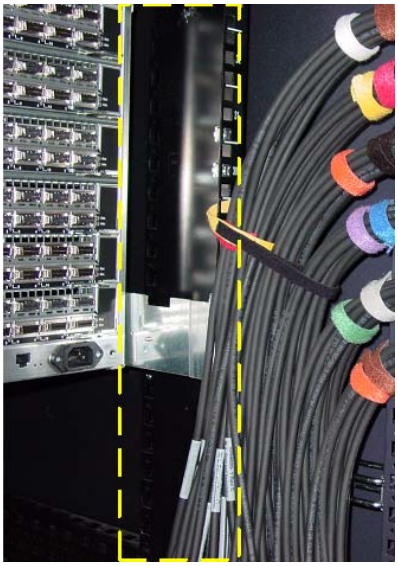
Figure 2-2 Attaching Cable Bundles to the Rack Post



NOTE:For rooms with under-floor cabling, attach the top bundle to the rack post first. For rooms with overhead cabling, attach the bottom bundle to the rack post first. The figures in this document represent an under floor example.

a. The vertical spacing of the bundles should leave enough room to access any SilverStorm 9000 series leaf card.

Figure 2-3 Vertical Spacing of Cable Bundles



b In order to reach the InfiniBand ports of each leaf card, leave approximately 24 inches of InfiniBand cable from the point each bundle is attached to the rack.

Figure 2-4 24-inch Cable Lengths

3. Connect InfiniBand cables to the ports on the SilverStorm 9120. Begin with the top bundle, which corresponds to the top leaf card in the chassis.

Figure 2-5 Cabled Chassis View 1

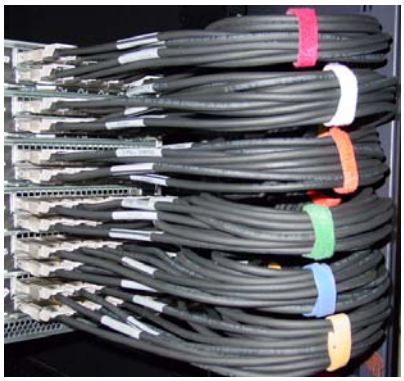


Figure 2-6 Cabled Chassis View 2

CAUTION!Be certain to maintain the minimum cable bend radius. See "Installing and Routing Cable" on page 2-6 for more information.

4. After cabling each leaf module, attach a Velcro strap at the midpoint of the bend radius of the cable bundle. This helps keeps the bundle together, as well as maintaining each bundle's respective position to the corresponding leaf card.

Figure 2-7 Velcro to Bend Radius Midpoint

Installing and Routing Cable

NOTE:Building and electrical codes vary depending on the location. Comply with all code specifications when planning the site and installing cable.
When running cable to the equipment, consider the following:

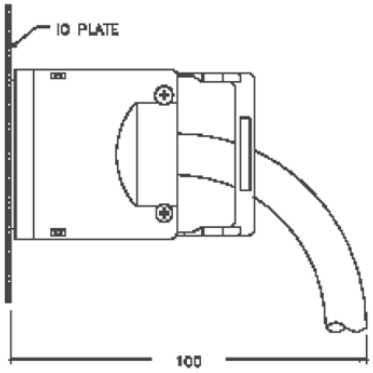
- Do not run cables where they can be stepped on or rolled over.
- Be sure cables are intact with no cuts, bends, or nicks.
- If the user is making a cable, ensure that the cable is properly crimped.
- Provide proper strain relief by adhering to the following guidelines:

Table 2-1. Cable Bend Radii

Assembly 90-Degree Bend Radii	
American Wire Gauge (AWG) Size Cable	**4X Bend Radius**
24	**5.00 inches**
26	**4.50 inches**
28	**4.00 inches**

- Temporary 90-degree bend can never be more than 0.5 inches tighter than the values listed above for any assembly.
- This is the absolute minimum sustained bend radius for each 4X cable AWG size. This measurement is the distance from the panel to the point where the cable makes a 90-degree bend. In other words, this number includes the 2" connector stand-off from the panel surface.

Figure 2-8 Bend Radius Measurement Diagram



- Support cable using a cable manager mounted above connectors to avoid unnecessary weight on the cable bundles.
- Bundle cable using Velcro straps to avoid injuring cables.
- Keep all ports and connectors free of dust.
- Untwisted Pair (UTP) cables can build up Electrostatic Discharge (ESD) charges when being pulled into a new installation. Before installing category 5 UTP cables, discharge ESD from the cable by plugging it into a port on a system that is not powered on.

When required for safety and fire rating requirements, plenum-rated cable can be used. Check the local building codes to determine when it is appropriate to use plenum-rated

Removing and Replacing FRUS

NOTE:For additional information on the following sections, refer to "Hot Swapping Compontents" in the SilverStorm 9000 Hardware Installation Guide.

Removing or Replacing a Leaf Card

Impacts:

- Logs:

  Messages indicating this event are logged in both the SM and switch logs. Following is an example:

```
W|   60 10:02:35.550B: Thread "IsmMadUPoll" (0x86816b50)
        Ism: Device has become inaccessible Switch Leaf 11

W|   60 10:02:35.810B: Thread "tChModWkr11" (0x860f1040)
        TMS: CMS Warning: Leaf 11 has been removed or power cycled

W|   60 10:04:00.020B: Thread "tSlot" (0x87d85b00)
        Osa: Leaf 11 was initializing, but is no longer present
```

- LEDs:

  None.

- Fabric Performance:

  If all hosts connected to the leaf were shutdown before removal, the fabric performance is unaffected.

- HCAs:

  The HCAs connected to the leaf will be down. All others will function normally.

Procedure/Hints

1. Remove the Velcro strap from the midpoint of the bend radius of the cable bundle corresponding to the leaf card to be removed/replaced.
2. Disconnect the InfiniBand cables from the ports of the leaf card.

3. Fold the bundle back to access the leaf card.

Figure 2-9 Removing a Leaf Card

4. Remove/replace the leaf card.
5. Reconnect the InfiniBand cables to the ports of the leaf card.
6. Reattach the Velcro strap to the midpoint of the bend radius of the cable bundle.

Removing or Replacing Spines

Impacts:

- Logs:

  Messages indicating this event are logged in both the SM and switch logs. Following is an example:

```
W|   60 10:09:16.030B: Thread "esm_top" (0x825c8860)
        ESM: Embedded SM Warning: topology_TrapDown: Switch node
'SilverStorm 9120 GUID=0x00066a0002000198 Spine 3, Chip
 B' (NodeGUID=0x00066A1004000542) has disappeared from fabric : 0
W|   60 10:09:16.030B: Thread "esm_top" (0x825c8860)
        ESM: Embedded SM Warning: topology_TrapDown: Switch node
'SilverStorm 9120 GUID=0x00066a0002000198 Spine 3, Chip
 A' (NodeGUID=0x00066A0004000542) has disappeared from fabric : 0
W|   60 10:09:17.180B: Thread "IsmMadUPoll" (0x86851520)
        Ism: Device has become inaccessible Switch Spine 3-A
W|   60 10:09:18.140B: Thread "IsmMadUPoll" (0x86851520)
        Ism: Device has become inaccessible Switch Spine 3-B
Devices initialized: leafs=12 spines=2 isMaster=1 SYS_MANAGER=1
initializing: leafs=0 spines=0
W|   60 10:09:20.690B: Thread "tChModWkr103" (0x86141df0)
        TMS: CMS Warning: Spine 3 has been removed or power cycled
```

- LEDs:

  None.

- Fabric Performance:

  Since each spine provides a certain amount of bandwidth to a fabric, removing it will affect the fabric performance.

- HCAs:

  All HCAs will function normally.

Removing or Replacing Power Supplies

Impacts:

- Logs:

  A message indicating this event will be logged. Following is an example:

```
MasterSpine1suppt-> W|   60 10:13:35.690B: Thread "tChPwrWkr06"
(0x861186c0)
        TMS: CMS Warning: Info, power supply 6 remove event has been
received!
```

- LEDs:

  If the chassis has redundant power supplies, the is no LED impact. In a non-redundant configuration, the chassis LED will become amber.

- Fabric Performance:

  None.

- HCAs:

  None.

Removing or Replacing Fans

Impacts:

- Logs:

  A message indicating this event will be logged. Following is an example:

```
W|   60 10:15:15.690B: Thread "tChFanWkr04" (0x8610a880)
        TMS: CMS Warning: Info, fan tray 4 remove event has been
received!
```

- LEDs:

  The chassis LED becomes amber.

- Fabric Performance:

  None.

- HCAs:

  None.

Removing or Replacing Cables

Impacts:

- Logs:

  None.

- LEDs:

  None.

- Fabric Performance:

  None.

- HCAs:

  None.

3 Switch Setup and Operations

For detailed information, refer to the following documents:

- SilverStorm 9000 Users Guide
- SilverStorm 9000 CLI Reference Guide
- SilverStorm 9000 Hardware Installation Guide
- Fast Fabric Users Guide

Switch Interfaces

A user can interface with a 9000-series switch in a number of ways:

- CLI: This textual user interface (TUI) runs on the switch and is accessible via the serial port or Ethernet port using a terminal emulator application such as Windows HyperTerminal.
- Chassis Viewer: This GUI resides on a switch and is accessible via the Ethernet port using an Internet browser.
- Fabric Viewer: This GUI application runs on either Windows or Linux hosts and interfaces with either the host or embedded subnet manager via Ethernet to manage the fabric. However, due to its rather limited functionality, the Fast Fabric toolset is preferred.
- Fast Fabric: A TUI running on an InfiniBand host using the QuickSilver HCA and host stack is a powerful tool to manage the fabric via Ethernet and InfiniBand protocols.

NOTE:For fabric-level management (i.e., one or more switches), use the Fabric Fabric toolset and Fabric Viewer only.

IP Address Settings & Management LAN Set-up

Although each managed spine has its own IP address, it is best to set and use the chassis IP address to access the ChassisViewer and CLI. For more information, refer to section “Changing the Switch IP Address and Default Gateway via the CLI” in the *SilverStorm 9000 Hardware Installation Guide*.

Switch Security

Each switch supports 3 levels of security:

- Level 1 is none
- Level 2 is Username/Password
- Level 3 is LDAP

Depending on a specific environment, a level should be selected. It is recommended that *at least* level 2 is enabled with telnet disabled so that only ssh can be used.

NOTE:It is recommened to change the adminstrator-level password. The new password must be included in the Fast Fabric configuration file. Additionally, make certain that the Fast Fabric configuration file can be accessed by only the root user.

Chassis and Broadcast Maximum Transmission Units

InfiniBand permits host channel adapters and switches to support MTUs of 256, 512, 1024, 2048 and/or 4096 bytes. These MTUs are the maximum sizes for InfiniBand packets. Since a given path through the fabric consists of multiple HCAs and switches, the MTU for a given path is limited to the smallest MTU of the components in the path. The MTU for a path is communicated to the application via path records that the application queries from the subnet manager and subnet agent (SA).

For InfiniBand multicast, the path can include many switches and HCAs. As such the maximum MTU allowed is limited to the smallest MTU of the components involved in multicast.

For InfiniBand user datagram-based protocols (e.g., Internet Protocol over InfiiniBand (IPoIB)/UD, InfiniBand multicast), the largest message that can be sent is limited by the MTU. If applications want to send larger messages via InfiniBand UD, an application specific implementation is required.

For InfiniBand RC-based protocols (e.g., VNIC, SRP, IPoIB/CM, SDP), the InfiniBand RC protocol hides the MTU limitations and permits messages up to 2GByte to be transferred. In the context of RC protocols, the MTU only affects performance. Actual performance effects can vary based on hardware used.

Different versions of MPI may use UD or RC protocols. Consult the MPI supplier for more information.

Switch/Chassis MTU:

The SilverStorm internally-managed chassis permit the MTU to be set at 2048 or 4096. This setting is configured into the switches and is discovered by the SM when the switch comes on line.

TCP/IP and MTU:

TCP/IP has the concept of MTU. In this context the MTU is the maximum packet size for the link layer below TCP/IP (e.g., Ethernet, IPoIB/UD, IPoIB/CM). For TCP applications, the MTU is transparent but can affect performance. For UDP applications, the MTU is visible and limits the size of the single frame messages that can be sent (however, IP can fragment larger messages as needed).

IPoIB and Broadcast MTU:

IPoIB makes use of InfiniBand multicast to perform IP multicast and broadcast. All IP networks make use of broadcast for ARP. Other protocols (DHCP, etc) and applications may also make use of IP multicast and/or broadcast. The QLogic SM permits the InfiniBand multicast group for IP multicast to be created in advance. This simplifies operations for the IPoIB participants and can ensure that IP multicast is consistently configured as desired.

The IPoIB broadcast MTU must be set consistently with the fabric hardware and the expectations of IPoIB participants.

When the QLogic SM is configured for strict InfiniBand multicast checking, InfiniBand multicast groups (including the IPoIB broadcast group), will not be permitted to be set to an MTU larger than any of the switches (e.g., the InfiniBand multicast MTU must be less than or equal to the MTU of every switch). This strict mode prevents server additions or boot ups from causing an unrealizable IPoIB network.

Similarly, when in strict InfiniBand multicast checking mode, the SM will generate **`NOTICE`** messages if a switch is added or rebooted with an MTU smaller than the present InfiniBand multicast MTUs. In this situation IPoIB connectivity for existing nodes could be impacted.

Hardware component naming conventions

For more information, refer to “Switch & Cable Labeling” on page 2-1. :

Every switch should be clearly labeled for easy identification. For example:

- For a 2-tier cluster, label the core switch as “Core #N” and the edge switch as “Edge #N.

Every cable should have labels on both ends clearly specifying each connection. For example:

- A connection between HCA port #1 of host #1 and port #5 of edge switch #3, the labels should have the following information:
  - Port #1/Host #1 <-> Port #5/Edge #3

Set-up NTP and Syslog Services

Syslog Setup

Set up a syslog host so that messages from the InfiniBand switches in the fabric, as well as the Fast Fabric management host can be stored and viewed.

NTP Setup

Setting up NTP allows the switch clock to be synchronized with the entire fabric, resulting in chronologically-ordered syslog log entries from all switches across the fabric.

Initial Bring-up Checks

- Make sure that all spines, leafs, power supply units, and fan trays are securely installed.
- Ensure that there are no red LEDs on any SilverStorm switches in the fabric.
- Access the CLI to make sure that the boot sequence completes cleanly. For multiple-switch fabrics, use the Fast Fabric toolset **`captureall`** command.
- Bring up ChassisViewer or use the CLI to clear all counters that may have incremented during boot.

4 Fabric Management

Best Practices for Fabric Management

For further information on fabric management, refer to the following documentation:

- Chassis Viewer: SilverStorm 9000 Users Guide
- CLI: SilverStorm 9000 CLI Reference Guide
- Fast Fabric: Fast Fabric Users Guide
- Fabric Manager and Fabric Viewer: QuickSilver Fabric Manager and Fabric Viewer Users Guide

Embedded Subnet Manager and Fabric Size

In fabrics comprised of System P servers and IBM GX/GX+ HCAs, use of the embedded subnet manager limits the connectivity of the number of external ports until the following criteria are met:

1. LPars in subnet must be less than or equal to 288 (240 when a 9240 is used)
   - If an LPar has more than 1 CA port into the subnet, then count that LPar once per CA port.
2. Max 144 total QLogic cable ports in subnet.

NOTE:This formula works regardless of IBM GX/GX+ HCA configuration.

HSM Minimal Server Configuration

Assuming a 1024-node fabric, with 1 instance per server, the typical requirements are:

- OS: SLES9, SLES10 or RHEL4/5
- CPU: dual processor Intel EMT64 or AMD64
- Memory: 2 GB
- Bus: PCI-X (at least one available slot for a HCA)

- Disk: 10 GB available

If multiple instances of the SM are running on a server, the additional multiplier depends on several factors:

Memory:

- Are the other instances backup SMs? (Add ~1/2 GB memory per instance).
- Are the other instances primary SMs? (Add 1 GB memory per instance).

Disk:

The file system requirements do not increase relative to instances. However, the user should consider if extra space is required due to verbose log settings, or local archiving of logs. The maximum recommended instances per server is four (4)

HSM Configuration:

The **`/directory/iview_fm config`** file provides a template with default values for each possible SM instance. The configuration file is commented, describing the purpose of each configuration setting. For more information, refer to the *QuickSilver Fabric Manager and Fabric Viewer Users Guide*.

Redundant Subnet Management

In order to ensure SM coverage in a subnet (e.g. plane), the configuration should include redundant subnet managers. In a fabric with more than one subnet manager, one SM acts in the "master" role while the rest act in a "standby" role. Should the master SM fail or otherwise lose control of managing the fabric or plane, the hierarchy of takeover of the standy SMs is as follows: the SM with the highest priority setting is the master; in cases of equal priorities, the SM with the lowest GUID is the master.

A host-based SM should be configured with multiple instances on a single host to act as master and standby subnet managers for multiple fabrics/planes. It is recommended to use the priority settings to control master/standby arbitration.

NOTE:Subnet managers (embedded and/or host-based) running on a single subnet MUST have the same configuration settings, with the exception of priorities.

Multiple Instances of the Host Subnet Manager

The number of instances of the host subnet manager is limited to the number of physical HCA external ports on that host. Each port should be connected to a separate subnet or plane in order to ensure maximum subnet management coverage among the instances. For example:

- In the host subnet manager configuration file, replace **`/directory/iview_fm config`** with **`/etc/sysconfig/iview_fm.config`** (note the period in the name).
- Highlight key attributes that must be changed from default (like device and port, etc.)

Subnet, Performance and Baseboard Manger Priority Values

The priority values of the SM, PM, and BM need to be set to the same value on the same node so that the master and standby ordering is consistent between the three managers.

For the host subnet manager this would be accomplished in the configuration file by setting the values of **`SM_n_priority`**, **`PM_n_priority`** and **`BM_n_priority`** to be the same, where **`n`** is the instance number in the configuration file (e.g., **`SM_0_priority`**, **`PM_0_priority`**, and **`BM_0_priority`**).

For the embedded subnet manager, the BM/PM priority setting defaults to a value of **`1`**. To set the embedded SM priority, use the CLI command **`smSetPriority`**.

Consistent priority settings are helpful in the case of a node failure (e.g., if the master node goes down, then all would fail over to a common standby) but not necessarily in the case of a process failure. For example, if the SM process dies, but the physical machine is still running and the BM and PM are running, then only the SM will fail over, resulting in a split fabric.

Maintenance Impacts on the Fabric

The following table illustrates the affect of various maintenace tasks on the fabric and the subnet manager.

Table 4-1. Fabric and SM Maintenance Tasks

Maintenance Task	SM Action	Impact to Fabric	Notes
Reboot End Node	**End node is removed from the fabric**	**Limited to who that end node was communicating with**	
Remove InfiniBand cable to End Node	**End node is removed from the fabric**	**Limited to who that end node was communicating with**	
Power-On 9240 leaf	**Activate all ports and any end nodes connected**	**No impact**	
Power-Off 9240 leaf	**All end nodes connected to it drop off the fabric**	**Limited to who those end nodes were communicating with**	
Remove inter-switch link	Move routes through that link to other choices	The link is broken for as long as it takes to discover and reprogram around the fabric changes. This will impact a subset of the end nodes as their LIDs are balanced across the inter-switch links	Not recommended if critical apps are running
Add an End Node	Program Node and add to fabric	No impact	
Add inter-switch link	Activate the 2 ports and use in routing	No impact	

Table 4-1. Fabric and SM Maintenance Tasks

Maintenance Task	SM Action	Impact to Fabric	Notes
Removing modules from core switches		The 24 inter-switch links are broken for as long as it takes to discover and reprogram around the fabric changes. This will impact a subset of the end nodes as their LIDs are balanced across the inter-switch links	Not recommended if critical apps are running
Adding Modules to Core switches	Activate and use in routing	Minimal impact if done in a responsible manner	
Removing Core switch	Fabric should be taken offline		
Adding Core switch	Fabric should be taken offline		

Additional notes regarding a fabric event impact:

- When doing maintenance on an InfiniBand fabric, it is important to realize that any "removal" event will result in some disruption to the fabric. The impact to the fabric grows with the affected component's proximity to the core. The disruption can be limited to as little as 5-20 seconds in duration in a typical 1000 node fabric if care is taken to ensure that multiple events are not inadvertently created during the "removal" event.
- It is generally recommended to attempt one change at time. For example:
  - Plugging in a cable
  - Unplugging a cable
  - Power cycling a core switch
  - Adding a host

Multiple, simultaneous insertion and removal events require the SM to wait for the fabric to settle before completing its routing changes. This can extend the overall outage.

- For cooling reasons any chassis card removed must be replaced with a card or blank within 2 minutes.

Verifying via log files that all ports/nodes are available after restarting a switch

While a switch is restarting, a variety of notices may be generated as the ports retrain the links. The SM performs subnet sweeps to detect added or removed ports. If a change is detected from sweep to sweep, then a fabric summary notice is generated containing the following:

- Number of added and/or removed switches
- Number of added and/or removed HCAs
- Number of added and/or removed end ports
- Number of added and/or removed total ports
- Number of added and/or removed subnet manager(s)

In addition to the notice, an informational fabric summary entry is generated for every subnet sweep. A subnet sweep occurs automatically every 5 minutes irregardless of subnet changes, yielding a 'snapshot' of node counts at every sweep interval. These informational entries contain the same counts as the notice entries, except that the informational entries contains totals. The notice entries contain *deltas*.

After approximately 2 minutes, the subnet refreshes and the SM performs a sweep that has no differences from the previous sweep. The informational fabric summary entry contains the 'correct' number of end nodes and/or HCAs.

NOTE:An easy way to determine this quiet state is when the informational fabric summary entries are unchanged from entry to entry.

Comparing the numbers from this informational fabric summary entry to the anticipated values is the recommended method for verifying subnet integrity. However, if the user does not know what the correct values should be, refer to an informational fabric summary entry prior to the restart.

5 Fast Fabric Usage and Fabric Diagnosis

Best Practices for Fast Fabric usage & Fabric diagnosis

In tems of switch management, the Fast Fabric toolset allows the user access CLI commands that can be executed across all switches in the fabric.

There is no direct tie to between Fast Fabric and the Fabric Manager. However, the recommended best practice it that when using the host-based SM, install the SM and Fast Fabric on the same nodes.

Most of the best practices for Fast Fabric usage and fabric diagnosis has been documented in the *Fast Fabric Users Guide*. Therefore, this section will not attempt to repeat the information. Instead, it will point to different sections of the *Fast Fabric Users Guide* describing the best practices:

- For users not familiar with Fast Fabric tools, refer to Section 2, *Fast Fabric Overview*.
- To properly set up and use the Fast Fabric for initial host drivers installation and verification, refer to section 3.2 *Set Up the Fabric* and section 3.3 *Using Fast Fabric*.
- Section 4, *Fast Fabric TUI Menu*, describes how to use Fast Fabric to manage different aspects of a fabric such as host setup for MPI, hosts management and chassis management.
- Section 5, *Detailed Descriptions of Command Line Tools*, provides detailed information of the tools necessary for fabric management and diagnosis.
- Section 6, *MPI Sample Applications*, lists all sample MPI applications such as bandwidth, latency, High Performance Linpack, and Pallas.

6 Emergency Power Off (EPO)

EPO Scenarios

- Full floor EPO .
- Server-only EPO.
- Server and switch EPO (leaving management servers up).

Emergency power off procedures

1. Stop any Fast Fabric scheduled health analysis tools, such as cron jobs.
2. Stop all subnet managers. This will limit logging events from the SM reporting fabric changes and loss of components.
3. Perform the EPO.

Power restoration

1. Power on all servers and/or switches.
2. Make sure the subnet manager(s) have been started and/or power on any management nodes used for SM and/or Fast Fabric.
3. Clear all port counters using the Fast Fabric command **`iba_report -C -o none`**.

   NOTE:This will work even if there are P5 nodes with a perfomance manager agent (PMA).

4. Perform an **`all_analysis`** to identify any missing, powered-off or damaged components in the fabric.

The **`iba_report`** options used by the **`fabric_analysis`** (and **`all_analysis`**) can be configured using **`FF_FABRIC_HEALTH`**. By default, these will clear the counters *after* checking for errors. Therefore, these will be clear in preparation for the next health check.

To manually clear these, all link counters in the fabric can be cleared via:

```
iba_report -C -o none
```

NOTE:This requires that all HCAs support a PMA.

Switch link counters can be cleared via:

```
cmdall -C 'ismPortStats -clear -noprompt'
```

NOTE:Where applicable, use `-H` and `-F` options to select the appropriate set of switches. By default `/etc/sysconfig/iba/chassis` will be used as the list of switches.