QLogic InfiniBand Cluster Planning Guide

D000051-000

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 planning criteria for a QLogic InfiniBand cluster.

This manual is organized as follows:

Section 1 describes the intended audience and technical support.

Section 2 describes InfiniBand cluster planning tasks.

Intended Audience

This manual is intended to provide network administrators and other qualified personnel a reference for planning an 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 Best Practices 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 Cluster Planning

Key Terminology

Following describes the key terminologies used in an InfiniBand fabric (or cluster) and the SilverStorm 9000 switch family.

- **Constant Bisectional Bandwidth (CBB)** - an InfiniBand fabric that does not support maximum bandwidth (i.e., 10Gbps for SDR or 20Gbps for DDR) because one or more nodes are oversubscribed (i.e., the number of inter switch links is less than number of hosts).
- **Core switch** - A backbone switch to interconnect all edge switches. For an example, refer to Figure 2-1 on page 2-5.
- **Double Data Rate (DDR)** - DDR = 20 Gbps.
- **Edge switch** - A switch interconnecting nodes and core switches. For an example, refer to Figure 2-1 on page 2-5.
- **Fabric** - An InfiniBand infrastructure that includes host channel adapters (HCAs), cables, switches and optional I/O gateways (i.e., Ethernet and Fibre Channel).
- **Fat Tree** - Defines a topology where multiple tiers of switches (or switch chips) are used to create a single, full bi-sectional bandwidth fabric.
- **Full Bisectional Bandwidth (FBB)** - describes an InfiniBand fabric with maximum bandwidth (i.e., 10Gbps for SDR or 20Gbps for DDR).
- **Fabric Manager (FM)** - is comprised of a subnet manager, a performance manager, and a baseboard manager and is capable of managing complex InfiniBand networks.
- **Inter switch link (ISL)** - A cable from one switch to another.
- **Leaf** - A modular edge switch located inside a SilverStorm 9000 Multi-Protocol Fabric Director (MPFD) switch. A leaf module contains 12 InfiniBand ports.
- **Managed Spine** - A multi-protocol fabric director (MPFD) spine with a management board and applicable software.
- **Node** - An HCA port or an InfiniBand switch chip.

- **Single Data Rate (SDR)** - SDR = 10Gbps.
- **Subnet Manager (SM)** - The main function of the SM is to assign a local ID (LID) to host nodes, automatically perform fabric reconfiguration following a failure and manage multicast groups.
- **Spine** - A modular core switch located inside a SilverStorm 9000 MPFD switch that contains a management board and applicable software.
- **Unmanaged Spine** - A spine module without a management board and software.

Cluster Planning

In order to design a functional InfiniBand cluster, many steps with many different parameters must be taken into consideration. The main steps are:

- Requirements gathering
- Define cluster characteristics
- Selecting InfiniBand switches
- Floor layout
- Cable length calculation
- Cable planning

NOTE:

For detailed information, refer to documents listed in “Related Materials” on page 1-1.

Requirements Gathering

Information for designing an InfiniBand cluster must be carefully and accurately gathered. The potential sources could be:

- Meeting with the customer
- A Request for Proposal (RFP) document
- A site survey
- Questions and answers exchanged with customers
- A survey form filled out by customers (See next page)

Presales Questionnaire for HPC Customers

Table 2-1. Presales Questionnaire

Questions	Answers
1. How many Fabrics are required?	
1.1 Number of nodes required for Fabric 1	
1.2 Number of nodes required for Fabric 2	
1.3 Number of nodes required for Fabric 3	
2.0 Is SDR or DDR required?	
3.0 Is FBB (Full Bisectional Bandwidth - 10.0/20.0 Gbps) or CBB (Constant Bisectional Bandwidth) required?	
3.1 If CBB is required, 3.3/6.6 or 5.0/10.0 Gbps	
4.0 Server	
4.1 Model	
4.2 HCA type	
4.3 Number of planes connected	
5.0 Rack/Cabinet size (42U typical) - Height, Width, and Depth	
6.0 Known floorplan/layout	
6.1 Number of racks per row	
6.2 Distance between racks	

Table 2-1. Presales Questionnaire

Questions	Answers
6.3 Number of rows	
6.4 Distance between rows	
6.5 Cables will be run under raised floor or above the racks with cable trays?	
6.5.1 If raised floor is used, how deep?	
6.5.2 If cable trays are used, how far above the racks are the overhead trays?	
7.0 InfiniBand Fabric Management	
7.1 Internal or external SM	

Determine Cluster Characteristics

Speed: Depending on customer requirements, either SDR or DDR InfiniBand switches can be selected. In the absence of any requirements, follow the guideline below to select the speed:

Table 2-2. Speed

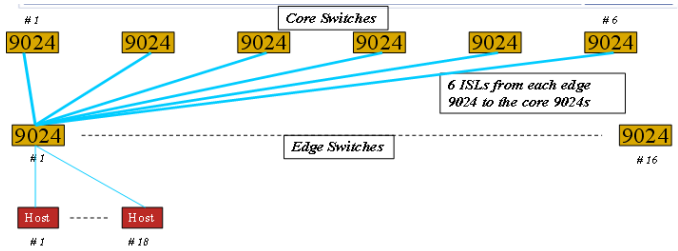
Factor	SDR	DDR
Cost	x	
Performance		x
New Technology		x

Cluster bandwidth: Depending on customer requirements, either CBB or FBB can be selected. In the absence of any requirements, follow the guideline below to select the bandwidth:

Table 2-3. Bandwidth

Factor	CBB	FBB
Cost	x	
Performance		x

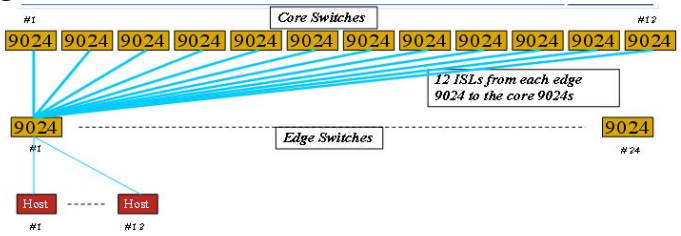
Figure 2-1 CBB Cluster with DDR 9024 Switches



18 hosts : 6 ISLs = 3:1 subscription rate

20Gbps / 3 = 6.6 Gbps (CBB)

Figure 2-2 FBB Cluster with DDR 9024 Switches



12 hosts : 12 ISLs = 1:1 subscription rate

20Gbps / 1 = 20 Gbps (FBB)

Topology: Depending on the cluster size (i.e., the number of nodes in the cluster), the topology can be either one tier or two tiers. Follow the guideline below to select the topology:

Table 2-4. Topology

Factor	One Tier	Two Tiers
Cost for Cluster <= 288 nodes per subnet	x	
Cluster <= 288 nodes per subnet	x	x
Cluster > 288 nodes per subnet		x

Figure 2-3 One Tier Cluster

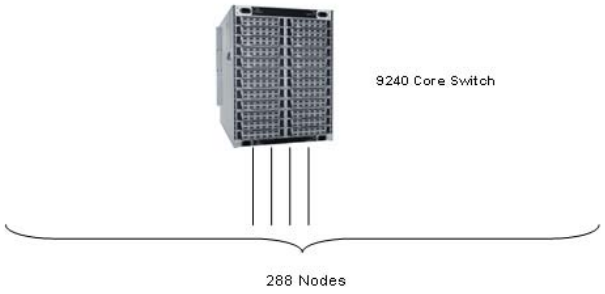
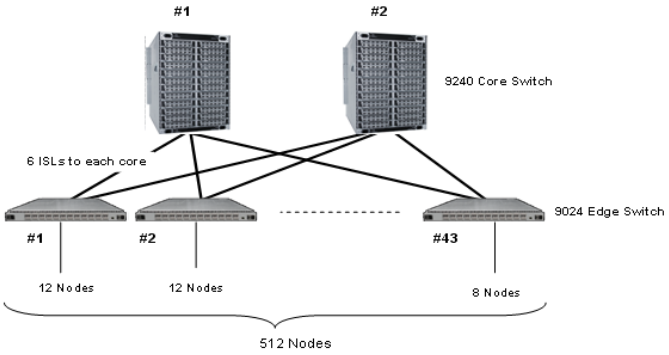


Figure 2-4 Two Tier Cluster



NOTE:Managing CBB and FBB clusters is the same. Managing a 1-tier cluster can be done using the switch GUI or CLI. Managing a 2-tier cluster is best done using the Fast Fabric toolset, which allows the user to silmultaneously manage multiple switches.

Redundancy: Depending on customer requirements, there are 3 fabric components that may require redundancy: power supplies, managed spines and Fabric Manager. In the absence of any requirements, follow the guidelines below:

Table 2-5. Power Supplies

Switch	Standard Configuration	Redundant Configuration
9240	**6 PS units**	**12 PS units**
9120	**3 PS units**	**6 PS units**
9080	**3 PS units**	**6 PS units**
9040	**2 PS units**	**4 PS units**
9020	**TBD**	**TBD**
9024	**1 PS unit**	**2 PS units**

Table 2-6. Managed Spines

Switch	Standard Configuration	Redundant Configuration
9240	**2 Managed Spines**	**4 Managed Spines**
9120	**1 Managed Spine**	**2 Managed Spines**
9080	**1 Managed Spine**	**2 Managed Spines**
9040	**1 Managed Spine**	**1 Managed Spine**
9020	**N/A**	**N/A**
9024	**N/A**	**N/A**

NOTE:Assuming a fully-populated MPFD chassis, a non-redundant power supply configuration will shut down a switch if a power supply unit(s) fails. If a managed spine in a non-redundaent configration fails, the user will be unable to manage the chassis.

NOTE:A 9240 MPFD has separate upper and lower hemispheres. Each hemisphere has its own management.

Table 2-7. Fabric Manager

Fabric	Standard Configuration	Redundant Configuration
<= 144 nodes	**1 Host-based FM**	**2 Host-based FM**
> 144 nodes	**1 Host-based FM**	**2 Host-based FM**

The Fabric Manager manages both physical and logical HCAs. Currently, the logical HCA is supported by the IBM P6 series. For more information refer to the applicable IBM documentation. Because of the importance of the subnet manager to a subnet, it is recommended to have redundant subnet managers.

NOTE:For more information on the Fabric Manager, refer to the QLogic Fabric Manager and Fabric Viewer Users Guide.

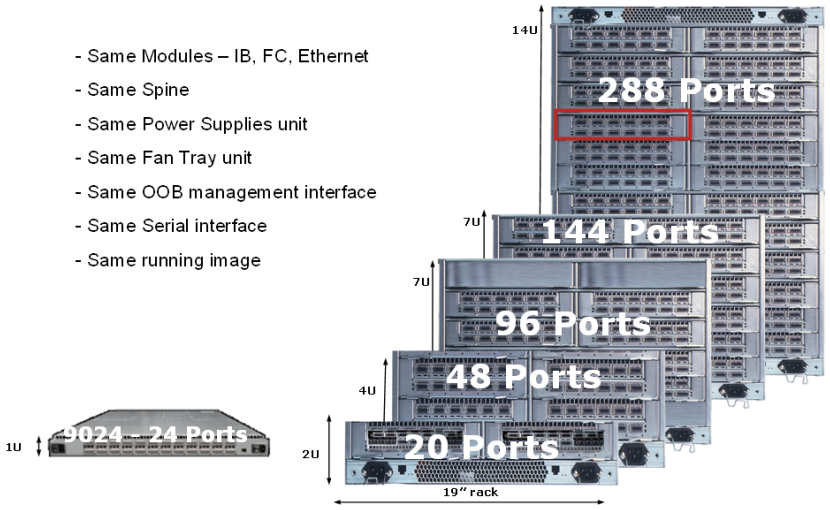
Cluster and Fabric Management

All fabrics require at least 1 subnet manager. Typically, two subnet managers will be used for redundancy. For smaller fabrics, the SM may be run within one or more of the core switches. For larger fabrics, the SM must be run on a host using the QuickSilver software and an HCA and OS supported by the software (e.g., SLES10). It is also recommended that all fabrics have the Fast Fabric tools installed on at least 1 node to facilitate fabric verification, analysis and monitoring. For more information, refer to the “Fabric Management” section of the *QLogic InfiniBand Best Practices Guide*.

Selecting InfiniBand Switches

Depending on the cluster size, it can be built using different SilverStorm InfiniBand switches (or switch family).

Figure 2-5 SilverStorm 9000 Switches



General Guidelines to Build a FBB Fabric

NOTE:The following numbers are based on a single subnet.

- 1 - 24 Nodes, use a 9024
- 25 - 48 Nodes, use a 9040
- 49 - 96 Nodes, use a 9080
- 97 - 144 Nodes, use a 9120
- 145 - 288 Nodes, use a 9240
- 289 to 1,728 use 9024 as leaf and 9120 at the core. For FBB, use the same number of cables between the leaf and core switches that are used between the leaf switches and server nodes.
- 1,729 to 3,456 use 9024 as leaf and 9240 at the core. For FBB, use the same number of cables between the leaf and core switches that are used between the leaf switches and server nodes.
- 2,437 to 10,368 use 9120 as leaf and 9240 as core
- 10,369 to 20,736 use 9240 as leaf and 9240 as core
- Over 20,736 then call QLogic

One-tier FBB Topology

To achieve FBB in the 9000-class of switches, the maximum number of spines must be installed regardless of how many leafs are to be used.

Table 2-8. Spines Required

Switch	Number of Spines for FBB
9240	**6**
9120	**3**
9080	**2**
9040	**1**
9020	**N/A**
9024	**N/A**

Two-tier FBB Topology

To achieve FBB, all of the core switches must be FBB (see Table 2-8) and all of the edge switches must have the following subscription ratio:

- Number of Computing Nodes/Number of ISLs = 1 (see Figure 2-2).

Floor Layout

For a large cluster, the floor layout planning is very important for two reasons:

1. Cooling considerations
2. Cable length determination

For such a floor layout, the following information must be determined:

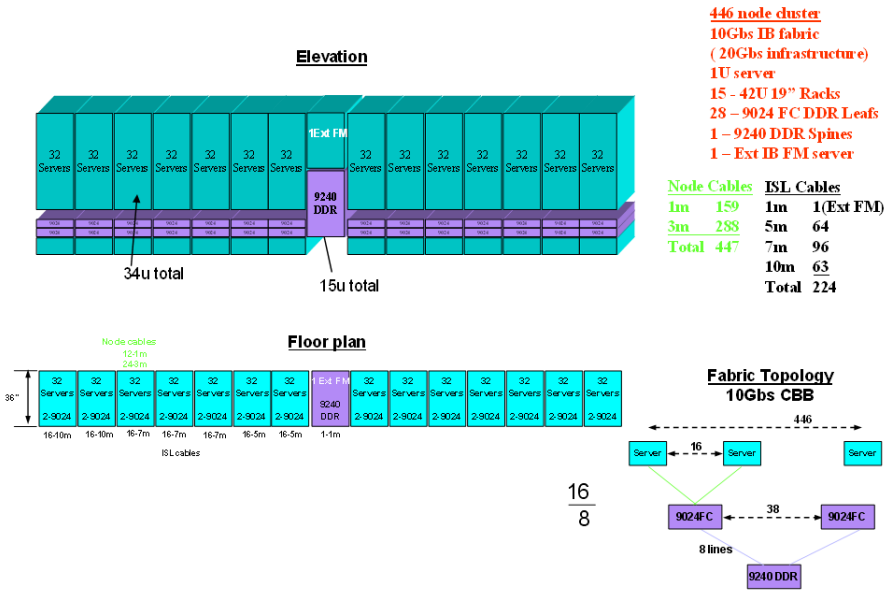
- Numbers of racks per row
- Numbers of rows
- Distance between two racks
- Distance between two rows
- Hot aisle and cool aisle

NOTE:

For switch specifcations, refer the the *SilverStorm 9000 Hardware Installation Guide*.

Following is an example:

Figure 2-6 Large Node Floor Layout



Calculate Cable Length

NOTE:For details on cables, refer to the sections "Switch and Cable Labeling" and "Cable Routing and Handling" in the QLogic InfiniBand Best Practices Guide.

Assuming only copper cables will be used, the available InfiniBand cable lengths offered are: 1, 2, 3, 5, 6, 7, 8, 9, 10, 12, and 20 Meters.

For IBM System p, cable options are as follows:

System p model number for IH servers with IBM GX HCA with QSFP connectors

- 6M passive QSFP-CX4 copper
- 10M active QSFP-CX4 copper
- 14M active QSFP-CX4 copper

System p model number for HV/HV8 servers with 12X HCA ports

- IBM "width changer" 12X-4X copper cable

BladeCenter-H with JS22 POWER blades & passive InfiniBand feed-through

- QLogic standard Gore CX4-CX4 copper cable offerings up to 8M (DDR only)

In order to calculate the cable length correctly, the following information needs to be collected:

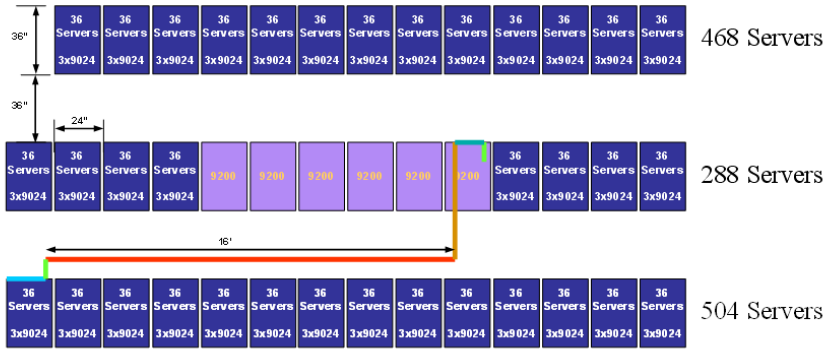
- Number of nodes
- Node size (1U, 2U, 4U) and type (shelf, blade)
- Rack size (42U typical height; 36 inches deep)
- Known floor plan / layout limitations
- If cables are run under a raised floor, need the depth of the raised floor
- If cables are run via overhead trays, need the height of the overhead trays above rack and/or cabinet
- Width of racks and/or cabinets
- Height of racks and/or cabinets
- Depth of racks and/or cabinets

Rules of thumb for using cables:

- SDR switches can support a maximum cable length of 20 meters.
- DDR switch:
  - Base (lower) ports:
    - Passive copper cables: 10 meters
    - Active copper: 14 meters
    - Fiber: 100 meters.
  - Mezzanine (upper) ports:
    - Passive copper cables: 8 meters
    - Active copper: 14 meters
    - Fiber: 100 meters.

Following is an example:

Figure 2-7 Cable Length Calculations



Example MAX Distance =

2 ft across the rack

3 ft Drop under raised floor

16 foot run (2' per rack)

6 feet across a row (allowing for standard 3' between racks)

2 ft across the rack

3 ft rise from under the floor

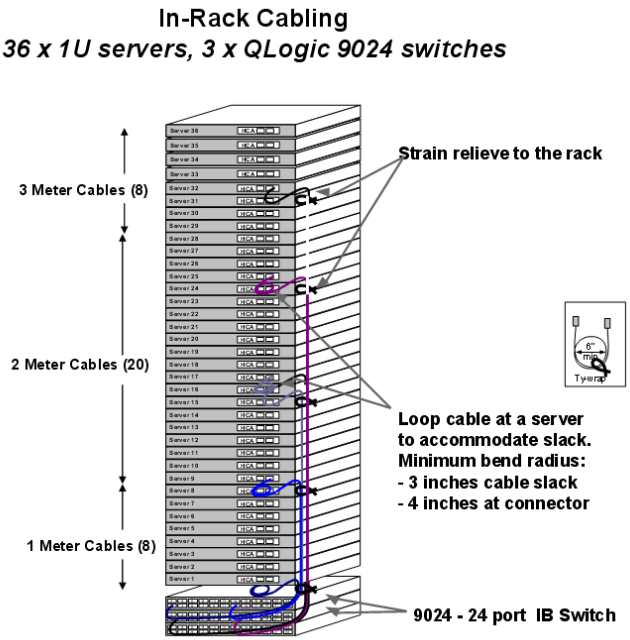
TOTAL = 2+3+16+6+2+3 = 32 feet ~ 10m

Cable Planning

- Determine the number (and size) of the compute nodes
- Identify the proper InfiniBand cable connectors such as CX4 (4x) or QSFP (12x)
- Be certain to allow sufficient slack to get the appropriate bend radius. For more information, refer to the section “Cable Routing & Handling” in the *QLogic InfiniBand Best Practices Guide*.
- Allow for space in the rack for the edge switches
- Assume that the first quarter of the rack can use 1m cables IF the edge switches are in the same rack
- Assume that the middle half of the rack is cabled using 2m cables
- Assume that the top quarter of the rack is cabled using 3m cables

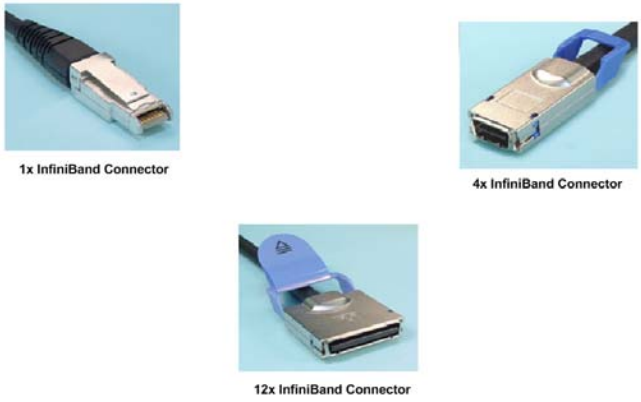
Following is a rack cabling example:

Figure 2-8 Rack Cabling



Following are InfiniBand cable connectors for various speeds:

Figure 2-9 InfiniBand Cable Connectors



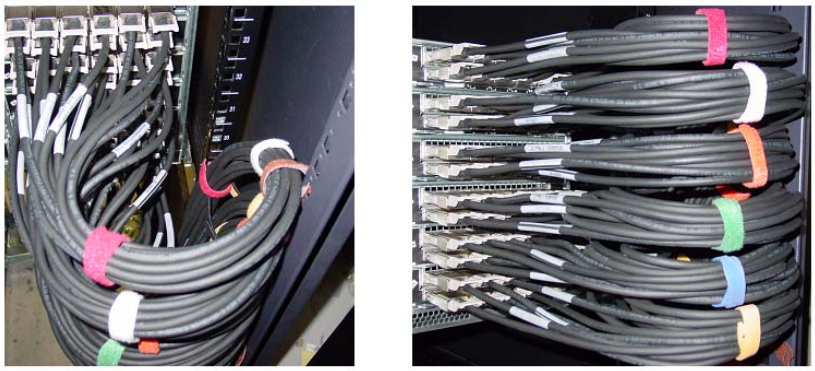
Cable Installation

Observe the following guidelines for cable installation:

- Cables should be routed and bundled to allow for the removal of a leaf.
- Cables should have at least three inch bend radius, otherwise damage to the cable may occur.

Following is an example:

Figure 2-10 Cable Installation



Quick Checklist

The following is a list of some basics to keep in mind at the customer site:

- Bring a laptop to customer site. If a laptop is not permitted on site, the customer should supply one.
- Bring a category 5 (Cat 5) LAN cable.
- Bring a RS 232 serial cable (provided by Qlogic) to access the switch.
- Have the customer pre-plan all cluster IP addresses before any installation takes place.
- Bring up one switch at time in order to change the default IP to the new IP address according to pre-planned list.