SGI® ICE™ X Installation and Configuration Guide

007–5917–001

The SGI Tempo systems management software stack, part of the SGI Management Center product, depends on several open source packages which require attribution. They are as follows:

c3:

C3 version 3.1.2: Cluster Command & Control Suite Oak Ridge National Laboratory, Oak Ridge, TN, Authors: M.Brim, R.Flanery, G.A.Geist, B.Luethke, S.L.Scott (C) 2001 All Rights Reserved NOTICE Permission to use, copy, modify, and distribute this software and # its documentation for any purpose and without fee is hereby granted provided that the above copyright notice appear in all copies and that both the copyright notice and this permission notice appear in supporting documentation. Neither the Oak Ridge National Laboratory nor the Authors make any # representations about the suitability of this software for any purpose. This software is provided "as is" without express or implied warranty. The C3 tools were funded by the U.S. Department of Energy.

conserver:

Copyright (c) 2000, conserver.com All rights reserved. Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer. - Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution. - Neither the name of conserver.com nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission. THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE REGENTS OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Copyright (c) 1998, GNAC, Inc. All rights reserved. Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met: - Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer. - Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution. - Neither the name of GNAC, Inc. nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission. THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE REGENTS OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Copyright 1992 Purdue Research Foundation, West Lafayette, Indiana 47907. All rights reserved. This software is not subject to any license of the American Telephone and Telegraph Company or the Regents of the University of California. Permission is granted to anyone to use this software for any purpose on any computer system, and to alter it and redistribute it freely, subject to the following

Record of Revision

Version	Description
001	May 2013 Original publication. This revision supports the SGI Management Center 1.7 release.

Contents

About This Guide . . . . . **xvii**
Related Publications xvii
Obtaining Publications xviii
Conventions xviii
Reader Comments xix

1. SGI ICE X System Software Overview **1**
SGI ICE X System Component Overview 1
SGI ICE X Networks 3

2. Customizing a Factory-installed SGI ICE X System **7**
About Customizing a Factory-installed SGI ICE X System 7
Obtaining Information 8
Changing the Password and Specifying Network Information 9
Completing the Customization 11

3. Installing and Configuring an SGI ICE X System **17**
About Performing a New Installation and Configuring the Software on an SGI ICE X System 18
Planning the Boot Slots and Disk Partitions 21
Boot Parameters 22
Partition Layout for a Two-slot SGI ICE X System (Default) 25
Partition Layout for a Five—slot SGI ICE X System 26
Partition Layout for a One-slot (Single—boot) SGI ICE X System 26
Preparing to Install Software on an SGI ICE X System 27

(Conditional) Setting a Static IP Address for the Baseboard Management Controller (BMC) in the System Admin Node (SAC) 29

(Optional) Configuring a Highly Available (HA) System Admin Controller (SAC) . . . 31

Booting the System 31

Installing the Operating System 33

Installing SUSE Linux Enterprise Server (SLES) 33

Installing Red Hat Enterprise Linux (RHEL) 37

Running the Cluster Configuration Tool 42

(Conditional) Configuring External Domain Name Service (DNS) Servers 51

Installing the SGI Management Center License Key 52

Synchronizing the Software Repository, Installing Software Updates, and Cloning the Images 53

Configuring the Switches 55

Configuring Management Switches With a MAC File 56

Configuring Management Switches Without a MAC File 59

(Conditional) Configuring the MCell Network 61

Configuring the Rack Leader Controllers (RLCs) and Service Nodes with the `discover` Command 65

(Optional) Configuring a Backup Domain Name Service (DNS) Server 68

Configuring the InfiniBand Subnetworks 69

4. Configuring Optional Features 75

About the Optional SGI ICE X Features 75

Configuring a Service Node as a Network Address Translation (NAT) Gateway 76

Configuring a File System on a Service Node for use with a Network File System (NFS) Server 80

Configuring a Service Node as a Network File System (NFS) Server 84

Configuring Service Nodes and/or Compute Nodes as Network Information Service (NIS) Clients to the House Network's NIS Server 88

Configuring a Service Node as a NIS Client 89

Configuring a Compute Node as a NIS Client 90
Method 1 — Configuring an Individual Compute Node as a NIS Client 90
Method 2 — Configuring the Master Compute Node Image as a NIS Client 92
Propagating a Node's Configuration to Another Node 94
RHEL Service Node House Network Configuration 95
Configuring a Service Node as a Network Information Service (NIS) Server 96
Configuring a Network Information Service (NIS) Master Server and One or More NIS Slave Servers 97
Configuring a Network Information Service (NIS) Client on a Service Node 99
Configuring a Rack Leader Controller (RLC) as a Network Information Server (NIS) Slave Server and Client (SLES) 100
Configuring the Compute Nodes as Network Information Service (NIS) Clients (SLES) . 102
NAS Configuration for Multiple IB Interfaces 103
Creating User Accounts (SLES) 106
Installing MPI on a Running SGI ICE X System 106
Troubleshooting Configuration Changes 109

5. Troubleshooting 111

Initial Installation Settings 112
System Discovery Overview 112
`configure-cluster` Command 113
`cmcdetected` Daemon 113
`discover` Command 114
`blademond` Daemon 114
Compute Nodes Are Taking Too Long To Boot 115
Verify the Bonding Mode on the Rack Leader Controller (RLC) 116
`cimage --push-rack` Pushes Too Many (or Too Few) Expansions 119
Cannot `ping` the CMCs from the Rack Leader Controller (RLC) 120

r1lead Configured with vlan1/vlan2 and Not vlan101 122
How to Make the blademond Daemon Start Over from Scratch 122
Log Files 123
CMC slot_map / blademond Debugging Hints 123
ssh Commands to Compute Nodes: ssh Key Failures / Known Hosts 125
Compute Node Hosts Seem to Actually be BMCs 125
Resolving CMC Slot Map Ordering Issues 125
In tmpfs Mode, File Has Date in the Future Warnings 126
Ensuring Hardware Clock Has the Correct Time 126
Configure Switches for a Rack Leader Controller (RLC) 127
Switch Wiring Rules 129
System Admin Controller (SAC) eth2 Link in the Bond is Down 130
No InfiniBand Interfaces on Rack Leader Controller (RLC), Service, or Compute Node Images 131
Troubleshooting a Network Address Translation (NAT) Configuration 132

Appendix A. YaST2 Navigation 135

Appendix B. Virtual Local Area Network (VLAN) Information 137

About VLANs 137
VLAN Ethernet Network Configurations 137
Head VLAN 138
Rack VLANs 139
SGI ICE X IP Address Ranges and VLANs for Management and Application Software . . 141
Component Naming Conventions 143
System Control Configuration 145

Appendix C. MCell Network IP Addresses 149

Index 155

Figures

Figure 1-1 Example SGI ICE X System 2

Figure 1-2 Three VLANs 4

Figure 3-1 SGI ICE X Software Installation Process 19

Figure 3-2 SAC Power On Button and DVD Drive 31

Figure 3-3 **Network Card Setup** Screen 35

Figure 3-4 **Initial Configuration Check** Screen 43

Figure 3-5 **Initial Cluster Setup** Screen with the initial screen 44

Figure 3-6 **Initial Cluster Setup Tasks** Screen 45

Figure 3-7 **Configure House DNS Resolvers** Screen 49

Figure 3-8 **Configure Switch Management Network** screen 50

Figure 3-9 **Configure Backup DNS Server (service node)** screen 69

Figure 3-10 Completed InfiniBand (`ib0`) Master / Standby Screen 71

Figure B-1 VLAN Overview 138

Figure B-2 HEAD VLAN Ethernet Connections 139

Figure B-3 RACKx VLAN Ethernet Connections 141

Figure B-4 Redundant Cascaded Switch Configuration 146

Figure B-5 Nonredundant Cascaded Switch Network Configuration 147

Examples

Example 5-1 `tcpdump` Command Examples 132

Procedures

Procedure 2-1 To obtain information for the customization 8
Procedure 2-2 To customize a factory-installed RHEL operating system 9
Procedure 2-3 To customize the cluster database 11
Procedure 3-1 To plan the boot slots and disk partitions 22
Procedure 3-2 To prepare for an installation 27
Procedure 3-3 Method 1 — To change from the BIOS 29
Procedure 3-4 Method 2 — To change the IP address from the SAC 29
Procedure 3-5 To boot the system 31
Procedure 3-6 To install SLES 11 SP2 on an SGI ICE X SAC 33
Procedure 3-7 To install RHEL 6 on an SGI ICE X SAC 38
Procedure 3-8 To run the cluster configuration tool 42
Procedure 3-9 To configure external DNS servers 52
Procedure 3-10 To license the SMC software 53
Procedure 3-11 To update the software 53
Procedure 3-12 To configure switches — with a MAC file 56
Procedure 3-13 To configure switches — without a MAC file 59
Procedure 3-14 To configure MCell switches 61
Procedure 3-15 To configure the RLCs and service nodes 65
Procedure 3-16 To enable a backup DNS 68
Procedure 3-17 To configure the InfiniBand network 69
Procedure 4-1 To enable NAT on a service node 76
Procedure 4-2 To configure an NFS home server on a service node 81
Procedure 4-3 To configure an NFS server on a service node 84

Procedure 4-4 To configure a service node as a NIS client 89
Procedure 4-5 To log into a compute node and configure that compute node as a NIS client 91
Procedure 4-6 To log into the SAC and edit the master compute node image 92
Procedure 4-7 To propagate NIS client configuration 94
Procedure 4-8 To configure a service node as a NIS master server 98
Procedure 4-9 To configure a service node as a NIS client 99
Procedure 4-10 To configure an RLC as a NIS slave server 100
Procedure 4-11 To configure the compute nodes as NIS clients 102
Procedure 4-12 To configure NAS 104
Procedure 4-13 Creating User Accounts on a NIS Server 106

About This Guide

This guide is a reference document for people who install and configure SGI® ICE™ X systems. It describes how to perform general system discovery, installation, configuration, and operations.

Related Publications

The following additional documentation might be useful to you:

- *SGI ICE X Administration Guide*, publication 007–5918–*xxx*

  This manual explains how to manage an SGI ICE X system.

- *SGI Management Center (SMC) Installation and Configuration*, publication 007–5643–*xxx*

  This manual is intended for system administrators. It describes how to install and configure the SGI Management Center.

- *SGI Management Center for SGI ICE*, publication 007–5718–*xxx*

  This manual describes how you can monitor and control a cluster using the SGI Management Center.

- *SGI ICE X System Hardware User Guide*, publication 007–5806–*xxx*

  This is the hardware user's guide for the SGI ICE X systems. It describes the hardware features of the SGI ICE X system, as well as, troubleshooting, upgrading, and repairing.

- *SGI Performance Suite X.X Start Here*, publication 007–5680–*xxx*

  This manual describes the most recent release of the SGI Performance Suite and lists the current SGI software and hardware manuals.

- Documentation from other sources:

  – SUSE documentation for SUSE Linux Enterprise Server 11 (SLES 11)

  – Red Hat documentation for Red Hat Linux Enterprise Server 6 (RHEL 6)

  – Intel compiler documentation

– Intel documentation about Xeon architecture

Obtaining Publications

You can obtain SGI documentation in the following ways:

- See the SGI Technical Publications Library at: http://docs.sgi.com. Various formats are available. This library contains the most recent and most comprehensive set of online books, release notes, man pages, and other information.
- Online versions of the *SGI Performance Suite X.X Start Here*, release notes, which contain the latest information about software and documentation for each SGI Performance Suite product, the list of RPMs distributed with each product can be found in the `/docs` directory on each SGI Performance Suite product media.
- You can view man pages by typing `man` *title* on a command line.

Conventions

The following conventions are used throughout this document:

Convention	Meaning
`command`	This fixed-space font denotes literal items such as commands, files, routines, path names, signals, messages, and programming language structures.
variable	Italic typeface denotes variable entries and words or concepts being defined.
`user input`	This bold, fixed-space font denotes literal items that the user enters in interactive sessions. (Output is shown in nonbold, fixed-space font.)
[ ]	Brackets enclose optional portions of a command or directive line.

...	Ellipses indicate that a preceding element can be repeated.

Reader Comments

If you have comments about the technical accuracy, content, or organization of this publication, contact SGI. Be sure to include the title and document number of the publication with your comments. (Online, the document number is located in the front matter of the publication. In printed publications, the document number is located at the bottom of each page.)

You can contact SGI in any of the following ways:

- Send e-mail to the following address:

  techpubs@sgi.com

- Contact your customer service representative and ask that an incident be filed in the SGI incident tracking system.

- Send mail to the following address:

  SGI
  Technical Publications
  46600 Landing Parkway
  Fremont, CA 94538

SGI values your comments and will respond to them promptly.

SGI ICE X System Software Overview

This chapter includes the following topics:

- "SGI ICE X System Component Overview" on page 1
- "SGI ICE X Networks" on page 3

SGI ICE X System Component Overview

The SGI Integrated Compute Environment (ICE) X systems provide an integrated compute node (blade) environment that can include thousands of compute nodes. The SGI Management Center (SMC) software for SGI ICE X systems enables you to provision, install, configure, and manage your system. This manual describes how to configure the software on the system for use at your site. This manual does not include an SGI ICE X hardware description. For hardware information, see the *SGI ICE X System Hardware User Guide*.

Figure 1-1 on page 2 shows a simple SGI ICE X system.

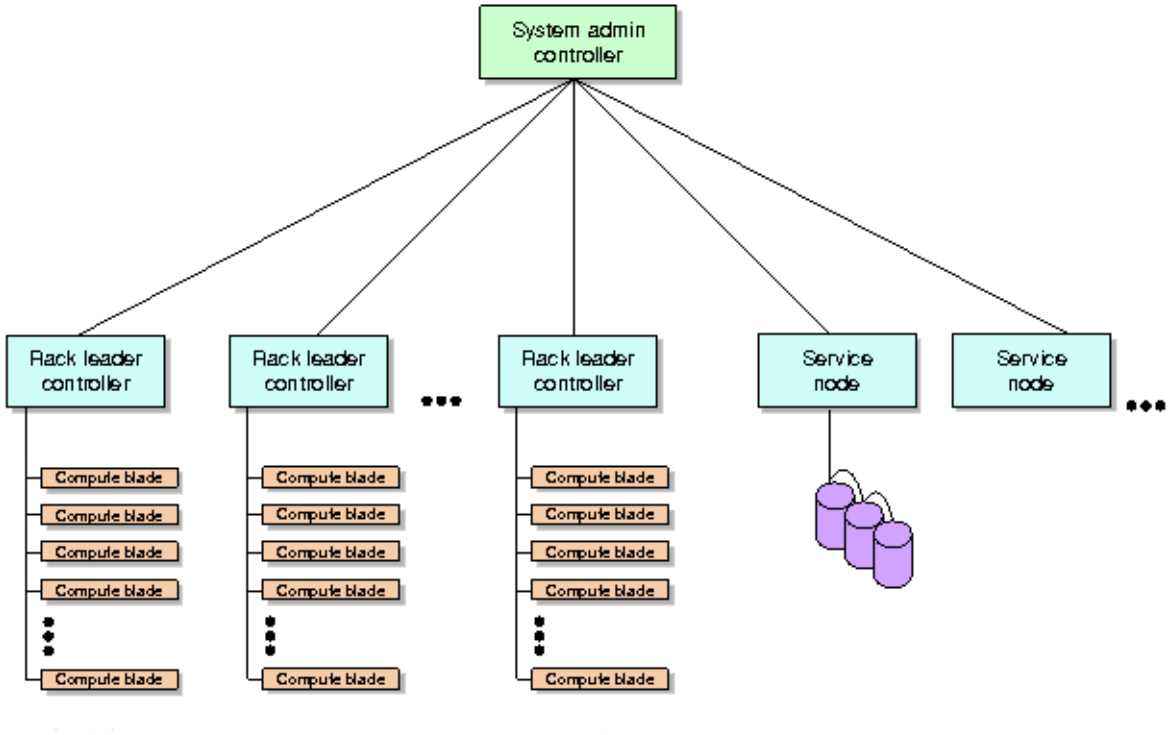

Figure 1-1 Example SGI ICE X System

SGI ICE X system processes are distributed across several controllers and service nodes. Each node has a specific role. These components are as follows:

- One system admin controller (SAC). The site system administrator can log into the SAC as the root user to monitor the system, to modify the master software images, and to perform general system administration tasks.
- One rack leader controller (RLC) per compute node (blade) rack. You can have multiple compute nodes in the racks, and you can have multiple racks. Each RLC manages a set of compute nodes in a particular rack.
- One or more rack leader controllers (RLCs). The RLCs manage the compute nodes (blades) in the rack enclosures.

- One or more service nodes. End users can log into a service node to run jobs. There can be several service nodes, and they can host single services or multiple services. The SGI Management Center software includes the service nodes, but SGI Management Center does not include the utilities that enable users to log in, run batch jobs, and so on. The service nodes can host one or more of the following types of services:
  - Login services. These services allow an end user to log in and then, for example, run or monitor MPI jobs.
  - Batch scheduling services. You can install schedulers such as Altair's PBS Professional or TORQUE.
  - I/O gateway. On a small system, you can combine the I/O gateway, login services, and batch scheduling on the same node. The I/O gateway services connect the SGI ICE X system to your house network. You can configure one or more of the following protocols on the service node: network file system (NFS), network address translation (NAT), network information service (NIS).
  - Storage. A storage service node is a network attached (NAS) appliance bundle that provides InfiniBand attached storage for the system.
  - Object storage server. This service is used in Lustre File Storage configurations.
  - Metadata server. This service is used in Lustre File Storage configurations.

The software distribution includes the master system image for the SAC. During installation and configuration process, the installation software creates the master system images for the RLCs, the service nodes, and the compute nodes. As you customize the system for your site, you can modify the node-specific system images on the SAC and push the updated images to the RLCs, to the service nodes, or to the compute nodes.

SGI ICE X Networks

The SGI ICE X system includes several virtual local area networks (VLANs). Figure 1-2 on page 4 is a logical representation of the SGI ICE X Ethernet network that shows three VLANs in an example SGI ICE X system.

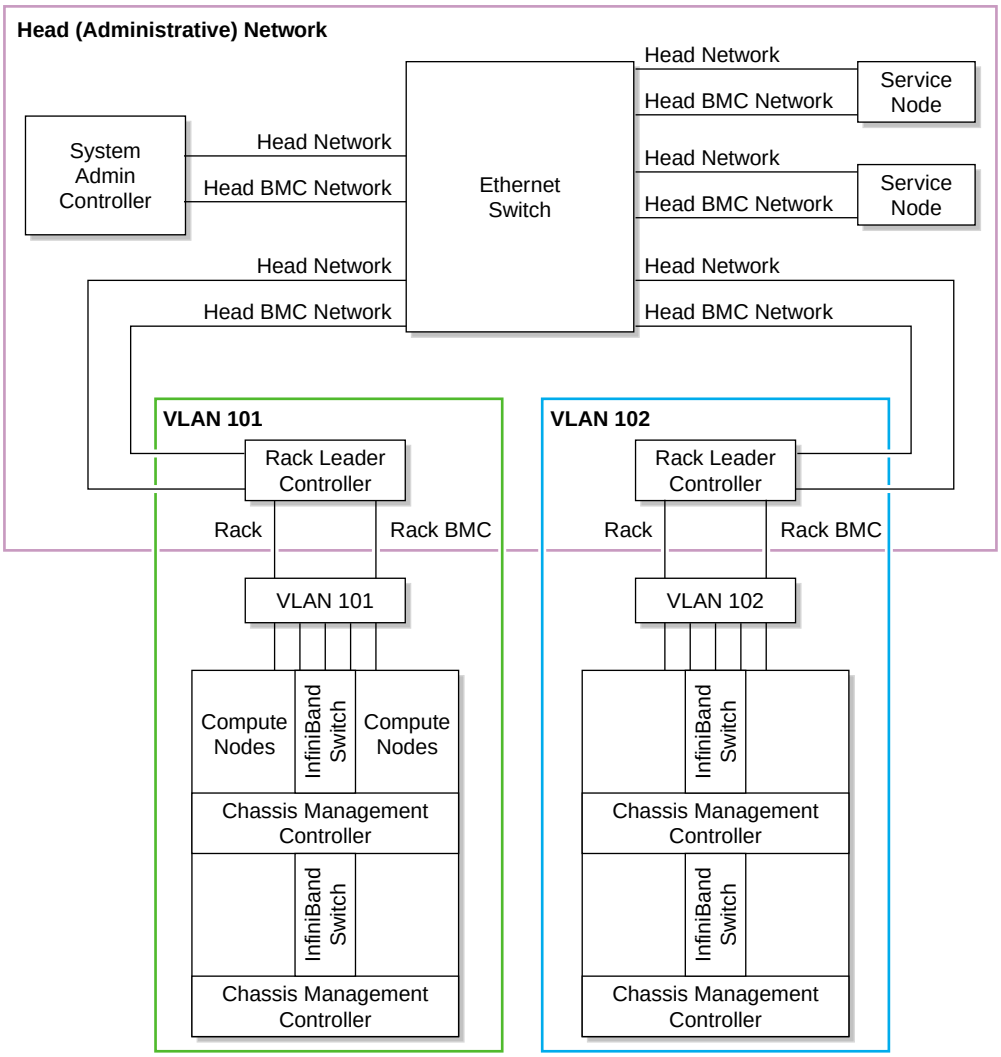

Figure 1-2 Three VLANs

In Figure 1-2 on page 4, the head network is VLAN 1. The SAC and service nodes are attached to the customer LAN.

The system components are attached to one or more of the following two internal networks:

- The high-performance network, which is designed for computation. This InfiniBand network facilitates communication to all compute nodes from the service nodes. It connects the following:
  - The compute nodes to each other. The IB network connects all the compute nodes (blades) to each other. The compute nodes are not part of the head network.
  - The service nodes to the compute nodes.
- The head network, which is the administrative network. This Ethernet network is designed for communication between the SAC, RLCs, and service nodes. These components communicate to each other directly within the head network. The head network includes the following Ethernet connections:
  - The SAC to the Ethernet switches.
  - The Ethernet switches to the SAC, RLCs, and service nodes.
  - The service nodes to the Ethernet switches.

One or two separate InfiniBand networks segregate traffic within the SGI ICE X system in a way that optimizes computing performance. When there are two InfiniBand networks, communication is segregated by network, as follows:

- `ib0`, which is typically used for Message Passing Interface (MPI) communication.
- `ib1`, which is typically used for storage traffic.

Users can log into the SAC and into the service nodes directly. If you need to reach the RLC, you need to log into the SAC first, and then `ssh`(1) into the RLC.

The RLCs and compute nodes are attached to a separate Ethernet network. An Ethernet network connects each RLC to each of the compute nodes that the RLC controls. Communication to the compute nodes always goes through the RLC. Only the RLC for a particular rack can communicate with the compute nodes on its own rack. Each compute node has its own Ethernet IP address within its rack. If you have multiple racks, compute node IP addresses are duplicated across the racks.

Customizing a Factory-installed SGI ICE X System

This chapter contains the following topics:

- "About Customizing a Factory-installed SGI ICE X System" on page 7
- "Obtaining Information" on page 8
- "Changing the Password and Specifying Network Information" on page 9
- "Completing the Customization" on page 11

About Customizing a Factory-installed SGI ICE X System

Your SGI ICE X system was tested and configured at the factory. At the factory, SGI configured the following:

- A factory-specified root password. One of the first steps in the configuration procedure is to change this root password on the system admin controller (SAC) node.
- Two slots. SGI configured the operating system that you ordered on slot 1. The operating system can be either Red Hat Enterprise Linux (RHEL) or SUSE Linux Enterprise Server (SLES). Slot 2 is blank.

  The SGI ICE X system supports a maximum of five slots. If you need more than two slots, you need to reconfigure the system. During the reconfiguration, you reinstall the operating system and perform many other tasks. For the reconfiguration procedure, see Chapter 3, "Installing and Configuring an SGI ICE X System" on page 17.
- A console configured with a serial port.

If you want to retain the factory-installed configuration, complete the following procedures:

- "Obtaining Information" on page 8
- "Changing the Password and Specifying Network Information" on page 9
- "Completing the Customization" on page 11

Obtaining Information

Your configuration session can proceed more quickly if you gather some information before you start. You need to update the factory-installed, system-wide root password and the time zone. In addition, you need to provide information about your site network for the SAC's `eth0` network interface card (NIC).

The following procedure explains the information that you need to gather.

Procedure 2-1 To obtain information for the customization

1. Complete the following table:

Information Needed	Specifics for this SGI ICE X System
Factory-installed password	_____
Password for this system at your site	_____
Time zone	_____
IP address	_____
Netmask	_____
Hostname	_____
Default route/Gateway	_____
Fully qualified domain name (FQDN)	_____
House NTP server	_____
First house (site) DNS resolver IP address	_____
(Optional) Second house DNS resolver IP address	_____
(Optional) Third house DNS resolver IP address	_____
House (site) domain	_____
SGI ICE X system subdomain name	_____

2. Proceed to the following:

"Changing the Password and Specifying Network Information" on page 9

Changing the Password and Specifying Network Information

The following procedure explains how to change the password on the SAC and how to update the operating system configuration files with your site's networking information.

Procedure 2-2 To customize a factory-installed RHEL operating system

1. Use the console attached to the SAC, and log into the SAC as the root user.
2. Use the `cpasswd` command to change the root passwords on all nodes.

   Obtain the password used for the factory installation from your SGI representative.

   The `cpasswd` command changes the root password on the SAC and on all other nodes. To obtain a usage statement, type the following command:

   ```
   # cpasswd --h
   ```

   For example:

   ```
   SAC:~ # cpasswd
   Enter new password:
   Enter new password (again):
   admin: updating /etc/shadow
   r1lead: updating /etc/shadow
   service0: updating /etc/shadow
   SAC:~ #
   ```

3. Change the system time zone.

   Type the following command:

   ```
   # system-config-date
   ```

   The `system-config-date` command starts a graphical user interface (GUI) tool. Within the GUI tool, change **only** the system time zone. The tool enables you to change other aspects of the configuration, but for this step, change only the system time zone.

Note: Do not use this tool to change the NTP server, the time, or other configuration data.

---

4. Use a text editor, such as `vi` or `vim`, to open file `/etc/sysconfig/network-scripts/ifcfg-eth0`.
5. Edit the `/etc/sysconfig/network-scripts/ifcfg-eth0` file.

   Add lines for the `IPADDR` and `NETMASK` values appropriate for your public (house) network. Also set `ONBOOT=yes`.

   For example:

   ```
   IPADDR=128.162.244.88
   NETMASK=255.255.255.0
   ONBOOT=yes
   ```

6. Save and close file `/etc/sysconfig/network-scripts/ifcfg-eth0`.
7. Use a text editor to create file `/etc/sysconfig/network`.
8. Add the following three lines to file `/etc/sysconfig/network`:

   ```
   NETWORKING=yes
   HOSTNAME=SAC_hostname
   GATEWAY=gateway_IP_address
   ```

   For *SAC_hostname*, type the hostname you want to assign to the SAC.

   For *gateway_IP_address*, type the IP address of the gateway for your house network.

   For example:

   ```
   NETWORKING=yes
   HOSTNAME=my-system-admin
   GATEWAY=128.162.244.1
   ```

9. Save and close file `/etc/sysconfig/network`.
10. Use a text editor to open file `/etc/hosts`.
11. Add a line in the following format to file `/etc/hosts`:

    SAC_IP SAC_FQDN SAC_hostname

    The variables in the preceding line are as follows:

    - For *SAC_IP*, type the IP address of the SAC.

- For *SAC_FQDN*, type the fully qualified domain name (FQDN) of the SAC.
- For *SAC_hostname*, type the hostname of the SAC.

For example, add the following line:

```
128.162.244.88 acme-admin.acme.usa.com acme-admin
```

12. Save and close file `/etc/hosts`.
13. Type the following command to set the SAC hostname:

    # **hostname *SAC_hostname***

    For *SAC_hostname*, type the hostname of the SAC.

    For example:

    # **hostname acme-admin**

14. Proceed to the following:

    "Completing the Customization" on page 11

Completing the Customization

The procedure in this topic explains how to use the cluster configuration tool to add information about your site's network to the cluster database.

For more information about the cluster configuration tool, see "Running the Cluster Configuration Tool" on page 42.

Procedure 2-3 To customize the cluster database

1. Type the following command to start the cluster configuration tool:

   # **/opt/sgi/sbin/configure-cluster**

2. On the cluster configuration tool's main menu select **Configure the Time Client/Server (NTP)** and select **OK**.

   The system guides you through the process specify your house NTP server in file `/etc/ntp.conf`. This process differs, depending on your platform, as follows:

   - On RHEL platforms, follow the the instructions that the cluster configuration tool presents to you.

- On SLES platforms, the cluster configuration tool opens a YaST2 menu. Follow the prompts in the YaST2 menu to set your NTP server.

3. On the cluster configuration tool's main menu select **Configure House DNS Resolvers** and select **OK**.

   You can specify up to three house DNS resolvers.

4. Select **Quit** and select **OK** to log out from the cluster configuration tool.

5. Type the following command to set the house (site) domain:

   # **cadmin --set-admin-domain *site_domain***

   For *site_domain*, specify the full name of your house domain. For example, `usa.acme.com`.

6. Type the following command to change the subdomain name for the SGI ICE X system:

   # **cadmin --set-subdomain *cluster_name***

   For *cluster_name*, specify the name of the system. For example, `sleet.usa.acme.com`.

   For more information about the `cadmin` command, type `cadmin -h` at the system prompt.

7. Type the following command to retrieve the path to the system images:

   # **cinstallman --show-images**

   An SGI ICE X system includes the following major nodes:

   - One system admin controller. This node is also called the *SAC* or the *admin node*.
   - One or more rack leader controllers. These nodes are also called the *RLCs*. There is one RLC per rack.
   - Several compute nodes. These nodes are also called *blades*. The compute nodes are housed in racks, and each rack has one RLC.
   - One or more service nodes. Users log into the system through the service nodes to run jobs.

The SAC hosts the master images for each of the preceding node types. If you need to change some aspect of a node's configuration, SGI recommends that you change the configuration in the master image and push out the changed master image to the affected nodes. This practice maintains consistency between the master images on the SAC and the production images on the nodes.

The following example shows how to retrieve the paths to the compute node images on the SAC:

```
# cinstallman --show-images
Image Name               BT Path
compute-rhel6.3          1  /var/lib/systemimager/images/compute-rhel6.3
service-rhel6.3          0  /var/lib/systemimager/images/service-rhel6.3
lead-rhel6.3             0  /var/lib/systemimager/images/lead-rhel6.3
```

The following steps explain how to update the master node images with your site's time zone information.

8. Type the cp(1) command three times (one for each master node image) to set the time zone in the system images.

   The format of this command is as follows:

```
cp /etc/localtime /var/lib/systemimager/images/image_name/etc
```

   For *image_name*, type the name of one of the system images you retrieved with the preceding `cinstallman` command. Type one of these commands for each system image.

   For example, type the following three commands:

```
# cp /etc/localtime /var/lib/systemimager/images/compute-rhel6.3/etc
# cp /etc/localtime /var/lib/systemimager/images/service-rhel6.3/etc
# cp /etc/localtime /var/lib/systemimager/images/lead-rhel6.3/etc
```

9. Type the following two commands to propagate the time zone information to the images on the RLCs and service nodes:

```
# pdcp -g leader /etc/localtime /etc/localtime
# pdcp -g service /etc/localtime /etc/localtime
```

   This step explains how to update the RLCs and service nodes. The following steps explain how to update the compute nodes.

10. Type the following command to stop the compute nodes:

    # **cpower --halt r*i*n***

11. (Optional) Provide information about the number of racks on your system.

    Perform this step if you have a small system with fewer than eight IRUs per RLC.

    The procedure pushes the updated compute image to all the compute nodes. This process can run for a long time on large systems. If you have a large number of IRUs, you need the system to perform expansions that enable you to change many compute nodes at a time. If you have fewer than either IRUs per RLC, however, the expansions are not needed. The following substeps explain how to prepare the system to work on a smaller number of compute nodes.

    a. Type the following command to retrieve the identifiers for the RLCs on your system:

       # **cnodes --leader**

    b. Type one or more of the following commands to suppress unnecessary processing:

       `cadmin --set-max-irus --node` *rlc_id number_of_racks*

       For *rlc_id*, specify the identifier for one of the RLCs in your system.

       For *number_of_racks*, specify the number of IRUs associated with this RLC.

    For example, the following command specifies that there is only one IRU associated with the RLC identified as `r1lead`:

    # **cadmin --set-max-irus --node r1lead 1**

12. Push the time zone changes to all the compute nodes.

    Use the `cimage` command in the following format:

    `cimage --push-rack compute` *image_name* `r\*`

    For *image_name*, specify the name of the compute node image.

    For example:

    # **cimage --push-rack compute-rhel6.3 r***

13. Type the following command to power-up the compute nodes:

    # **cpower --boot r*i*n***

14. (Optional) Configure additional features.

    The SGI ICE X system supports several optional features, for example, networking features such as network address translation. For information about how to configure optional features, see the following:

    Chapter 4, "Configuring Optional Features" on page 75

Installing and Configuring an SGI ICE X System

This chapter contains the following topics:

- "About Performing a New Installation and Configuring the Software on an SGI ICE X System" on page 18
- "Planning the Boot Slots and Disk Partitions" on page 21
- "Preparing to Install Software on an SGI ICE X System" on page 27
- "(Conditional) Setting a Static IP Address for the Baseboard Management Controller (BMC) in the System Admin Node (SAC)" on page 29
- "(Optional) Configuring a Highly Available (HA) System Admin Controller (SAC)" on page 31
- "Booting the System" on page 31
- "Installing the Operating System" on page 33
- "Running the Cluster Configuration Tool" on page 42
- "(Conditional) Configuring External Domain Name Service (DNS) Servers" on page 51
- "Installing the SGI Management Center License Key" on page 52
- "Synchronizing the Software Repository, Installing Software Updates, and Cloning the Images" on page 53
- "Configuring the Switches" on page 55
- "Configuring the Rack Leader Controllers (RLCs) and Service Nodes with the `discover` Command" on page 65
- "(Optional) Configuring a Backup Domain Name Service (DNS) Server" on page 68
- "Configuring the InfiniBand Subnetworks" on page 69

About Performing a New Installation and Configuring the Software on an SGI ICE X System

SGI installs operating system software on each SGI ICE X system before factory shipment occurs. The topics in this chapter include the additional procedures that you need to complete in order to configure the system for your site.

If you want to completely reinstall the operating system and all other software, the topics in this chapter enable you to complete that task. For example, you might need to reinstall the operating system to meet site requirements or to recover a system in case of a disaster.

Figure 3-1 on page 19 depicts the software installation process.

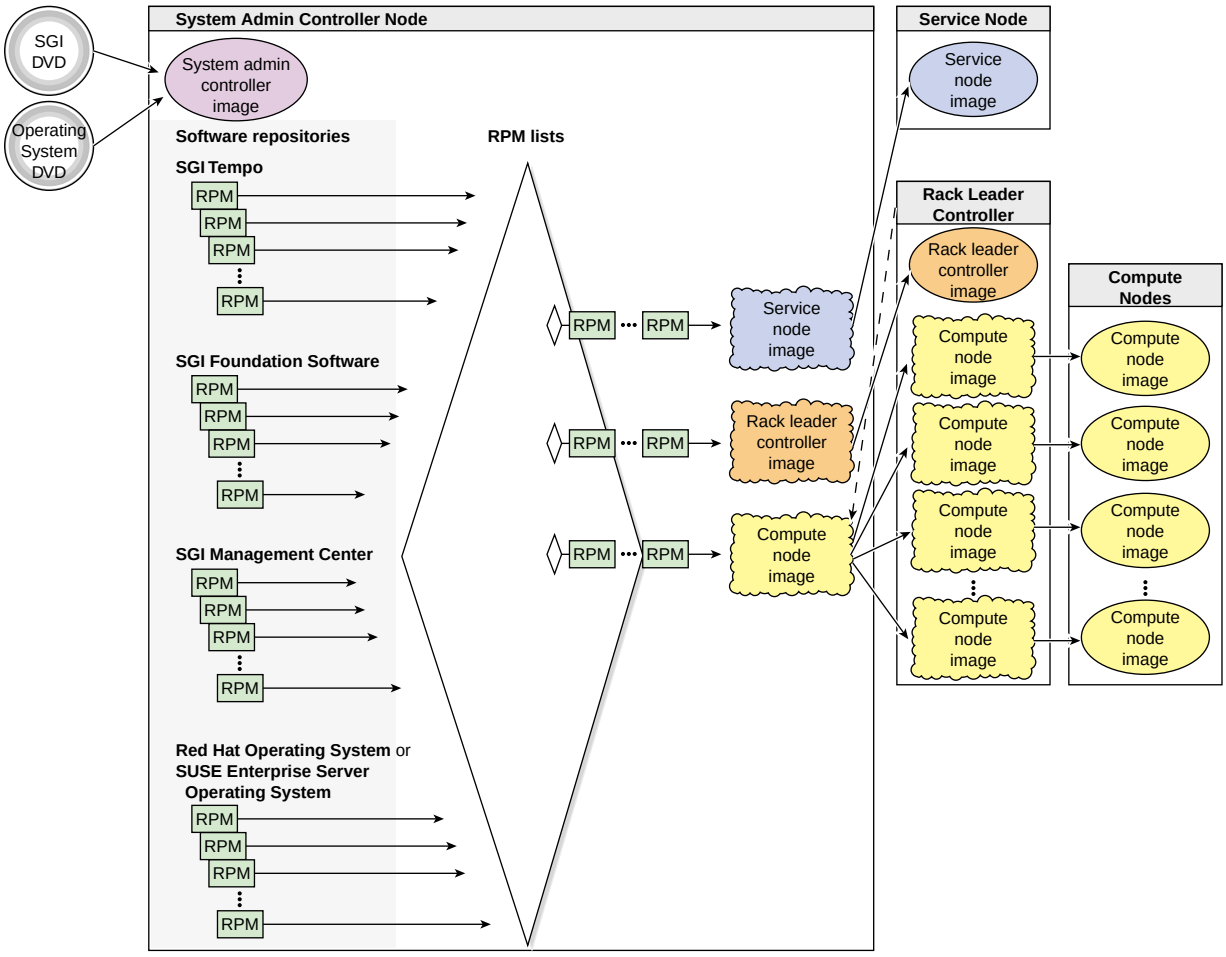

Figure 3-1 SGI ICE X Software Installation Process

.

Table 3-1 on page 19 shows the installation and configuration procedures to follow if you want to install the SGI ICE X system from scratch. In this case, you reinstall the operating system on the nodes and configure everything yourself.

Table 3-1 SGI ICE X System Installation and Configuration Process

Step	Task	See
1	Plan the boot slots and partitions.	"Planning the Boot Slots and Disk Partitions" on page 21
2	Prepare to install the SGI ICE X software.	"Preparing to Install Software on an SGI ICE X System" on page 27
3	(Conditional) Configure a static address for the baseboard management controller (BMC) on the system admin controller (SAC). Perform this step only if your site practices require a static IP on the BMC.	"(Conditional) Setting a Static IP Address for the Baseboard Management Controller (BMC) in the System Admin Node (SAC)" on page 29
4	(Optional) Configure a highly available system admin controller	"(Optional) Configuring a Highly Available (HA) System Admin Controller (SAC)" on page 31
5	Boot the system.	"Booting the System" on page 31
6	Install the operating system on the system admin controller (SAC) node. You can install either the SUSE Linux Enterprise Server (SLES) or Red Hat Enterprise Server (RHEL) operating system.	"Installing the Operating System" on page 33
7	Run the cluster configuration tool. Complete the initial cluster configuration tasks, which include the following: • Set up software repositories for required and optional software. • Install the SAC software. • Configure network settings. • Configure the NTP server. • Set up the initial SAC infrastructure. • Configure the house network DNS resolvers.	"Running the Cluster Configuration Tool" on page 42
8	(Conditional) Configure external domain name service (DNS). If you want to configure network address translation, you also need to configure an external DNS.	"(Conditional) Configuring External Domain Name Service (DNS) Servers" on page 51
9	Install the SGI Management Center (SMC) license key.	"Installing the SGI Management Center License Key" on page 52

Step	Task	See
10	Sync the repository updates, apply the latest patches to the newly installed software, and clone the images.	"Synchronizing the Software Repository, Installing Software Updates, and Cloning the Images" on page 53
11	Configure the switches.	"Configuring the Switches" on page 55
12	Use the `discover` command to install and configure the rack leader controller and service node software.	"Configuring the Rack Leader Controllers (RLCs) and Service Nodes with the `discover` Command" on page 65
13	(Optional) Configure a backup domain name service (DNS) server on a service node.	"(Optional) Configuring a Backup Domain Name Service (DNS) Server" on page 68
14	Configure the InfiniBand subnetworks.	"Configuring the InfiniBand Subnetworks" on page 69
15	Configure optional features.	Chapter 4, "Configuring Optional Features" on page 75

Planning the Boot Slots and Disk Partitions

On a multiple-slot SGI ICE X system, the system admin controller (SAC), the rack leader controller (RLC), and the service nodes all have the same disk layout. If you insert an operating system installation disk, power-on the SAC, and type **`install`** at the `boot:` prompt, the default behavior is for the system to create two slots and to write the initial installation to slot 1. After the system is installed, you cannot increase the number of slots without destroying the data on the disks. You can configure up to five slots, for a total of five root/boot directory pairs.

One feature of a multi-boot system is that you can install a different operating systems, or different operating system versions, into different slots. You can boot the system with the operating system of your choice. This configuration might be useful if you want to test an operating system or other software. The following are some other characteristics of single-boot systems and multi-boot systems:

Multi-boot	Single-boot
RLCs and service nodes boot from their own disk. Data is retained in the master boot record (MBR).	RLCs and service nodes boot from the boot partition in the slot that is currently configured as the boot slot. Only the SAC retains data in the MBR.
RLC and service node software is reinstalled from the SAC.	Software on the RLCs and service nodes is reinstalled over the network.
As you increase the number of slots, you decrease the amount of disk space per slot. SGI recommends a minimum of 100 GB per slot.	A single slot uses all available disk space.

Procedure 3-1 To plan the boot slots and disk partitions

1. Select your boot parameters and plan the disk partitions.

   The following topics describe the boot parameters you can specify, disk partitioning, and operations such as cloning:

   - "Boot Parameters" on page 22
   - "Partition Layout for a Two-slot SGI ICE X System (Default)" on page 25
   - "Partition Layout for a Five—slot SGI ICE X System" on page 26
   - "Partition Layout for a One-slot (Single—boot) SGI ICE X System" on page 26

2. Proceed to the following:

   "Preparing to Install Software on an SGI ICE X System" on page 27

Boot Parameters

When you use the SGI system admin controller (SAC) installation DVD to boot an SGI ICE X system, you can specify parameters at the `boot:` prompt. By default, if you type **`install`** at the boot prompt and the installer detects a SAC with exactly one blank disk, the installer partitions the SAC with two slots, and the installer writes the initial installation to slot 1.

If you specify more than one boot parameter, use a space to separate each parameter.

The following list shows the result of pressing Enter at the boot prompt and shows the result of the different boot options:

Parameter **Effect**

Press `Enter`

If you press `Enter` at the `boot:` prompt, without specifying any boot parameters, the installer creates two slots, but it installs a root directory on only one slot. The software repository is not configured, so you when you run the `configure-cluster` command, you need to use the tool to manually enter the distribution ISO images in the `/tftpboot` directory on the SAC.

You can use the second slot for a clone or preproduction testing.

`console=`*specs*

Specify console characteristics. Use when you connected to the compute node through a serial console. By default, this is set to `ttyS1,38400n8`.

The BIOS console and the IPMI console settings are configured in the SAC BIOS. If you configure the SAC BIOS to be different from the default, use this boot option to specify your console's characteristics.

`destructive=1`

Permits partitioning operations that are potentially destructive. Use this parameter only if you want to repartition a disk that contains data. Use in conjunction with the `re_partition_with_slots` parameter.

If the system encounters data in a partition, nothing destructive happens unless you also specify the `destructive=1` parameter.

`install`

Creates two slots on each compute node and installs the operating system from the DVD to slot 1. Both slots contain a root directory, and the installer writes the software distribution ISO images to each slot.

This is the recommended boot option for an initial configuration.

`install_slot=`*slot*

Specifies the slot into which the operating system is installed first. Specify `1` (default), `2`, `3`, `4`, or `5` for *slot*. By default, slot 1 is installed first. To direct the initial installation to a slot other than slot 1, specify this parameter.

If you specify a *slot* that appears to have data on it, no repartitioning is performed unless you also specify `destructive=1`.

Example: You can specify `install_slot=2` to direct the first installation into slot 2. In this case, after the boot completes, type the following `cadmin` command to set the default slot to slot 2:

```
# cadmin --set-default-root --slot 2
```

In this example, if you do not specify the default slot (slot 2) during the first boot of the SAC, then the SAC attempts to boot from an empty slot 1. The boot fails. If this happens, restart the boot, and select the install slot during the boot.

`netinst=`*path*

Specifies the NFS path to an ISO image for a network installation.

`re_partition_with_slots=`*slots*

Partitions the SAC system drive with between one and five slots. Specify `1`, `2` (default), `3`, `4`, or `5` for *slots*.

The SAC system drive must be blank in order for this parameter to have an effect. If the SAC system drive contains data and you want to reconfigure the system, also specify the `destructive=1` parameter.

For example, the following parameter creates five slots:

```
re_partition_with_slots=5
```

The installer creates partitions on the rack leader controllers (RLCs) and service nodes to mimic the SAC. If an RLC or service node is discovered to have a slot count that does not match the SAC, the system reinitializes the partitions on the RLC and service nodes. Likewise, if you change the number of slots on the SAC, the system updates the disk partitioning on the RLC and service nodes, too.

`serial`

Configures the system for serial console operations for the installation and later operations. If you do not specify `serial`, the system sends output to the VGA.

Also see `vga` in this list.

`vga`

Configures the system so that you can use the VGA screen for the installation and later operations.

If you press `Enter` at the boot prompt, the result is the same as if you had typed `vga`.

Also see `serial` in this list.

Partition Layout for a Two-slot SGI ICE X System (Default)

Table 3-2 on page 25 shows the layout for a default two-slot SGI ICE X system.

Table 3-2 Example Partition Layout for a Two–slot SGI ICE X System (Default Layout)

Partition	File System Type	File System Label	Notes
1	swap	`sgiswap`	Partition layout for multiple slots.
2	ext3	`sgidata`	SGI data partition.
3	-	N/A	Extended partition. Makes logicals out of the rest of the disk.
5	ext3	`sgiboot`	Slot 1 `/boot` partition.
6	ext3 or XFS	`sgiroot`	Slot 1 `/` partition.
7	ext3	`sgiboot`	Slot 2 `/boot` partition.
8	ext3 or XFS	`sgiroot`	Slot 2 `/` partition.

Partition Layout for a Five—slot SGI ICE X System

Table 3-3 on page 26 shows a sample partition layout for a five-slot SGI ICE X system. This layout yields five boot partitions.

Table 3-3 Example Partition Layout for a Five–slot SGI ICE X System

Partition	File System Type	File System Label	Notes
1	swap	`sgiswap`	Partition layout for multiple slots.
2	ext3	`sgidata`	SGI data partition.
3	-	N/A	Extended partition. Makes logicals out of the rest of the disk.
5	ext3	`sgiboot`	Slot 1 `/boot` partition.
6	ext3 or XFS	`sgiroot`	Slot 1 `/` partition.
7	ext3	`sgiboot`	Slot 2 `/boot` partition.
8	ext3 or XFS	`sgiroot`	Slot 2 `/` partition.
9	ext3	`sgiboot`	Slot 3 `/boot` partition.
10	ext3 or XFS	`sgiroot`	Slot 3 `/` partition.
11	ext3	`sgiboot`	Slot 4 `/boot` partition.
12	ext3 or XFS	`sgiroot`	Slot 4 `/` partition.
13	ext3	`sgiboot`	Slot 5 `/boot` partition.
14	ext3 or XFS	`sgiroot`	Slot 5 `/` partition.

Partition Layout for a One-slot (Single—boot) SGI ICE X System

Table 3-4 on page 27 shows a sample partition layout for a single-boot SGI ICE X system. This layout shows one slot, which yields one boot partition. If you configure a single-slot system and later decide to add another partition, the addition process destroys all the data on your system.

Table 3-4 Example Partition Layout for a Single-boot SGI ICE X System

Partition	File System Type	File System Label	Notes
1	ext3	`sgiboot`	`/boot` partition.
2	-	N/A	Extended partition. Makes logicals out of the rest of the disk.
5	swap	`sgiswap`	Swap partition.
6	ext3 or XFS	`sgiroot`	`/` partition.

If you upgrade an existing SGI ICE X system to a more current release of the SGI Management Center software, your SAC might have a partition layout that differs from the layouts shown in the preceding tables.

Preparing to Install Software on an SGI ICE X System

The following procedure explains the information you need to obtain before you begin working with the SGI ICE X system. Your installation session can proceed more quickly if you gather information before you begin.

Procedure 3-2 To prepare for an installation

1. Contact your site's network administrator, and obtain network information.

   Obtain the following information to use when you configure the baseboard management controller (BMC):

   - (Optional) The current IP address of the BMC on the system admin node (SAC). You can set the BMC address from a serial console if you do not have this information.
   - The address you want to set for the BMC.
   - The netmask you want to set for the BMC.
   - The default gateway you want to set for the BMC.

   Your network administrator can provide an IP address, a hostname, or a fully qualified domain name (FQDN) for each of the preceding addresses.

Obtain the following information to use when you configure the network for the SGI ICE X system:

- Hostname
- Domain name
- IP address
- Netmask
- Default route
- Root password

Obtain the following information about your site's house network:

- IP addresses of the domain name servers (DNSs)

2. Familiarize yourself with the boot parameters, and determine which boot parameters you want to use.

   You can configure your SGI ICE X system to boot from one, two (default), three, four, or five partitions. This enables you to configure your SGI ICE X system as either a single-boot computer system or as a multiple-boot computer system. A multiple-boot computer system has two or more partitions, so it has more than one root directory (`/`) and more than one boot directory (`/boot`). In an SGI ICE X system, these root and boot directories are paired into multiple *slots*. A multiple-slot disk layout is also called a *cascading dual-root layout* or a *cascading dual-boot layout*.

   The installation procedure explains how to create a default, two-slot SGI ICE X system and directs you to use the `install` boot parameter. If you want to create only one slot, or if you want to create three or more slots, you need to specify different boot parameters.

   The installer creates the same disk layout on all nodes. For more information about boot parameters, disk layouts, and so on, see "Planning the Boot Slots and Disk Partitions" on page 21.

3. (Optional) Obtain the MAC file for your system from your SGI representative.

   The MAC file contains MAC address information for the nodes. If you have these addresses, the node discovery process can complete more quickly.

(Conditional) Setting a Static IP Address for the Baseboard Management Controller (BMC) in the System Admin Node (SAC)

Perform the procedure in this topic only if your site practices require a static IP address for the BMC.

When you set the IP address for the BMC on the SAC, you ensure access to the SAC when the site DHCP server is inaccessible. This procedure is required if you want to enable high availability. If you want to configure a highly available SAC, make sure to perform this topic's procedure on the BMCs on each of the two SACs.

The following procedures explain how to set a static IP address.

Procedure 3-3 Method 1 — To change from the BIOS

1. Use the BIOS documentation for the SAC.

Procedure 3-4 Method 2 — To change the IP address from the SAC

1. Log into the SAC as the root user.
2. Type the following command to retrieve the current network settings:

   ```
   ipmitool lan print 1
   ```

3. In the output from the preceding command, look for the `IP Address Source` line and the `IP Address` line.

   For example:

   ```
   IP Address Source       : DHCP Address
   IP Address              : 128.162.244.59
   ```

   Note the IP address in this step and decide whether or not this IP address is acceptable. The rest of this procedure explains how to keep this IP address or to set a different static IP address.

4. Type the following command to specify that you want the BMC to have a static IP address:

   ```
   ipmitool lan set 1 ipsrc static
   ```

   This step specifies that the IP address on the BMC is a static IP address, and this step sets the IP address to the IP address that is currently assigned to the BMC. If you want to set the IP address to a different IP address, proceed to the following

step. If the current IP address is acceptable, you do not need to perform the next step.

5. (Optional) Set a different IP address.

   Perform this step if you want to set the static IP address to be a different IP address than the one that is set currently.

   Type the following commands:

   ```
   ipmitool lan set 1 ipaddr ip_addr
   ipmitool lan set 1 netmask netmask
   ipmitool lan set 1 defgw gateway
   ```

   For *ip_addr*, specify the IP address you want to assign to the BMC.

   For *netmask*, specify the netmask you want to assign to the BMC.

   For *gateway*, specify the gateway you want to assign to the BMC.

6. Proceed to one of the following:

   - If you want to configure a highly available SAC, proceed to the following:

     "(Optional) Configuring a Highly Available (HA) System Admin Controller (SAC)" on page 31

   - If you want to configure a traditional SAC, proceed to the following:

     "Booting the System" on page 31

Example. Assume that you want to set a static IP address on the SAC BMC, and you do not want to accept the currently assigned IP address as the permanent, static, IP address. Type the following command to set a static IP address on `sleet-bmc`:

```
# ipmitool lan set 1 ipsrc static
```

Type the following commands to set the IP address on `sleet-bmc` to 100.100.100.100:

```
# ipmitool lan set 1 ipaddr 100.100.100.100
# ipmitool lan set 1 netmask 255.255.255.0
# ipmitool lan set 1 defgw 128.162.244.1
```

(Optional) Configuring a Highly Available (HA) System Admin Controller (SAC)

SGI enables you to configure the SAC and your rack leader controllers (RLCs) as highly available nodes in an SGI ICE X system. If you want to enable high availability, contact your SGI representative.

Booting the System

The following procedure explains how to boot the system and begin the installation.

Procedure 3-5 To boot the system

1. (Conditional) Power-off the system admin controller (SAC).

   Perform this step only if the SAC is powered on at this time.

2. Power-on the SAC.

   As Figure 3-2 on page 31 shows, the power-on button is on the right of the SAC.

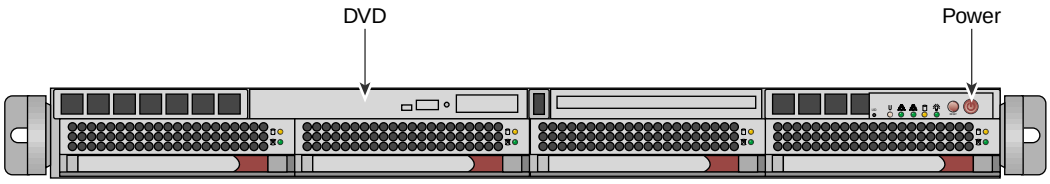

Figure 3-2 SAC Power On Button and DVD Drive

3. (Optional) Back up the cluster configuration snapshot in the `/opt/sgi/var/ivt` directory on the SAC.

   If you back up your system's factory configuration now, you can use the interconnect verification tool (IVT) to verify your hardware configuration later.

   You can use `ftp` or `scp` to copy the IVT files to another server at your site, or you can write the IVT files to a USB stick. For more information about the IVT, see the *SGI ICE X Administrator Guide*.

4. (Optional) Configure the system so that you can perform the installation from a VGA screen and can perform later operations from a serial console.

   If you want to enable this capability, perform the following steps:

   - Use a text editor to open file `/boot/grub/menu.lst`.
   - Search the file for the word `kernel` at the beginning of a line.
   - Add the following to the `kernel` line: `console=`*type*.

     For example:

```
kernel /boot/vmlinuz-2.6.16.46-0.12-smp root=/dev/disk/by-label/sgiroot console=ttyS1,38400n8
splash=silent showopts
```

   - Add the `console=`*type* parameter to the end of every `kernel` line.

   Later, if you want to access the SAC from only a VGA, you can remove the `console=` parameters.

5. Insert the SGI Admin Node Autoinstallation DVD into the DVD drive on the system admin controller (SAC).

   The autoinstallation message appears, and at the end is the `boot:` prompt.

6. At the `boot:` prompt, type **`install`** and, optionally, other boot parameters.

   "Planning the Boot Slots and Disk Partitions" on page 21 explains the other, optional, boot parameters.

   Monitor the installation. This can take several minutes.

7. Remove the operating system installation DVD.

8. At the # prompt, type **`reboot`**.

   This is the first boot from the SAC's hard disk. The SAC displays log messages during the boot. If you want to suppress the log message output to the screen, edit file `/etc/syscontrol.conf` and add the following line to the top of the file (line 1):

   ```
   kernel.printk = 2 4 1 7
   ```

   In the preceding `kernel.printk` line, the spaces between the numbers `2 4 1 7` are Tab characters.

9. Proceed to one of the following:

"Installing the Operating System" on page 33

Installing the Operating System

The SGI ICE X platform supports both the SUSE Linux Enterprise Server (SLES) and Red Hat Enterprise Linux (RHEL) operating systems. Use one of the following procedures to install your operating system software on the system admin controller (SAC) node:

- "Installing SUSE Linux Enterprise Server (SLES)" on page 33
- "Installing Red Hat Enterprise Linux (RHEL)" on page 37

Installing SUSE Linux Enterprise Server (SLES)

The SLES YaST2 interface enables you to install the SLES operating system on the SGI ICE X system. To navigate the YaST2 modules, use key combinations such as `Tab` (forward) `Shift + Tab` (backward). You can use the arrow keys to move up, down, left, and right. To use shortcuts, press the `Alt` + the highlighted letter. Press `Enter` to complete or confirm an action. `Ctrl + L` refreshes the screen. For more information about navigation, see Appendix A, "YaST2 Navigation" on page 135.

The following procedure explains how to use YaST2 to install SLES 11 SP2 on an SGI ICE X system. Use the following keys to navigate the YaST2 interface:

Procedure 3-6 To install SLES 11 SP2 on an SGI ICE X SAC

1. Connect to the system admin controller (SAC) by one of the following methods:
   - Through the intelligent platform management interface (IMPI) tool
   - Through the console attached to the SGI ICE X system
   - Through a separate keyboard, video display terminal, and mouse
2. On the **Language and Keyboard Layout** screen, complete the following steps:
   - Select your language
   - Select your keyboard layout
   - Select **Next**.
3. On the **Welcome** screen, select **Next**.

4. On the **Hostname and Domain Name** screen, complete the following steps:
   - Type the hostname for this SGI ICE X system.
   - Type the domain name.
   - Clear the box next to **Change Hostname via DHCP**. The box appears with an X in it by default, but you need to clear this box.
   - Select **Assign Hostname to Loopback IP**. Put an X in this box.
   - Select **Next**.
5. On the **Network Configuration** screen, complete the following steps:
   - Select **Change**. A pop-up window appears.
   - On the pop-up window, choose **Network Interfaces.**
6. On the **Network Settings** screen, complete the following steps:
   - Highlight the first network interface card that appears underneath **Name**.
   - Select **Edit**.
7. On the **Network Card Setup** screen, specify the system admin controller's (SAC's) house/public network interface.

   Figure 3-3 on page 35 shows the **Network Card Setup** screen.

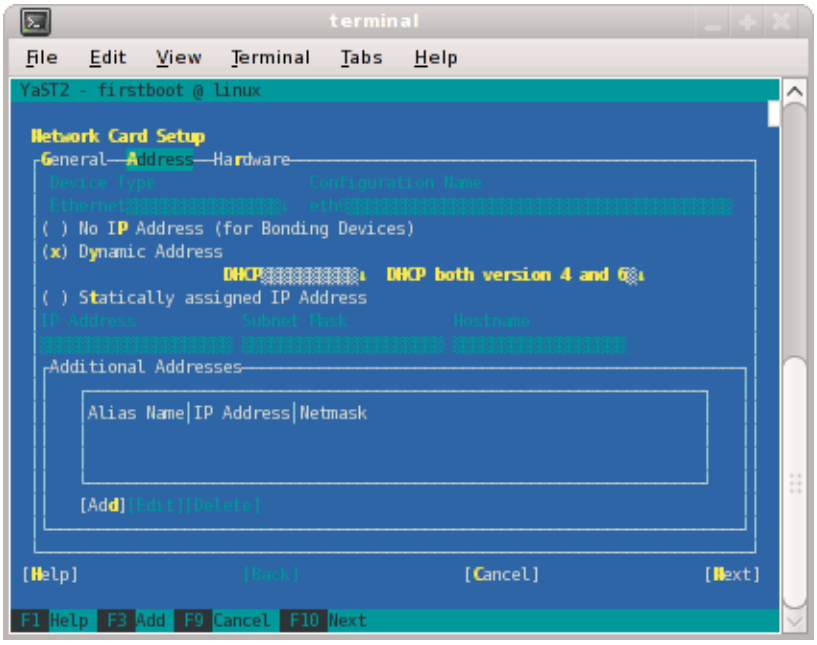

Figure 3-3 Network Card Setup Screen

Complete the following steps:

- Select **Statically Assigned IP Address**. SGI recommends a static IP address, not DHCP, for system admin nodes (SACs).
- In the **IP Address** field, type the system's IP address.
- In the **Subnet Mask** field, type the system's subnet mask.
- In the **Hostname** field, type the system's fully qualified domain name (FQDN). SGI requires you to type an FQDN, not the system's shorter hostname, into this field. For example, type `mysystem-admin.mydomainname.com`. Failure to supply an FQDN in this field causes the `configure-cluster` command to fail.
- Select **Next**.

You can specify the default route, if needed, in a later step.

8. On the **Network Settings** screen, complete the following steps:
   - Select **Hostname/DNS**.
   - In the **Hostname** field, type the system's fully qualified domain name (FQDN).
   - In the **Domain Name** field, type the domain name for your site.
   - Put an X in the box next to **Assign Hostname to Loopback IP**.
   - In the **Name Servers and Domain Search List**, type the name servers for your house network.
   - Back at the top of the screen, select **Routing**.

     The **Network Settings > Routing** screen appears.
   - In the **Default Gateway** field, type your site's default gateway.
   - Select **OK**.
9. On the **Network Configuration** screen, click **Next**.

   The **Saving Network Configuration** screen appears and saves your configuration.
10. On the **Clock and Time Zone** screen, complete the following steps:
    - Select your region.
    - Select your time zone.
    - (Optional) In the **Hardware Clock Set To** field, choose **Local Time** or accept the default of **UTC**.
    - Select **Next**.

    This step synchronizes the time in the BIOS hardware with the time in the operating system. Your choice depends on how the BIOS hardware clock is set. If the clock is set to GMT, which corresponds to UTC, your system can rely on the operating system to switch from standard time to daylight savings time and back automatically.
11. On the **Password for System Administrator "root"** screen, complete the following steps:
    - In the **Password for root User** field, type the password you want to use for the root user.

This password becomes the root user's password for all the nodes on the ICE X system. These nodes are as follows:

- SAC
- Rack leader controller (RLC)
- Service nodes
- Compute nodes (blades)

- In the **Confirm password** field, type the root user's password again.
- In the **Test Keyboard Layout** field, type a few characters.

  For example, if you specified a language other than English, type a few characters that are unique to that language. If these characters appear in this plain text field, you can use these characters in passwords safely.

- Select **Next**.

12. On the **User Authentication Method** screen, select one of the authentication methods and select **Next**.

    Typically, users accept the default (**Local**).

13. On the **New Local User** screen, create additional user accounts or select **Next**.

    If you do not create additional users, select **Yes** on the **Empty User Login** warning pop-up window, and select **Next**.

14. On the **Installation Completed** screen, select **Finish**.

15. Type a tilde character (~) and then a period character (.) to exit from the IPMI tool.

16. Log into the SAC and confirm that the system is working as expected.

    If necessary, restart YaST2 to correct settings.

Installing Red Hat Enterprise Linux (RHEL)

This section describes how to configure Red Hat Enterprise Linux 6 on the system admin controller (SAC).

Procedure 3-7 To install RHEL 6 on an SGI ICE X SAC

1. Connect to the system admin controller (SAC) by one of the following methods:
   - Through the intelligent platform management interface (IMPI) tool
   - Through the console attached to the SGI ICE X system
   - Through a separate keyboard, video display terminal, and mouse
2. Use a text editor, such as `vi` or `vim`, to open file `/etc/sysconfig/network-scripts/ifcfg-eth0`.
3. Add lines for the `IPADDR`, `NETMASK`, and `NETWORK` values appropriate for your public (house) network to file `/etc/sysconfig/network-scripts/ifcfg-eth0`.

   For example:

   ```
   IPADDR=128.162.244.88
   NETMASK=255.255.255.0
   NETWORK=128.162.244.0
   ```

4. Save and close file `/etc/sysconfig/network-scripts/ifcfg-eth0`.
5. Use a text editor to create file `/etc/sysconfig/network`.
6. Add the following three lines to file `/etc/sysconfig/network`:

   ```
   NETWORKING=yes
   HOSTNAME=SAC_hostname
   GATEWAY=gateway_IP_address
   ```

   For *SAC_hostname*, type the hostname you want to assign to the SAC.

   For *gateway_IP_address*, type the IP address of the gateway for your house network.

   For example:

   ```
   NETWORKING=yes
   HOSTNAME=my-system-admin
   GATEWAY=128.162.244.1
   ```

7. Save and close file `/etc/sysconfig/network`.
8. Use a text editor to open file `/etc/hosts`.

9. Add a line in the following format to file `/etc/hosts`:

   SAC_IP SAC_FQDN SAC_hostname

   The variables in the preceding line are as follows:

   - For *SAC_IP*, type the IP address of the SAC.
   - For *SAC_FQDN*, type the fully qualified domain name (FQDN) of the SAC.
   - For *SAC_hostname*, type the hostname of the SAC.

   For example, add the following line:

```
128.162.244.88 my-system-admin.domain-name.mycompany.com my-system-admin
```

10. Save and close file `/etc/hosts`.
11. Type the following command to set the SAC hostname:

    # **hostname *SAC_hostname***

    For *SAC_hostname*, type the hostname of the SAC.

    For example:

    # **hostname my-system-admin**

12. Use a text editor to create file `/etc/resolv.conf`.
13. Add lines to file `/etc/resolv.conf` that specify the search domain and the domain name service (DNS) servers at your site.

    Later in the configuration process, when you run the cluster configuration tool, the tool uses the DNS servers you specify in this step for its defaults.

    Specify lines with the following format:

    `search` *search_domain*
    `nameserver` *name_server_IP*
    `nameserver` *name_server_IP*

    The following is an example `resolv.conf` file:

```
search mydomain.com
nameserver 192.168.0.1
nameserver 192.168.0.25
```

14. Type the following ncsd(8) command to force the invalidation of the name service cache daemon:

    # **nscd -i hosts**

15. Type the following commands, in the order shown, to restart services:

    # **/etc/init.d/network restart**
    # **/etc/init.d/rpcbind start**
    # **/etc/init.d/nfslock start**

16. Type the following command to retrieve the SAC's current time zone information:

    # **strings /etc/localtime | tail -1**
    CST6CDT,M3.2.0,M11.1.0

    The previous output shows the SAC set to US Central time. If the output you see is not correct for this SGI ICE X system, perform the following steps:

    a. Type the following command to change to the directory that contains the time zone configuration files:

       # **cd /usr/share/zoneinfo**

    b. Select a file from that directory that describes the time zone for the SAC.

    c. Type commands to enable the new time zone configuration file.

       For example:

       # **/bin/cp -l /usr/share/zoneinfo/*time_zone_file* /etc/localtime.$$**
       # **/bin/mv /etc/localtime.$$ /etc/localtime**

       For *time_zone_file*, type the name of the time zone file that you need from the */usr/share/zoneinfo* directory.

       For example, type the following commands to change the SAC's time zone to US Pacific time:

       # **/bin/cp -l /usr/share/zoneinfo/PST8PDT /etc/localtime.$$**
       # **/bin/mv /etc/localtime.$$ /etc/localtime**

    d. Type the following command to confirm the time zone:

       # **strings /etc/localtime | tail -1**
       PST8PDT,M3.2.0,M11.1.0

17. (Conditional) Edit file `/etc/ntp.conf` to direct requests to the network time protocol (NTP) server at your site.

    Complete the following steps if you want to direct requests to your site's NTP server instead of to the public time servers of the `pool.ntp.org` project:

    a. Use a text editor to open file `/etc/ntp.conf`.

    b. Insert a pound character (#) into column 1 of each of each line that includes `rhel.pool.ntp.org`.

    Note: Do not edit or remove entries that serve the cluster networks.

    c. Add a line at the end of the file that points to your site's NTP server.

    The following is an example of a correctly edited file:

    ```
    # Use public servers from the pool.ntp.org project.
    # Please consider joining the pool (http://www.pool.ntp.org
    # server 0.rhel.pool.ntp.org
    # server 1.rhel.pool.ntp.org
    # server 2.rhel.pool.ntp.org
    server ntp.mycompany.com
    ```

    The preceding output has been truncated at the right for inclusion in this guide.

    d. Type the following command to restart the NTP server:

    ```
    # /etc/init.d/ntpd restart
    ```

18. Type the following command to register the SAC with the Red Hat Network (RHN):

    ```
    # /usr/bin/rhn_register
    ```

19. (Conditional) Type a tilde character (~) and then a period character (.) to exit from the IPMI tool.

    Perform this step if you connected to the system through the IPMI tool.

20. Proceed to the following:

    "Running the Cluster Configuration Tool" on page 42

Running the Cluster Configuration Tool

The cluster configuration tool enables you to configure, or reconfigure, your SGI ICE X system. The procedure in this topic explains the general, required configuration steps for SGI ICE X systems. If your SGI ICE X system includes optional components, or if your site has specific requirements, later procedures explain how to use the cluster configuration tool to create a more customized environment.

The following procedure explains how to complete the following required cluster configuration steps:

- Create repositories for software installation files and updates.
- Install the system admin node (SAC) cluster software.
- Configure the cluster subdomain and examine other network settings. The cluster subdomain is likely to be different from the `eth0` domain on the SAC itself.
- Configure the NTP server.
- Install the cluster's software infrastructure. This step can take 30 minutes.
- Configure the house network's DNS resolvers.

Procedure 3-8 To run the cluster configuration tool

1. Locate your site's SGI software distribution DVDs or verify the path to your site's online software repository.

   You can install the software from either physical media or from an ISO on your network.

2. From the VGA screen, or through an `ssh` connection, log into the system admin controller (SAC) as the root user.

   SGI recommends that you run the cluster configuration tool either from the VGA screen or from an `ssh` session to the system admin controller (SAC). Avoid running the `configure-cluster` command from a serial console.

3. Type the following command to start the cluster configuration tool:

   # **`/opt/sgi/sbin/configure-cluster`**

4. On the cluster configuration tool's **Initial Configuration Check** screen, select **OK** on the initial window.

   Figure 3-4 on page 43 shows the initial window.

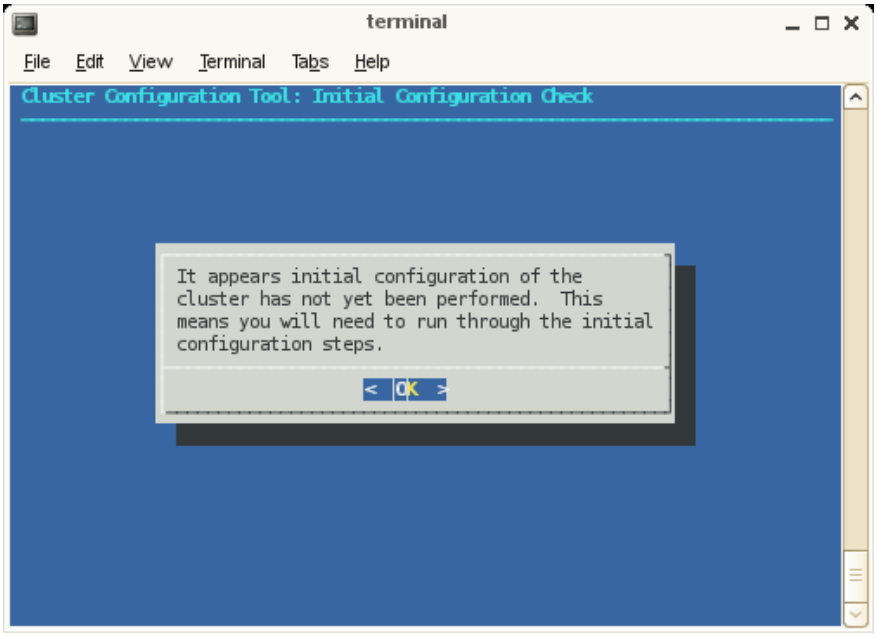

Figure 3-4 Initial Configuration Check Screen

The cluster configuration tool recognizes a configured cluster. If you start the tool on a configured SGI ICE X system, it opens into the **Main Menu**.

5. On the **Initial Cluster Setup** screen, select **OK** on the screen.

   Figure 3-5 on page 44 shows the window.

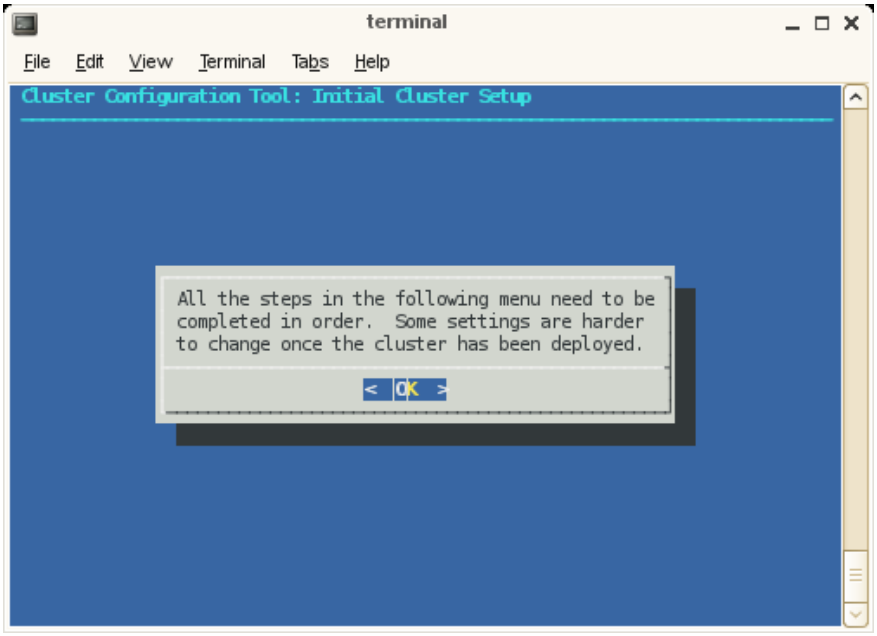

Figure 3-5 Initial Cluster Setup Screen with the initial screen

6. On the **Initial Cluster Setup** screen, select **R Repo Manager: Set Up Software Repos**, and click **OK**.

   Figure 3-6 on page 45 shows the **Initial Cluster Setup** screen with the task menu. This procedure guides you through the tasks you need to perform for each of the menu selections on the **Initial Cluster Setup** screen.

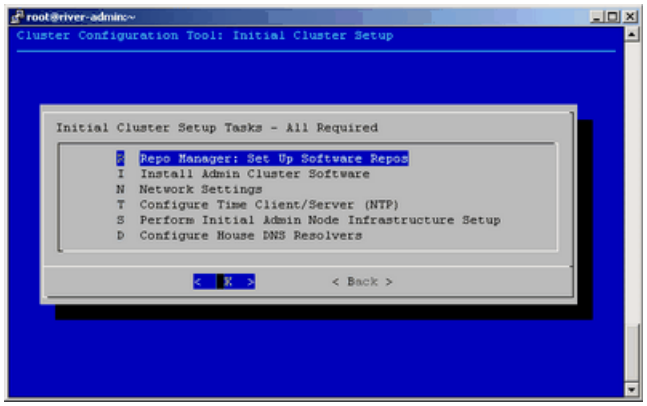

Figure 3-6 Initial Cluster Setup Tasks Screen

The next few steps create software repositories for the initial installation packages and for updates. You need to create repositories for the following software:

- The operating system software, either RHEL or SLES
- SGI Foundation Suite
- SGI Management Center (SMC) for SGI ICE X
- Additional software packages for which you hold licenses, such as the Message Passing Toolkit, the SGI Performance Suite, and any others

7. On the **One or more ISOs were embedded on the ...** screen, select **Yes**.
8. On the **Repositories are created ...** screen, press **Enter**.
9. On the **You will now be prompted to add additional media ...** screen, select **OK**.
10. On the **Would you like to register media with Tempo? ...** screen, select **Yes**.
11. On the **Please either insert the media in your DVD drive ...** screen, select either **Insert DVD** or **Use Custom path/url**.

    Proceed as follows:

    - To install the software from DVDs, perform the following steps:

      a. Insert a DVD.

b. Select **Mount inserted DVD**.

c. On the **Media registered successfully with crepo ...** screen, select **OK**, and eject the DVD.

d. On the **Would you like to register media with Tempo? ...** screen, select **Yes** if you have more software that you need to register.

If you select **Yes**, repeat the preceding this sequence for the next DVD.

If you select **No**, proceed to the next step.

- To install the software from a network location, perform the following steps:

a. Select **Use custom path/URL**.

b. On the **Please enter the full path to the mount point or the ISO file ...** screen, type the full path in *server_name*:*path_name*/*iso_file* format. This field also accepts a URL or an NFS path. Select **OK** after typing the path.

c. On the **Media registered successfully with crepo ...** screen, select **OK**.

d. On the **Would you like to register media with Tempo? ...** screen, select **Yes** if you have more software that you need to register.

If you select **Yes**, repeat the preceding tasks in this sequence for the next DVD.

If you select **No**, proceed to the next step.

12. On the **Initial Cluster Setup Tasks** screen, select **I Install Admin Cluster Software**, and select **OK**.

    This step installs the cluster software that you wrote to the repositories.

13. On the **Initial Cluster Setup Tasks** screen, select **N Network Settings**, and select **OK**.

14. On the **Cluster Network Settings** screen, select **S Configure Subnet Addresses**, and select **OK**.

15. On the **Warning: Changing the subnet IP addresses ...** screen, click **OK**.

16. Review the settings on the **Subnet Network Addresses** screen, and modify these settings only if absolutely necessary.

    Select either **OK** or **Back** if you accept the defaults.

    If your site has network requirements that conflict with the defaults, you need to change the network settings. On the **Update Subnet Addresses** screen, the **Head Network** field shows the SAC's IP address. SGI recommends that you do not change the IP address of the SAC or rack leader controllers (RLCs) if at all possible. You can change the IP addresses of the InfiniBand network (**IB0** and **IB1**) to match the IP requirements of the house network, and then select **OK**.

17. On the **Cluster Network Settings** screen, select **D Configure Cluster Domain Name**, and select **OK**.

18. On the **Please enter the domain name for this cluster.** pop-up window, type the domain name, and select **OK**.

    The domain you type becomes a subdomain to your house network..

    For example, type `ice.americas.sgi.com`.

19. On the **Cluster Network Settings** screen, select **Back**.

20. On the **Initial Cluster Setup** screen, select **T Configure Time Client/Server (NTP)**, and select **OK**.

21. Configure your NTP server.

    On the subsequent screens, you set the SAC as the time server to the SGI ICE X system. For this step, the installer screens differ on RHEL platforms and SUSE platforms.

    On RHEL platforms, complete the following step:

    - On the **A new ntp.conf has been put in to position ...** screen, select **OK**.

    On SLES platforms, complete the following steps:

    - On the **A new ntp.conf has been put in to position ...** screen, select **OK**.
    - Use the YaST interface and the SLES documentation to guide you through the NTP configuration.
    - On the **This procedure will replace your ntp configuration file ...** screen, select **Yes**.

22. On the **Initial Cluster Setup Tasks** menu, select **S Perform Initial Admin Node Infrastructure Setup**, and select **OK**.

23. On the **A script will now perform the initial cluster ...** screen, select **OK**.

    This step runs a series of scripts that configure the SAC on the SGI ICE X system. The scripts also create the root images for the RLCs, service nodes, and compute nodes. The scripts run for approximately 30 minutes. At the end, the script issues a line that includes **install-cluster completed** in its output.

    The final output of the script is as follows:

    ```
    /opt/sgi/sbin/create-default-sgi-images Done!
    ```

    The output of the `mksiimage` commands are stored in a log file at the following location:

    ```
    /var/log/cinstallman
    ```

24. On the **Initial Cluster Setup Complete** window, select **OK**.

25. On the **One or more ISOs were embedded on the admin install DVD and copied to ...**, screen, select **OK**.

26. On the **Initial Cluster Setup** menu, select **D Configure House DNS Resolvers**, and select **OK**.

    Figure 3-7 on page 49 shows the **Configure House DNS Resolvers** screen.

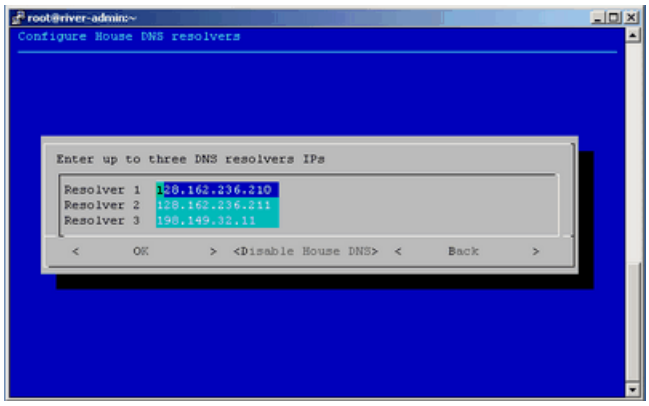

Figure 3-7 Configure House DNS Resolvers Screen

The system autopopulates the values on the **Configure House DNS Resolvers** screen to match the DNS specifications on the SAC. The DNS resolvers you specify here enable the service nodes to resolve host names on your network. You can set the DNS resolvers to the same name servers used on the SAC itself.

Perform one of the following actions:

- To accept these settings, select **OK**, and then select **Yes**.
- To change the settings, type in different IP addresses, and select **OK**, and then select **Yes**.
- To disable house network resolvers, select **Disable House DNS**.

  On the **Setting DNS Forwarders to ...** screen, select **Yes**.

27. On the **Initial Cluster Setup** screen, select **Back**.

    This action returns you to the cluster configuration tool main menu.

28. On the **Main Menu**, select **S Configure Switch Management Network (optional)**, and select **OK**.

    On an SGI ICE X system, the switch management network enables the Ethernet switch to control all VLANs and trunking.

29. On the pop-up window that appears, select **Y yes**, and select **OK**.

Figure 3-8 on page 50 shows the selection pop-up window:

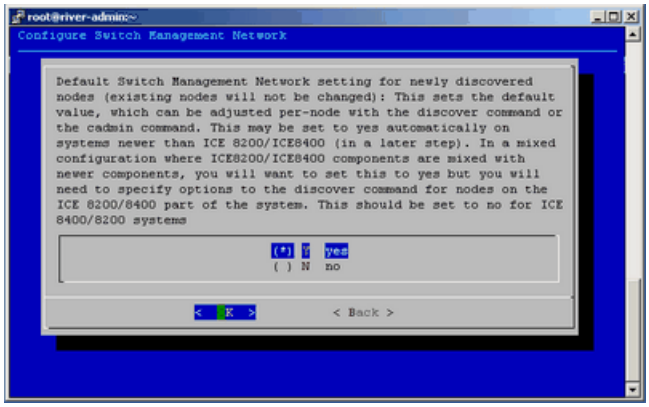

Figure 3-8 Configure Switch Management Network screen

30. (Conditional) On the **Main Menu**, select **N Configure MCell Network (optional)**, and select **OK**.

    Perform this step if your SGI ICE X system contains MCells.

31. (Conditional) On the screen that appears, select **Y yes**, and select **OK**.

    Perform this step if your SGI ICE X system contains MCells.

32. Select **Quit**.

33. Type the `cattr list -g` command to verify the features you configured with the cluster configuration tool.

    The following example output is appropriate for an SGI ICE X system with MCells. If your system does not include MCells, the `mcell_network` value should display `no`. The output is as follows:

```
# cattr list -g
global
   redundant_mgmt_network  : yes
   switch_mgmt_network     : yes
   mcell_network           : yes
   discover_skip_switchconfig : no
```

```
max_rack_irus           : 8
blademond_scan_interval : 120
my_sql_replication      : yes
rack_vlan_start         : 101
rack_vlan_end           : 1100
head_vlan               : 1
mcell_vlan              : 3
replication_file        : mysql-bin.000132
replication_position    : 6580967
```

If you need to respecify any global values, start the cluster configuration tool again, and correct your specifications. To start the cluster configuration tool, type the following command:

/opt/sgi/sbin/configure-cluster

34. Proceed to one of the following:

    - "(Conditional) Configuring External Domain Name Service (DNS) Servers" on page 51
    - "Installing the SGI Management Center License Key" on page 52

(Conditional) Configuring External Domain Name Service (DNS) Servers

Perform the procedure in this section if you want to enable network address translation (NAT) gateways for the SGI ICE X system. A later procedure explains how to configure NAT on one of your service nodes. If you want to enable NAT, perform the procedure in this topic at this time.

When external DNS and NAT are enabled, the host names for the compute nodes (blades) in the cluster resolve through external DNS servers. The compute nodes need to be able to reach your house network.

Note: You cannot configure this feature after you run the `discover` command. If you configure this feature after you run the `discover` command, the IP addresses assigned previously on the discovered nodes remain.

The following procedure explains how to configure external DNS servers.

Procedure 3-9 To configure external DNS servers

1. Obtain a large block of IP addresses from your network administrator.

   This feature requires you to reserve a block of IP addresses on your house network. If you want to use external DNS servers, all nodes on the InfiniBand networks, both the `ib0` and `ib1` networks are included. The external DNS is enabled to provide addresses for all rack leader controllers (RLCs), all service nodes, and all compute nodes.

2. Through an `ssh` connection, log into the system admin controller (SAC) as the root user.

3. Type the following command to start the cluster configuration tool:

   # **`configure-cluster`**

4. Select **E Configure External DNS Masters (optional)**, and select **OK**.

5. On the **This option configures SGI Tempo to look up the IP addresses for the InfiniBand networks from external DNS servers ...** screen, select **Yes**.

6. On the **Enter up to five external DNS master IPs** screen, type the IP addresses of up to five external DNS servers on your house network, and select **OK**.

7. On the **Setting external DNS masters to *ip_addr***, select **Yes**.

8. Proceed to the following:

   "Installing the SGI Management Center License Key" on page 52

Installing the SGI Management Center License Key

The SGI Management Center (SMC) software runs on the system admin controller (SAC). SMC provides a graphical user interface for system configuration, operation, and monitoring.

For more information about using SMC, see *SGI Management Center (SMC) System Administrator Guide.*

For more information about licensing, see the licensing FAQ on the following website:

http://www.sgi.com/support/licensing/faq.html

The following procedure explains how to obtain and install the license key for SMC.

Procedure 3-10 To license the SMC software

1. Use a text editor to open file `/etc/lk/keys.dat`.
2. Copy and paste the license key exactly as it was given to you.

   There can be line breaks in the license key.
3. Save the file.
4. Type the following command to restart the SMC daemon:

   # **`service mgr restart`**
5. Type the following command to start SMC:

   # **`mgrclient`**
6. Proceed to the following:

   "Synchronizing the Software Repository, Installing Software Updates, and Cloning the Images" on page 53

Synchronizing the Software Repository, Installing Software Updates, and Cloning the Images

The following procedure explains how to update the software in the repositories that you created with the cluster configuration tool. The following procedure assumes that the SGI ICE X system has a connection to the internet. If you need to perform this procedure on a secure SGI ICE X system, you need to modify this procedure. For a secure system, obtain the software updates from SGI SupportFolio manually and use the `crepo` command to install the software manually.

Procedure 3-11 To update the software

1. Through an `ssh` connection, log into the system admin controller (SAC) as the root user.
2. Type the following command to retrieve information about the network interface card (NIC) bonding method on the SAC:

   # **`cadmin --show-mgmt-bonding --node admin`**

   If bonding has been set appropriately, the command returns `802.3ad`.

If the command does not return `802.3ad`, type the following commands to set the bonding appropriately and reboot the system:

```
# cadmin --set-mgmt-bonding --node admin 802.3ad
# reboot
```

3. Type the following command to retrieve the new images from SGI SupportFolio and the operating system vendor:

   ```
   # sync-repo-updates
   ```

   For RHEL-based systems, make sure the system is subscribed as `rhel-x86_64--server-6`.

   This step requires that the system be connected to the internet. Contact your SGI representative if this update method is not acceptable for your site.

4. Type the `cinstallman --show-images` command to retrieve the image names.

   For example:

   ```
   # cinstallman --show-images
   Image Name              BT Path
   compute-rhel6.3         1  /var/lib/systemimager/images/compute-rhel6.3
   service-rhel6.3         0  /var/lib/systemimager/images/service-rhel6.3
   lead-rhel6.3            0  /var/lib/systemimager/images/lead-rhel6.3
   ```

5. (Optional) Clone the images.

   Perform this step if you want to back up the current images before they are installed.

   Type the following command:

   `cinstallman --create-image --clone --source` *src_image_name* `--image` *image*

   For *src_image_name*, specify the name of the source image. For example: `lead-rhel6.3`.

   For *image*, specify a file name for the copied file (the clone). For example: `lead-rhel6.3.backup`

   For example:

```
# cinstallman --create-image --clone --source compute-rhel6.3 --image compute-rhel6.3.backup
# cinstallman --create-image --clone --source service-rhel6.3 --image service-rhel6.3.backup
```

```
# cinstallman --create-image --clone --source lead-rhel6.3 --image lead-rhel6.3.backup
```

6. Type a series of `cinstallman --update-image` commands to update the software images.

   For *image*, specify the software packages displayed in the previous step.

   For example, to install the packages shows in Procedure 3-11, step 4 on page 54, type the following commands:

   ```
   # cinstallman --update-image --image compute-rhel6.3
   # cinstallman --update-image --image service-rhel6.3
   # cinstallman --update-image --image lead-rhel6.3
   ```

7. Proceed to the following:

   - "Configuring the Switches" on page 55

Configuring the Switches

The `discover` command initializes and configures the system components for the SGI ICE system. You use the `discover` command to configure the SGI ICE system's management switches first, and then if you have MCells, you configure the MCell switches. After you configure the switches, you can configure the nodes.

Switch configuration can proceed much more quickly if you have a media access control (MAC) file. For new SGI ICE X systems, you can obtain a MAC file from your SGI representative.

The MAC file shows the MAC addresses of the components in your environment. Switch discovery and configuration can complete more quickly if you obtain this file. Without this file, you need to power cycle each switch manually.

The following is an example MAC file:

```
r1lead 00:30:48:9e:f2:59 00:30:48:F2:7E:A2
r2lead 00:25:90:01:4e:3c 00:25:90:01:6c:cc
service0 00:25:90:00:3b:8f 00:25:90:01:6e:9e
mgmtsw0 00:26:f3:c3:7a:40 00:26:f3:c3:7a:40
```

The content of a MAC file is as follows:

Column	Content
1	The component's ID.
2	For nodes, column 2 contains the MAC address of baseboard management controller (BMC) on the component. For switches, column 2 contains the MAC address of the first network interface card (NIC), `eth0`. Switches do not have a BMC.
3	The MAC address of the first network interface card (NIC), `eth0`. For switches, columns two and three are identical in the MAC file.

The following list explains the procedures you need to follow to configure the switches:

Procedure	Perform if ...
"Configuring Management Switches With a MAC File" on page 56	If you have a MAC file.
"Configuring Management Switches Without a MAC File" on page 59	If you do not have a MAC file.
"(Conditional) Configuring the MCell Network" on page 61	If you have MCells. This extra procedure configures the MCell switches separately from the rest of the ICE X switches.

Configuring Management Switches With a MAC File

The following procedure explains how to configure your switches when you have each switch's MAC information in a MAC file.

Procedure 3-12 To configure switches — with a MAC file

1. Complete the following steps to prepare for switch configuration:
   - Visually inspect your system. Note the types of switches you have and their identifiers. At a minimum, you have one management switch stack. You might also have InfiniBand switches and other management switch stacks attached to MCells. In this procedure, you configure only the management switches.

- Make sure that only the system admin controller (SAC) is powered on. All other nodes and switches should be connected to a power source, but they should not be powered on. That is, make sure that all chassis management controllers (CMCs), all rack leader controllers (RLCs), all service nodes, all switches, and so on, are not powered on.
- Unplug the cables to the slave switches. For each management switch stack, verify that only one cable goes from the first switch stack to the second switch stack.

  This cable should connect the master switch in the upper switch to the master switch in the stack immediately below. Each switch can have a cable plugged into its slave switch, but make sure that the cables that connect the slave switches to each other are unplugged. This prevents looping.

2. Through an `ssh` connection, log in as root to the SAC, and write the MAC file to a location on your SAC.

   For example, write it to `/var/tmp/mac_file`.

3. Power-on all the management switches.
4. Type the following command to discover management switch 0, which is attached to the SAC:

   # **discover --mgmtswitch 0 --macfile *path***

   For *path*, type the full path to the location of the MAC file.

5. Type additional `discover` commands for each additional switch.

   The formats for these additional commands are as follows:

   # **discover --mgmtswitch *num* --macfile *path***

   For *num*, type the identifier for the switch.

   For *path*, type the full path to the location of the MAC file.

   For example:

   # **discover --mgmtswitch 1 --macfile /var/tmp/mac_file**

6. Plug in the cables to the slave switches.

7. Type the following command to retrieve information about the switches that you discovered, and examine the output for errors:

   # **cnodes --mgmtswitch**

   If the output is very long, direct the output to a file that you can examine with a text editor. For example:

   # **cnodes --mgmtswitch > switch_file**

8. Power on the CMCs on the system.

   The `cmcdetectd` daemon runs on the SAC. It configures the top level switches so that the CMCs are on the appropriate rack VLAN.

9. Save the switch configuration.

   This step explains how to save the switch configuration to a file on the system admin controller (SAC). In the future, if you need to replace the switch, you can save configuration time if you push this configuration file out to the new switch.

   Type the following command to save the switch configuration to a file on the SAC:

   `switchconfig pull_switch_config -s` *switch_ID* `-f` *file*

   For *switch_ID*, specify the switch name.

   For *file*, specify the name of the file to receive the switch configuration information. The command writes the file to the `/tftpboot/`*file*`.cfg`. If your *file* specification ends in `.cfg`, the command does not append another `.cfg` string to the file name.

   For example, the following command writes the configuration file for `mgmtsw0` to file `/tftpboot/mgmtsw0_startup1.cfg` on the SAC:

   ```
   switchconfig pull_switch_config -s mgmtsw0 -f mgmtsw0_startup1
   ```

10. After all management switches have been discovered, proceed to one of the following:

    - If you have MCells, proceed to the following:

      "(Conditional) Configuring the MCell Network" on page 61

    - If you do not have MCells, proceed to the following:

"Configuring the Rack Leader Controllers (RLCs) and Service Nodes with the `discover` Command" on page 65

Configuring Management Switches Without a MAC File

The following procedure explains how to configure your switches when you do not have the switch information in a MAC file.

Procedure 3-13 To configure switches — without a MAC file

1. Complete the following steps to prepare for switch configuration:

   - Visually inspect your system. Note the types of switches you have and their identifiers. At a minimum, you have one management switch stack. You might also have InfiniBand switches and other management switch stacks attached to MCells. In this procedure, you configure only the management switches.
   - Make sure that only the system admin controller (SAC) is powered on. All other nodes should be connected to a power source, but they should not be powered on. That is, make sure that all chassis management controllers (CMCs), all rack leader controllers (RLCs), all service nodes, and so on, are not powered on.
   - Unplug the cables to the slave switches. For each management switch stack, verify that only one cable goes from the first switch stack to the second switch stack.

     This cable should connect the master switch in the upper switch to the master switch in the stack immediately below. Each switch can have a cable plugged into its slave switch, but make sure that the cables that connect the slave switches to each other are unplugged. This prevents looping.

2. Through an `ssh` connection, log in to the SAC as the root user, and type the following command:

   ```
   # discover --mgmtswitch 0
   ```

3. When prompted, connect the switch to a power source.

   The command discovers the MAC address of the switch after you connect the switch to a power source.

4. Type additional `discover` commands for each additional switch.

The formats for these additional commands are as follows:

discover --mgmtswitch *num*

For *num*, type the identifier for the switch.

For example:

discover --mgmtswitch 1

Plug in the additional switches as directed by the system prompts.

5. Type the following command to save the MAC address to your MAC file:

   # **discover --show-macfile > *path***

   For *path*, type the full path to the location of the MAC file. For example, `/var/tmp/mac_file`.

6. Repeat the preceding steps for each switch that is attached to your SGI ICE system.

7. Plug in the cables to the slave switches.

8. Type the following command to retrieve information about the switches that you discovered, and examine the output for errors:

   # **cnodes --mgmtswitch**

   If the output is very long, direct the output to a file that you can examine with a text editor. For example:

   # **cnodes --mgmtswitch > switch_file**

9. Power on the CMCs on the system.

   The `cmcdetectd` daemon runs on the SAC. It configures the top level switches so that the CMCs are on the appropriate rack VLAN.

10. Save the switch configuration.

    This step explains how to save the switch configuration to a file on the system admin controller (SAC). In the future, if you need to replace the switch, you can save configuration time if you push this configuration file out to the new switch.

Type the following command to save the switch configuration to a file on the SAC:

`switchconfig pull_switch_config -s` *switch_ID* `-f` *file*

For *switch_ID*, specify the switch name.

For *file*, specify the name of the file to receive the switch configuration information. The command writes the file to the `/tftpboot/`*file*`.cfg`. If your *file* specification ends in `.cfg`, the command does not append another `.cfg` string to the file name.

For example, the following command writes the configuration file for `mgmtsw0` to file `/tftpboot/mgmtsw0_startup1.cfg` on the SAC:

```
switchconfig pull_switch_config -s mgmtsw0 -f mgmtsw0_startup1
```

11. After all management switches have been discovered, proceed to one of the following:

    - If you have MCells, proceed to the following:

      "(Conditional) Configuring the MCell Network" on page 61

    - If you do not have MCells, proceed to the following:

      "Configuring the Rack Leader Controllers (RLCs) and Service Nodes with the `discover` Command" on page 65

(Conditional) Configuring the MCell Network

Perform the procedure in this topic if you have an SGI ICE X system that includes MCells.

An SGI ICE X system contains cooling distribution units (CDUs) and cooling rack controllers (CRCs). The CDUs and CRCs have statically assigned IP addresses. These IP addresses are critical to associating the individual rack units (IRUs) with specific CDUs or CRCs. For information about these IP addresses, see the following:

Appendix C, "MCell Network IP Addresses" on page 149

The following procedure explains how to configure the switches attached to the MCells.

Procedure 3-14 To configure MCell switches

1. Gather information about the MCell switches in your SGI ICE X system.

Visually inspect your system. Note the switches identifiers, and note the port identifiers.

2. Log in as the root user to the system admin controller (SAC).
3. Use the `cattr list` command to retrieve information about the VLANs that are configured at this time.

   For example:

   ```
   # cattr list -g mcell_vlan
   global
     mcell_vlan              : 3
   ```

4. Use the `switchconfig set` command in the following format to configure the ports on which the CDUs and the CRCs are connected to the MCells.

   `switchconfig set -v num=`*vlan_num* `-b=none -d` *vlan_num* `-p` *ports* `-s` *switch*

   Type an individual `switchconfig set` command for each switch on the SGI ICE X system network.

   The arguments are as follows:

Argument	Specification
vlan_num	The VLAN number of the MCell network. For *vlan_num*, use the output from the `cattr list` command as shown earlier in this procedure. The default is 3, and SGI recommends that you do not change this value. This argument appears in two places in the `switchconfig` command.
ports	Specify the target ports. The command configures both the target ports and the corresponding redundant ports.

switch The ID number of the management switch to which the CDU or CDC is attached. For example: `mgmtsw0`.

To determine this value, you need to visually inspect the switch, as follows:

- Locate each CDU or CDC. The following are example labels for CDUs: `DU01`, `DU02`, and so on.
- Follow the cable that connects each CDU or CDC to a switch. The following is an example label for a cable that connects each CDU to a switch: `DU01-LAN1 | 101MSW0A-36`.
- Note the label on the switch. For example, the switch label could be one of the following: `MSW0A`, `MSW0B`, `MSW1A`, `MSW1B`, and so on. The `A` and `B` on the switch labels identify the master switch and slave switch in the stack. The `switchconfig set` command operates on the entire switch stack, so you need to note only the characters on the label that precede the `A` and `B`. Use the following table to determine the value you need to use for *switch*:

switch	**Label**
`mgmtsw0`	`MSW0A` or `MSW0B`
`mgmtsw1`	`MSW1A` or `MSW1B`
`mgmtsw2`	`MSW2A` or `MSW2B`
`mgmtsw3`	`MSW3A` or `MSW3B`
`mgmtsw4`	`MSW4A` or `MSW4B`
`mgmtsw5`	`MSW5A` or `MSW5B`
`mgmtsw6`	`MSW6A` or `MSW6B`
`mgmtsw7`	`MSW7A` or `MSW7B`
`mgmtsw8`	`MSW8A` or `MSW8B`

mgmtsw9	MSW9A or MSW9B

If you make a mistake in your configuration, you can disable the ports from the VLANs you configured. The following example command removes the configuration of VLAN 3 from the target ports and the redundant ports:

```
switchconfig unset -p 1/31,1/32,1/33 -s mgmtsw0 -v num=3
```

Example 1. The following command configures virtual local area network (VLAN) 3 on management switch 0 for target ports 1/31, 1/32, and 1/33 and for redundant ports 2/31, 2/32, and 2/33.

```
switchconfig set -v num=3 -b=none -d 3 -p 1/31,1/32,1/33 -s mgmtsw0
```

Example 2. The following command configures VLAN 3 on management switch 0 for target ports 2/31, 2/32, and 2/33 and for redundant ports 1/31, 1/32, and 1/33.

```
switchconfig set -v num=3 -b=none -d 3 -p 2/31,2/32,2/33 -s mgmtsw0
```

5. Repeat the preceding step for each CDU and each CRC attached to your system.

6. Save the switch configuration.

   This step explains how to save the switch configuration to a file on the system admin controller (SAC). In the future, if you need to replace the switch, you can save configuration time if you push this configuration file out to the new switch.

   Type the following command to save the switch configuration to a file on the SAC:

   `switchconfig pull_switch_config -s` *switch_ID* `-f` *file*

   For *switch_ID*, specify the switch name.

   For *file*, specify the name of the file to receive the switch configuration information. The command writes the file to the `/tftpboot/`*file*`.cfg`. If your *file* specification ends in `.cfg`, the command does not append another `.cfg` string to the file name.

   For example, the following command writes the configuration file for `mgmtsw0` to file `/tftpboot/mgmtsw0_startup1.cfg` on the SAC:

   ```
   switchconfig pull_switch_config -s mgmtsw0 -f mgmtsw0_startup1 --debug
   ```

The `--debug` parameter is optional. When specified, the command writes debugging information to `/var/log/switchconfig`.

7. After all switches have been discovered, proceed to the following:

   "Configuring the Rack Leader Controllers (RLCs) and Service Nodes with the `discover` Command" on page 65

Configuring the Rack Leader Controllers (RLCs) and Service Nodes with the `discover` Command

The `discover` command finds and configures rack leader controllers (RLCs), service nodes, and external switches. This procedure explains how to use the `discover` command to configure the RLCs and service nodes.

The following procedure explains how to use the `discover` command to identify and configure the RLCs and service nodes on the SGI ICE X system.

Procedure 3-15 To configure the RLCs and service nodes

1. Visually inspect your SGI ICE X system and note the labels on the RLCs and service nodes.

   RLCs are numbered starting with 1.

   Service nodes are numbered starting with 0.

2. Make sure all RLCs and service nodes are plugged in.

   Do not power-on the nodes at this time. When the node is plugged in and connected to a power source, the baseboard management controller (BMC) is started, and that is all that is required at this time.

3. Through an `ssh` connection, log into the system admin controller (SAC) as the root user.

4. Retrieve the DHCP option code in use, and reset the code if necessary.

   The RLCs and service nodes determine the integrated Ethernet devices by only accepting DHCP leases from the SGI Management Center (SMC) for SGI ICE X. SMC for SGI ICE X uses option code 149 by default. In rare situations, a house network DHCP server could be configured to use this option code. In this case, nodes that are connected to the house network can mistake a house DHCP server

as being an SMC for SGI ICE X DCHP server, which can lead to an installation failure. Change this option code only if absolutely necessary.

Type the following command to retrieve the option code that is in use:

```
# cadmin --show-dhcp-option
```

To change the `dhcp` option code number used for this operation, type a command such as the following:

```
# cadmin --set-dhcp-option 150
```

This command sets the DHCP option code to 150.

5. (Conditional) Power on all the nodes.

   Perform this step if you plan to specify a media access control (MAC) file as input to the `discover` command.

   A MAC file contains the MAC addresses for RLCs and service nodes. If you use a MAC file, the configuration process can complete more quickly. Contact your SGI representative to find out if a MAC file is available. For more information about MAC files, see "Configuring the Switches" on page 55.

6. Type one or more `discover` commands to configure the RLCs and service nodes on your SGI ICE X system.

   The following is the general format for the `discover` command:

   ```
   # discover --rack[set] rnum --service snum --macfile file
   ```

   The variables in the `discover` command are as follows:

   - For *rnum*, specify the ID number for the rack (or racks) that you want to configure.

     If you want to configure one rack, specify `--rack` and the ID number that corresponds to that rack. For example, `--rack 2`.

     If you want to configure a range of racks, specify `--rackset`, the starting rack ID number, a comma (`,`), and the ending rack ID number. For example, `--rackset 1,3`.

   - For *snum*, specify the ID number for the service node that you want to configure.

     For example, specify `--service0`, `--service1`, and so on.

- For *macfile*, specify the full path to the media access control (MAC) input file.

To retrieve more information about the `discover` command, type `discover --h`.

Example 1. The following command uses a MAC file to configure rack 1 and service node 0:

```
# discover --rack 1 --service 0 --macfile mymacfile
```

Example 2. If you have one rack and one service node, type the following command:

```
# discover --rack 1 --service 0
```

Example 3. If you have five racks and one service node, type the following command:

```
# discover --rackset 1,5 --service 0
```

7. (Conditional) When prompted to do so by the system, power up each individual rack or service node.

   Perform this step if you did not use a MAC file as input to the `discover` command.

   The system prompt for this action is as follows:

   ```
   At this time, please turn on the power to this service node.
   Do not turn the system on.
   ```

   The blue light on each component turns on when configuration is complete.

   You can use the `console`(1) command if you want to watch the installation progress. The sessions are also logged.

8. Type the following command to update the configuration files:

   ```
   # update-configs
   ```

9. Proceed to one of the following:

   - If you want to configure a backup domain name service (DNS) server, proceed to the following:

     "(Optional) Configuring a Backup Domain Name Service (DNS) Server" on page 68

- To configure the InfiniBand subnetworks, proceed to the following:

  "Configuring the InfiniBand Subnetworks" on page 69

(Optional) Configuring a Backup Domain Name Service (DNS) Server

Typically, the DNS on the system admin controller (SAC) provides name services for the SGI ICE X system. When you configure a backup DNS, however, the compute nodes can use a service node as a secondary DNS server if the SAC is down, being serviced, or is otherwise not available. You can configure a backup DNS only after you run the `discover` command to configure the cluster. This is an optional feature.

The following procedure explains how to configure a service node to act as a DNS.

Procedure 3-16 To enable a backup DNS

1. Through an `ssh` connection, log into the system admin controller (SAC) as the root user.
2. Type the following command to retrieve a list of available service nodes:

   # **`cnodes --service`**

   The service node you want to use as a backup DNS must be configured in the system already. That is, you must have run the `discover` command to configure the service nodes.
3. Type the following command to start the cluster configuration tool:

   # **`/opt/sgi/sbin/configure-cluster`**
4. On the **Main Menu** screen, select **B Configure Backup DNS Server (optional)**, and select **OK**.
5. On screen that appears, type the identifier for the service node that you want to designate as the backup DNS, and select **OK**.

   Figure 3-9 on page 69 shows how to specify `service0` as the backup DNS.

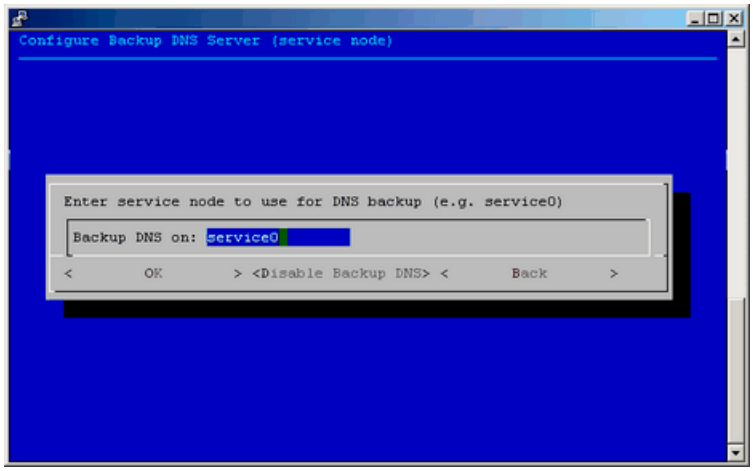

Figure 3-9 Configure Backup DNS Server (service node) screen

To disable this feature, select **Disable Backup DNS** from the same menu and select **Yes** to confirm your choice.

Configuring the InfiniBand Subnetworks

The InfiniBand network on SGI ICE X systems uses Open Fabrics Enterprise Distribution (OFED) software. For information about OFED, see http://www.openfabrics.org. For more information about the InfiniBand fabric implementation on SGI ICE X systems, see the *SGI ICE X Administrator's Guide*.

Each SGI ICE X system has two InfiniBand fabric network cards, `ib0` and `ib1`. Each subnetwork has a subnet manager, which runs on a rack leader controller (RLC). The following procedure explains how to specify which RLC you want to configure as the master and which you want to configure as the stand-by.

Procedure 3-17 To configure the InfiniBand network

1. Through an `ssh` connection, log into the system admin controller (SAC) as the root user.

2. Type the following command to disable InfiniBand switch monitoring:

   ```
   % cattr set disableIbSwitchMonitoring true
   ```

   The system sometimes issues InfiniBand switch monitoring errors before the InfiniBand network has been fully configured. The preceding command disables InfiniBand switch monitoring.

3. Use one of the following methods to access the InfiniBand network configuration tool:

   - Type the following command to start the cluster configuration tool:

     ```
     # configure-cluster
     ```

     After the cluster configuration tool starts, select **F Configure InfiniBand Fabric**, and select **OK**.

   - Type the following command to start the InfiniBand management tool:

     ```
     # tempo-configure-fabric
     ```

   Both of the preceding methods lead you to the same InfiniBand configuration page. On the InfiniBand configuration pages, **Quit** takes you to the previous screen.

4. Select **A Configure InfiniBand ib0**, and select **Select**.

5. On the **Configure InfiniBand** screen, select **A Configure Topology**, and select **Select**.

6. On the **Topology** screen, select the topology your system uses, and select **Select**.

   The menu selections are as follows:

   - **H HYPERCUBE**
   - **E EHYPERCUBE** (Enhanced Hypercube)
   - **F FAT TREE**
   - **G BFTREE**

7. On the **Configure InfiniBand** screen, select **B Master / Standby**, and select **Select**.

8. On the **Master / Standby** screen, type the RLC identifiers for the master (primary) and the standby (backup, secondary) subnetwork, and select **Select**.

   If you have only one rack leader controller (RLC), type `r1lead` in the **MASTER** field, and leave the **STANDBY** field blank.

   If you have more than one RLC, specify different RLCs in the **MASTER** and **STANDBY** fields.

   For example, Figure 3-10 on page 71 shows a completed screen.

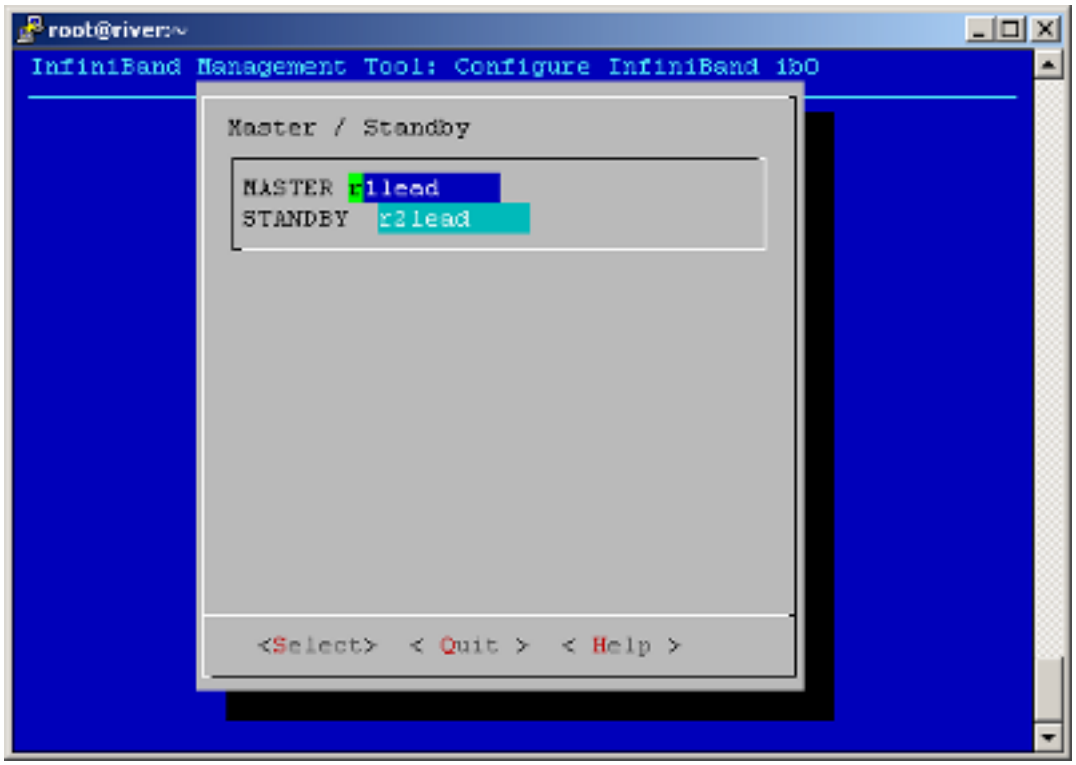

Figure 3-10 Completed InfiniBand (`ib0`) Master / Standby Screen

9. On the **Configure InfiniBand** screen, select **Commit**.

   Wait for the comfirmatory messages to appear in the window before you continue.

The next few steps repeat the preceding steps, but this time you configure the `ib1` interface.

10. On the InfiniBand Management Tool main menu screen, select **B Configure InfiniBand ib1**, and select **Select**.
11. On the **Configure InfiniBand** screen, select **A Configure Topology**, and select **Select**.
12. On the **Topology** screen, select the topology your system uses, and select **Select**.

    Select the topology that exists on your system. The menu selections are as follows:
13. On the **Configure InfiniBand** screen, select **B Master / Standby**, and select **Select**.
14. On the **Master / Standby** screen, type the RLC identifiers for the master (primary) and the standby (backup, secondary) subnetwork, and select **Select**.

    If you have only one rack leader controller (RLC), type `r1lead` in the **MASTER** field, and leave the **STANDBY** field blank.

    If you have two RLCs, you can flip the specifications for `ib1`. For example, assume that for `ib0`, you specified **MASTER** as `r1lead` and **STANDBY** as `r2lead`. For `ib1`, you can specify **MASTER** as `r2lead` and **STANDBY** as `r1lead`.

    If you have three or more RLCs, specify different RLCs in the **MASTER** and **STANDBY** fields.
15. On the **Configure InfiniBand** screen, select **Commit**.

    Wait for the comfirmatory messages to appear in the window before you continue.
16. On the InfiniBand Management Tool main menu screen, select **C Administer Infiniband ib0**, and select **Select**.
17. On the **Administer InfiniBand** screen, select **Start**, and select **Select**.
18. On the **Start SM master_ib0 on ib0 succeeded!** screen, select **OK**.
19. Select **Quit** to return to the InfiniBand Management Tool main menu screen.

    The next few steps repeat the preceding steps, but this time you start the `ib1` interface.
20. On the InfiniBand Management Tool main menu screen, select **D Administer Infiniband ib1**, and select **Select**.

21. On the **Administer InfiniBand** screen, select **Start**, and select **Select**.
22. On the **Start SM master_ib1 on ib1 succeeded!** screen, select **OK**.
23. On the **Administer InfiniBand** screen, select **Status**, and select **Select**.

    The **Status** option returns information similar to the following:

    ```
    Master SM
    Host = r1lead
    Guid = 0x0002c9030006938b
    Fabric = ib0
    Topology = hypercube
    Routing Engine = dor
    OpenSM = running
    ```

24. Wait for the status messages to stop, and press `Enter`.
25. Select **Quit** on the menus that follow to exit the configuration tool.
26. Type the `cnodes --leader` command to retrieve the list of rack leader controller (RLC) IDs.

    In the next few steps, you verify that the InfiniBand network is working.

    For example:

    ```
    SAC:~ # cnodes --leader
    r1lead
    r2lead
    ```

27. Through an `ssh`(1) connection, log into a leader node.

    For example:

    ```
    SAC:~ # ssh r2lead
    ```

28. Type the following command to retrieve the IDs of the compute nodes (blades):

    ```
    r2lead:~ # cnodes --compute
    r2i0n0
    r2i0n1
    r2i0n2
    r2i0n3
    ```

29. Type a `ping`(8) command to make sure that the RLC can reach its compute nodes.

For example:

```
r2lead:~ # ping -c1 r2i0n0
```

If the ping(8) is successful, the InfiniBand network is configured properly.

30. Type the following command to reenable InfiniBand switch monitoring:

```
% cattr unset disableIbSwitchMonitoring
```

31. (Optional) Configure additional features.

    The SGI ICE X system supports several optional features, for example, networking features such as network address translation. For information about how to configure optional features, see the following:

    Chapter 4, "Configuring Optional Features" on page 75

Configuring Optional Features

This chapter includes the following topics:

- "About the Optional SGI ICE X Features" on page 75
- "Configuring a Service Node as a Network Address Translation (NAT) Gateway" on page 76
- "Configuring a File System on a Service Node for use with a Network File System (NFS) Server" on page 80
- "Configuring a Service Node as a Network File System (NFS) Server" on page 84
- "Configuring Service Nodes and/or Compute Nodes as Network Information Service (NIS) Clients to the House Network's NIS Server" on page 88
- "RHEL Service Node House Network Configuration " on page 95
- "Configuring a Service Node as a Network Information Service (NIS) Server" on page 96
- "Installing MPI on a Running SGI ICE X System" on page 106
- "Troubleshooting Configuration Changes" on page 109

About the Optional SGI ICE X Features

You can configure one or more of the following optional features in this chapter for your SGI ICE X system:

- "Configuring a Service Node as a Network Address Translation (NAT) Gateway" on page 76
- "Configuring a File System on a Service Node for use with a Network File System (NFS) Server" on page 80
- "Configuring a Service Node as a Network File System (NFS) Server" on page 84
- "Configuring Service Nodes and/or Compute Nodes as Network Information Service (NIS) Clients to the House Network's NIS Server" on page 88
- "RHEL Service Node House Network Configuration " on page 95

- "Configuring a Service Node as a Network Information Service (NIS) Server" on page 96
- "Installing MPI on a Running SGI ICE X System" on page 106

Configuring a Service Node as a Network Address Translation (NAT) Gateway

The procedure in this topic explains how to configure network address translation (NAT) for your ICE X system. The procedure configures NAT on a service node. There is no need to configure NAT on compute nodes or any of the other node types.

Complete the procedure in this topic if you want to run a network file system (NFS) client or a network information service (NIS) client (also known as a *yp client*) on the compute nodes. The procedure in this topic configures NAT on node `service0`. This procedure assumes a typical configuration in which `eth0` on `service0` is connected to the house network. If you have trouble with the following procedure, see the following information in the troubleshooting chapter:

"Troubleshooting a Network Address Translation (NAT) Configuration" on page 132

The following procedure explains how to enable NAT on a service node.

Procedure 4-1 To enable NAT on a service node

1. Through an `ssh` connection, log into the system admin controller (SAC) as the root user.
2. Type the `cnodes --service` command to retrieve information about the service nodes in your system.

   This command shows the service nodes that are available. You can configure NAT on any of these nodes. This example uses `service0`.

   For example:

   ```
   # cnodes --service
   service0
   service1
   service2
   ```

3. Through an `ssh` connection, log into one of the service nodes as the root user.

For example:

```
# ssh service0
```

4. Type the following command to change to the directory where the NAT configuration script resides:

```
# cd /opt/sgi/docs/setting-up-NAT
```

5. Type the following command to enable execute permission on the file named README:

```
# chmod 755 README
```

6. Type the following command to run the README file:

```
# ./README
net.ipv4.ip_forward = 1
+ iptables-restore
+ modprobe ip_conntrack_tftp
+ modprobe ip_nat_tftp
```

Output similar to the preceding appears on your screen when the README script runs correctly.

7. Type the `ifconfig ib0` command to retrieve the IP address of InfiniBand network card `ib0`.

For example:

```
service0:~ # ifconfig ib0
ib0       Link encap:InfiniBand  HWaddr 80:00:04:04:FE:C0:00:00:00:00:00:00:00:00:00:00:00:00:00:00
          inet addr:10.148.0.2  Bcast:10.148.255.255  Mask:255.255.0.0
          inet6 addr: fe80::202:c902:26:403d/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:65520  Metric:1
          RX packets:1973872 errors:0 dropped:0 overruns:0 frame:0
          TX packets:1612831 errors:0 dropped:97 overruns:0 carrier:0
          collisions:0 txqueuelen:256
          RX bytes:232879516 (222.0 Mb)  TX bytes:582347073 (555.3 Mb)

service0:~ #
```

The IP address of `ib0` is `10.148.0.2`.

8. Type `logout` to log out from the service node and return to the SAC.

9. Type the following command to change to the directory where the compute node update script resides:

   # **cd /opt/sgi/share/per-host-customization/global**

   The system runs these scripts at startup. The next few steps explain how to edit the `sgi-static-routes` file to point to `ib0`.

10. Use a text editor to open file `sgi-static-routes`.

    The next few steps in this procedure modify the file. As a precaution, you can copy the file to a backup location before you begin to edit.

11. Search for a line that begins with `echo "default`.

    This line should include the IP address of `ib0` and the literal string `ib0`. The line might be correct in the file, but if necessary, edit the line. For this example, edit the line to be as follows:

    ```
    echo "default 10.148.0.2 ib0 -" >>${imagedir}/etc/sysconfig/network/routes
    ```

    The `sgi-static-routes` script customizes the network routing based upon the rack, the individual rack unit (IRU), and the slot of the compute blade. Some examples are available in the script. The next few steps boot the compute nodes.

12. Type the following command to shut down and stop all the compute nodes:

    SAC:~ # **cpower --halt r*i*n***

13. Type the following command to propagate the changes:

    SAC:~ # **cimage --push-rack**

14. Type the following command to power-up all the compute nodes:

    SAC:~ # **cpower --up r*i*n***

    When you power-up the computes nodes, the `sgi-static-routes` script runs and updates the default route information configured for NAT.

15. Type the following command to retrieve a list of the rack leader controller (RLC) nodes:

    SAC:~ # **cnodes --leader**

16. Through an `ssh`(1) connection, log into one of the leader nodes.

For example:

```
SAC:~ # ssh r1lead
```

17. Type the following command to retrieve a list of the compute nodes attached to this RLC:

```
r1lead:~ # cnodes --compute
```

18. Through an `ssh`(1) connection, log into one of the compute nodes.

For example:

```
r1lead:~ # ssh r1i1n0
```

Note: To log into a compute node, always log into the rack leader controller (RLC) first. You cannot log into a compute node directly from the SAC or from a service node.

---

19. Type the `ping`(8) command in the following format to verify that the compute node can access the service node through the InfiniBand `ib0` subnetwork:

`ping -c 1` *ib0_IP_addr*

In the preceding format, note the following:

- The `-c 1` parameter restricts the output to one `ECHO_REQUEST` packet.
- For *ib0_IP_addr*, specify the IP address of the InfiniBand `ib0` subnetwork. This is the IP address you retrieved in the following previous step:

  Procedure 4-1, step 7 on page 77

For example:

```
r1i3n0:~ # ping -c 1 10.148.0.2
PING 10.148.0.2 (10.148.0.2) 56(84) bytes of data.
64 bytes from 10.148.0.2: icmp_seq=1 ttl=64 time=3.90 ms

--- 10.148.0.2 ping statistics ---
1 packets transmitted, 1 received, 0% packet loss, time 0ms
rtt min/avg/max/mdev = 3.904/3.904/3.904/0.000 ms
```

20. Type the following command to verify the InfiniBand address on the compute node:

```
r1i3n0:~ # ifconfig ib0
ib0       Link encap:InfiniBand  HWaddr 80:00:04:04:FE:C0:00:00:00:00:00:00:00:00:00:00:00:00:00:00
          inet addr:10.148.0.57  Bcast:10.148.255.255  Mask:255.255.0.0
          inet6 addr: fe80::230:487a:c4e0:1/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:65520  Metric:1
          RX packets:125967 errors:0 dropped:0 overruns:0 frame:0
          TX packets:143324 errors:0 dropped:7 overruns:0 carrier:0
          collisions:0 txqueuelen:256
          RX bytes:33073064 (31.5 Mb)  TX bytes:22602331 (21.5 Mb)
```

21. Type the following command to verify the default gateway to the service node:

    ```
    r1i3n0:~ # netstat -rn
    ```

    Make sure that the default route shown in the output is to the service node (that is, to the NAT).

22. On the service node, type the following command(s) to verify that the service node can communicate with the compute nodes:

    ```
    service0:~ # date
    Mon Dec  3 12:14:13 CST 2012
    service0:~ # date ; cexec --pipe date
    Mon Dec  3 12:14:23 CST 2012
    blades r1i3n0: Mon Dec  3 12:14:24 CST 2012
    blades r1i3n1: Mon Dec  3 12:14:24 CST 2012
    blades r1i3n2: Mon Dec  3 12:14:24 CST 2012
    blades r1i3n3: Mon Dec  3 12:14:24 CST 2012
    service0:~ # cexec --pipe date
    blades r1i3n0: Mon Dec  3 12:14:48 CST 2012
    blades r1i3n1: Mon Dec  3 12:14:48 CST 2012
    blades r1i3n2: Mon Dec  3 12:14:48 CST 2012
    blades r1i3n3: Mon Dec  3 12:14:48 CST 2012
    ```

Configuring a File System on a Service Node for use with a Network File System (NFS) Server

The procedure in this topic explains how to configure a file system on a service node. This procedure assumes the SUSE Linux Enterprise Server (SLES) platform. If you

use the Red Hat Enterprise Linux (RHEL) platform, use your operating system documentation to complete this procedure.

Procedure 4-2 To configure an NFS home server on a service node

1. Through an `ssh` connection, log into the system admin controller (SAC) as the root user, and then log into the service node as the root user.

   The example in this procedure assumes that you want to configure `service0` as an NFS server.

   For example:

   ```
   # ssh mycluster
   root@sac # ssh service1
   root@service1 #
   ```

2. Retrieve the name of the root device.

   The following example shows the command to use and example output:

```
# ls -l /dev/disk/by-label/sgiroot
lrwxrwxrwx 1 root root 10 2008-03-18 04:27 /dev/disk/by-label/sgiroot -> ../../sda2
```

   The preceding command retrieves information that shows that the root device is named `sda`.

   Make sure you know which device is your root device. Do not take any actions that can repartition or otherwise destroy your root device.

3. Retrieve the names of the disk partitions, and use `by-id` notation.

   The steps in this procedure avoid using `/dev/sd`*X* notation in device names because device names in that style are not persistent. Those device names can change as you adjust disks and RAID volumes in your system. For example, you may assume that `/dev/sda` is the system disk and that `/dev/sdb` is a data disk. This is not always the case. To avoid accidental destruction of your root disk, the instructions in this procedure use `by-id` notation.

   Your goal is to retrieve the names of the non-root disk partitions. You can choose one of these partitions to host the NFS services. The following example shows the command to use and example output:

```
# ls -l /dev/disk/by-id
total 0
```

```
lrwxrwxrwx 1 root root  9 2012-03-20 04:57 ata-MATSHITADVD-RAM_UJ-850S_HB08_020520 -> ../../hdb
lrwxrwxrwx 1 root root  9 2012-03-20 04:57 scsi-3600508e000000000307921086e156100 -> ../../sda
lrwxrwxrwx 1 root root 10 2012-03-20 04:57 scsi-3600508e000000000307921086e156100-part1 -> ../../sda1
lrwxrwxrwx 1 root root 10 2012-03-20 04:57 scsi-3600508e000000000307921086e156100-part2 -> ../../sda2
lrwxrwxrwx 1 root root 10 2012-03-20 04:57 scsi-3600508e000000000307921086e156100-part5 -> ../../sda5
lrwxrwxrwx 1 root root 10 2012-03-20 04:57 scsi-3600508e000000000307921086e156100-part6 -> ../../sda6
lrwxrwxrwx 1 root root  9 2012-03-20 04:57 scsi-3600508e0000000008dced2cfc3c1930a -> ../../sdb
lrwxrwxrwx 1 root root 10 2012-03-20 04:57 scsi-3600508e0000000008dced2cfc3c1930a-part1 -> ../../sdb1
lrwxrwxrwx 1 root root  9 2012-03-20 09:57 usb-PepperC_Virtual_Disc_1_0e159d01a04567ab14E72156DB3AC4FA \
 -> ../../sr0
```

The preceding output shows that ID `scsi-3600508e000000000307921086e156100` is in use by your system disk. This in-use status is revealed in the symbolic link that points to `../../sda`. This is the root disk device. Do not consider this disk device for NFS use.

The other disk in the listing has ID `scsi-3600508e0000000008dced2cfc3c1930a` and is linked to `/dev/sdb`. You can configure the NFS services on this disk because it is a separate physical disk and is not `sda`, which is the root disk.

The next few steps create a filesystem on the disk.

4. Create a new `msdos` label on the disk.

   This procedure uses the `parted`(8) utility in a command-line driven manner. If you prefer, you can use `parted`(8) interactively, or you can use a different partitioning tool.

   For example, the following command creates a new label on `/dev/disk/by-id/scsi-3600508e0000000008dced2cfc3c1930a`:

```
# parted /dev/disk/by-id/scsi-3600508e0000000008dced2cfc3c1930a mkpart primary ext2 0 249GB
Information: Don't forget to update /etc/fstab, if necessary.
```

5. Retrieve the size of the disk.

   For example, type the following command:

```
# parted /dev/disk/by-id/scsi-3600508e0000000008dced2cfc3c1930a print
Disk geometry for /dev/sdb: 0kB - 249GB
Disk label type: msdos
Number  Start   End     Size    Type      File system  Flags
Information: Don't forget to update /etc/fstab, if necessary.
```

6. Create a partition that spans the size of the disk.

   For example, type the following command:

```
# parted /dev/disk/by-id/scsi-3600508e0000000008dced2cfc3c1930a mkpart
primary ext2 0 249GB
Information: Don't forget to update /etc/fstab, if necessary.
```

7. Create a filesystem on the disk.

   You can choose the filesystem type.

   Example 1. This example shows how to create an ext3 filesystem. The number of blocks and the bytes-per-node ratio determine the default number of inodes that the command creates, but the command accepts parameters that enable you to control the number and size of the inodes. It can take 10 minutes or more to create one 500–GB filesystem using default `mkfs.ext3` command line parameters. The following example command uses the `-N` option to reduce the number of inodes to 20 million inodes:

```
# mkfs.ext3 -N 20000000 /dev/disk/by-id/scsi-3600508e0000000008dced2cfc3c1930a-part1
```

   Example 2. This example shows how to create an XFS filesystem. Generally, you can create an XFS file system in less time than it takes to create an ext3 filesystem. The command is as follows:

```
# mkfs.xfs /dev/disk/by-id/scsi-3600508e0000000008dced2cfc3c1930a-part1
```

8. Use a text editor to open file `/etc/fstab`.

9. Add a line at the end of file `/etc/fstab` that defines the new filesystem.

   Make sure to use the `by-id` path for the device. This `fstab` entry enables the operating system to mount the filesystem automatically the next time the system reboots.

   Example 1. The following line defines the ext3 filesystem that was created in Procedure 4-2, step 7 on page 83:

```
/dev/disk/by-id/scsi-3600508e0000000008dced2cfc3c1930a-part1       /home   ext3    defaults        1
```

Example 2. The following line defines the XFS filesystem that was created in Procedure 4-2, step 7 on page 83:

```
/dev/disk/by-id/scsi-3600508e0000000008dced2cfc3c1930a-part1       /home   xfs    defaults        1
```

10. Save and close the `/etc/fstab` file.
11. Type the following command to mount the new filesystem:

    # **mount -a**

12. Proceed to the following:

    "Configuring a Service Node as a Network File System (NFS) Server" on page 84

Configuring a Service Node as a Network File System (NFS) Server

The following procedure explains how to configure a service node as an NFS server.

Procedure 4-3 To configure an NFS server on a service node

1. Through an `ssh` connection, log into the system admin controller (SAC) as the root user.
2. Through an `ssh` connection, log into one of the service nodes as the root user.

   For example:

   ```
   SAC:~ # ssh service0
   ```

3. Type the following command to determine whether the `nfsserver` service is enabled:

   ```
   service0:~ # chkconfig --list | grep nfs
   nfs                     0:off  1:off  2:off  3:on   4:off  5:on   6:off
   nfsserver               0:off  1:off  2:off  3:off  4:off  5:off  6:off
   ```

   The second line of output indicates that the NFS server is not enabled on `service0`.

4. Type the following command to turn on the `nfsserver` service:

   ```
   service0:~ # chkconfig nfsserver on
   insserv: Service dbus is missed in the runlevels 4 to use service openibd
   ```

5. Type the following command to retrieve the list of file systems that NFS can export:

```
service0:~ # cat /etc/exports
# See the exports(5) manpage for a description of the syntax of this file.
# This file contains a list of all directories that are to be exported to
# other computers via NFS (Network File System).
# This file used by rpc.nfsd and rpc.mountd. See their manpages for details
# on how make changes in this file effective.
/home *(rw,sync,no_subtree_check)
```

6. Type the following command to create a directory:

```
service0:~ # mkdir /home
```

   This step creates an example directory. Alternatively, you could specify an entire file system, rather than the directory `/home`. This could be the file system that you created in the following procedure:

   "Configuring a File System on a Service Node for use with a Network File System (NFS) Server" on page 80

7. Type the following command to start the NFS server:

```
service0:~ # /etc/init.d/nfsserver start
Starting kernel based NFS server: idmapd mountd statd nfsd sm-notify done
```

8. Type the `exportfs -av` command to export the test NFS directory, `/home`.

   For example:

```
service0:~ # exportfs -av
exporting *:/home
```

9. Type the following command to create a file named `testfile` in the `/home` directory and to write `test` to `testfile`:

```
service0:~ # echo test >/home/testfile
```

10. Type the following command to make sure that file `testfile` was created correctly:

```
service0:~ # cat /home/testfile
test
```

11. Type the following command to retrieve the IP address of `ib0` on the service node:

```
service0:~ # ifconfig ib0
ib0       Link encap:InfiniBand  HWaddr 80:00:04:04:FE:C0:00:00:00:00:00:00:00:00:00:00:00:00:00:00
          inet addr:10.148.0.2  Bcast:10.148.255.255  Mask:255.255.0.0
          inet6 addr: fe80::202:c902:26:403d/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:65520  Metric:1
          RX packets:1974162 errors:0 dropped:0 overruns:0 frame:0
          TX packets:1613139 errors:0 dropped:97 overruns:0 carrier:0
          collisions:0 txqueuelen:256
          RX bytes:232925385 (222.1 Mb)  TX bytes:582398138 (555.4 Mb)
```

    You need the information in the `inet addr` field in this output. In this example, the IP address is `10.148.0.2`. You use this address in a later step.

12. Through an `ssh`(1) connection, log into one of the leader nodes.

    If necessary, type a `cnodes --leader` command to retrieve the ID of one of the system's RLCs.

    For example:

    ```
    service0:~ # ssh r1lead
    ```

13. Through an `ssh`(1) connection, log into one of the compute nodes.

    For example:

    ```
    r1lead:~ # ssh r1i3n0
    ```

14. Type the following command to retrieve mount information and display the NFS server's file system export list:

    `showmount -e` *ib0_IP_service0*

    For *ib0_IP_service0*, specify the IP address of `ib0` on service node 0.

    For example:

    ```
    r1i3n0:~ # showmount -e 10.148.0.2
    Export list for 10.148.0.2:
    /home *
    ```

15. Type the following command to create the mount point:

    ```
    r1i3n0:~ # mkdir /tmp/mnt
    ```

16. Type the following command to mount the file system on compute node `r1i3n0`:

    ```
    mount -t nfs ib0_IP_service0:/home /tmp/mnt
    ```

    For *ib0_IP_service0*, specify the host name, fully qualified domain name (FQDN), or IP address of the InfiniBand `ib0` subnetwork.

    For example:

    ```
    r1i3n0:~ # mount -t nfs 10.148.0.2:/home /tmp/mnt
    ```

17. Type the `cd /tmp/mnt` command to change to the mount point of the NFS directory.

18. Type the following command to display mount information:

    ```
    r1i3n0:/tmp/mnt # mount | grep 10.148.0.
    10.148.0.2:/home on /tmp/mnt type nfs (rw,addr=10.148.0.2)
    ```

19. Type the following command to make sure you can access the test file on `service0` from the compute node:

    ```
    r1i3n0:/tmp/mnt # cat /tmp/mnt/testfile
    test
    ```

20. Type `logout`, to log out from the compute node and return to the service node.

21. Type `logout` to log out from the service node and return to the RLC node.

22. Type `logout` to log out from the RLC node and return to the SAC.

23. Use the `cd`(1) command to change to the following directory:

    ```
    /opt/sgi/share/per-host-customization/global
    ```

    The following steps explain how to add the new file system to the `sgi-fstab` file and ensure that the new file system mounts.

24. Use a text editor to open the following file on the SAC:

    ```
    sgi-fstab
    ```

25. Within file `sgi-fstab`, add a line for file system's mount point, and then save and close the file.

26. Type the following command to shut down and stop all the compute nodes:

    ```
    SAC:~ # cpower --halt r*i*n*
    ```

    The next steps restart the compute nodes and mount the new filesystem on all the compute nodes. These steps ensure that the file system mounts on all compute nodes when you restart the system.

27. Type the following command to propagate the changes:

    ```
    SAC:~ # cimage --push-rack
    ```

28. Type the following command to power-up all the compute nodes:

    ```
    SAC:~ # cpower --up r*i*n*
    ```

    When you power-up the computes nodes, the file system mounts on all compute nodes. If the file system does not mount, see the troubleshooting information in the following topic:

    "Troubleshooting a Network Address Translation (NAT) Configuration" on page 132

Configuring Service Nodes and/or Compute Nodes as Network Information Service (NIS) Clients to the House Network's NIS Server

Perform the procedures in this topic if you want to configure your service nodes or compute nodes as NIS clients to your house network's NIS server. You can perform the procedures in this topic at any time after you configure network address translation (NAT) on a service node.

The compute nodes are enabled to access the house network at this point because previous procedures configured the default gateway on the compute nodes to a service node and because you configured the service node to run NAT. For information about how to configure NAT, see the following:

"Configuring a Service Node as a Network Address Translation (NAT) Gateway" on page 76

The following procedures explain how to configure the service nodes and the compute nodes as NIS clients:

- "Configuring a Service Node as a NIS Client" on page 89

- "Configuring a Compute Node as a NIS Client" on page 90
- "Propagating a Node's Configuration to Another Node" on page 94

Configuring a Service Node as a NIS Client

The following procedure explains how to configure a service node as a NIS client.

Procedure 4-4 To configure a service node as a NIS client

1. Through an `ssh` connection, log into the system admin controller (SAC) as the root user.
2. Through an `ssh` connection, log into one of the service nodes as the root user.

   For example:

   ```
   SAC:~ # ssh service0
   ```

3. Type the following command to start the `ypbind`(8) service:

   ```
   service0:~ # chkconfig ypbind on
   ```

4. Open file `/etc/yp.conf` in a text editor.
5. Add information about your site's house NIS server to file `/etc/yp.conf`, and then save and close the file.

   For example:

   ```
   domain duluth server 100.100.100.100
   ```

   The preceding example specifies NIS server `100.100.100.100` in the `duluth` domain.

6. Type the following command to start the NIS client on `service0`:

   ```
   service0:~ # /etc/init.d/ypbind start
   ```

7. Type the following command to verify that the service node client is communicating with the NIS server:

   ```
   service0:~ # ypwhich
   ```

   The output should contain the address of the NIS server, for example `100.100.100.100`.

8. Proceed as follows:

   - If you have other service nodes that you want to configure as NIS clients, repeat this procedure on those other service nodes.
   - If you want to configure compute nodes as NIS clients, proceed to the following:

     "Configuring a Compute Node as a NIS Client" on page 90

Configuring a Compute Node as a NIS Client

The following procedures explain how to configure a compute node as a NIS client. There is more than one way to accomplish this task, so choose from the following procedures:

- Method 1 — If you want to log into an existing compute node and configure only that one compute node as a NIS client, perform the following procedure:

  "Method 1 — Configuring an Individual Compute Node as a NIS Client" on page 90

- Method 2 — If you want to edit the master compute node image on the system admin controller, you can push the resulting master compute node image to any number of compute nodes. Use this procedure if you want all the compute nodes to be configured as NIS clients. Perform the following procedure:

  "Method 2 — Configuring the Master Compute Node Image as a NIS Client" on page 92

- Method 3 — If you want to propagate one node's image to another node, you can change the start-up scripts that run when you boot the system. This method assumes that you used one of the previous methods (Method 1 or Method 2) to configure an initial node and that you want to copy the initial node's configuration to another node. You can use this method to update the image on any kind of node. This method clones an image from one node to another node. Perform the following procedure:

  "Propagating a Node's Configuration to Another Node" on page 94

Method 1 — Configuring an Individual Compute Node as a NIS Client

The following procedure configures an individual compute node as a NIS client.

Procedure 4-5 To log into a compute node and configure that compute node as a NIS client

1. Through an `ssh` connection, log into the system admin controller (SAC) as the root user.
2. Through an `ssh`(1) connection, log into one of the rack leader controllers (RLCs).

   If necessary, type a `cnodes --leader` command to retrieve the ID of one of the system's RLCs.

   For example:

   ```
   SAC:~ # ssh r1lead
   ```

3. Through an `ssh`(1) connection, log into one of the compute nodes.

   If necessary, type a `cnodes --compute` command to retrieve the ID of one of the system's compute nodes.

   For example:

   ```
   r1lead:~ # ssh r1i3n0
   ```

4. Type the following command to start the `ypbind`(8) service:

   ```
   r1i3n0:~ # chkconfig ypbind on
   ```

5. Open file `/etc/yp.conf` in a text editor.
6. Add information about your site's house NIS server to file `/etc/yp.conf`, and then save and close the file.

   For example:

   ```
   domain duluth server 100.100.100.100
   ```

   The preceding example specifies NIS server `100.100.100.100` in the `duluth` domain.

7. Type the `ypwhich` command to verify that the service node client is communicating with the NIS server.

   The output should contain the address of the NIS server.

   For example:

   ```
   100.100.100.100
   ```

8. Type the `logout` command until you have returned to the SAC.

Method 2 — Configuring the Master Compute Node Image as a NIS Client

The following procedure configures the master compute node image on the system admin controller (SAC) as a NIS client. You can propagate this image to other compute nodes after you complete the following procedure.

Procedure 4-6 To log into the SAC and edit the master compute node image

1. Through an `ssh` connection, log into the system admin controller (SAC) as the root user.
2. Type the following command to locate the compute node images:

```
SAC:~ # cinstallman --show-images
Image Name               BT Path
compute-sles11sp2        1  /var/lib/systemimager/images/compute-sles11sp2
service-sles11sp2        0  /var/lib/systemimager/images/service-sles11sp2
lead-sles11sp2           0  /var/lib/systemimager/images/lead-sles11sp2
```

3. In a text editor, open the `yp.conf` file for the compute nodes.

   For example:

```
SAC:~ # vi /var/lib/systemimager/images/compute-sles11sp2/etc/yp.conf
```

4. Add information about your site's house NIS server to file `yp.conf`, and then save and close the file.

   For example:

```
domain duluth server 100.100.100.100
```

   The preceding example specifies NIS server `100.100.100.100` in the `duluth` domain.

5. Type the following command to power-down the compute nodes:

```
SAC:~ # cpower --off    r1i*n*
```

6. Type the following command to push the changes to the compute node image to the compute nodes on your system:

```
SAC:~ # cimage --push-rack compute-sles11sp2
```

7. Type the following command to boot the compute nodes:

   ```
   SAC:~ # cpower --on r1i*n*
   ```

8. Through an `ssh`(1) connection, log into one of the rack leader controllers (RLCs).

   If necessary, type a `cnodes --leader` command to retrieve the ID of one of the system's RLCs.

   For example:

   ```
   SAC:~ # ssh r1lead
   ```

9. Through an `ssh`(1) connection, log into one of the compute nodes.

   If necessary, type a `cnodes --compute` command to retrieve the ID of one of the system's compute nodes.

   For example:

   ```
   r1lead:~ # ssh r1i3n0
   ```

10. Type the `ypwhich` command to verify that the service node client is communicating with the NIS server.

    The output should contain the address of the NIS server.

    For example:

    ```
    100.100.100.100
    ```

11. (Optional) Troubleshoot the NIS configuration.

    Use one or more of the following commands to view or set the current root image on the SAC:

    ```
    cadmin --show-root-labels
    cadmin --show-default-root
    cadmin --show-current-root
    cadmin --set-root-label --slot 2 --label "xxxxx"
    cadmin --set-default-root --slot 2
    ```

12. Type the `logout` command until you have returned to the SAC.

Propagating a Node's Configuration to Another Node

You can copy the configuration of one node to another node. When the second node boots, the image is copied from the initial node to the secondary node. This topic explains how to clone the image from one node to another node.

For example, if you have a service node that is configured as a NIS client, you can copy (or clone) the image from the initial node to a second node. The procedure in this topic uses NIS client configuration information as an example, but you can use this procedure to propagate other system characteristics.

The following procedure explains how to propagate changes to multiple nodes.

Procedure 4-7 To propagate NIS client configuration

1. Through an `ssh` connection, log into the system admin controller (SAC) as the root user.
2. Use the `cd`(1) command to change to the following directory:

   `/opt/sgi/share/per-host-customization/global`
3. Use a text editor to open the configuration file that you need.
4. Add the information you need to the configuration file.

   For example, within file `sgi-fstab`, add a line for file system's mount point, and then save and close the file.
5. Shut down the nodes you want to reconfigure.

   For example, type the following command to shut down and stop all the compute nodes:

   ```
   SAC:~ # cpower --halt r*i*n*
   ```

   The next steps restart the compute nodes and mount the new filesystem on all the compute nodes. These steps ensure that the file system mounts on all compute nodes when you restart the system.
6. Type the following command to propagate the changes:

   ```
   SAC:~ # cimage --push-rack
   ```
7. Power up the nodes.

For example, type the following command to power-up all the compute nodes:

```
SAC:~ # cpower --up r*i*n*
```

When you power-up the computes nodes, the system uses the NIS server specifications for all compute nodes and starts the `ypbind` service.

RHEL Service Node House Network Configuration

If you plan to put your service node on the house network, you need to configure it for networking. For this, you may use the `system-config-network` command. It is better to use the graphical version of the tool if you are able. Use the `ssh -X` command from your desktop to connect to the system admin controller (SAC) and then again to connect to the service node. This should redirect graphics over to your desktop.

Some helpful hints are, as follows:

- On service nodes, the cluster interface is `eth0`. Therefore, do not configure this interface as it is already configured for the cluster network.
- Do not make the public interface a `dhcp` client as this can overwrite the `/etc/resolv.conf` file.
- Do not configure name servers, the name server requests on a service node are always directed to the rack leader controller (RLC) for resolution. If you want to resolve network addresses on your house network, just be sure to enable the **House DNS Resolvers** using `configure-cluster` command on the system admin controller (SAC).
- Do not configure or change the search order, as this again could adjust what cluster management has placed in the `/etc/resolv.conf` file.
- Do not change the host name using the RHEL tools. You can change the hostname using the `cadmin` tool on the SAC.
- After configuring your house network interface, you can use the `ifupethX` command to bring the interface up. Replace `X` with your house network interface.
- If you wish this interface to come up by default when the service node reboots, be sure `ONBOOT` is set to yes in `/etc/sysconfig/network-scripts/ifcfg-ethX` (again, replace `X` with the proper value). The graphical tool allows you to adjust this setting while the text tool does not.

- If you happen to wipe out the `resolv.conf` file by accident and end up replacing it, you may need to issue this command to ensure that DNS queries work again:

  # **`nscd --invalidate hosts`**

- Having a single, small server provide filesystems to the whole SGI ICE X system could create network bottlenecks that the hierarchical design of SGI ICE X is meant to avoid, especially if large files are stored there. Consider putting your home filesystems on an NAS file server.

- If you want to use NAS server for scratch storage or make home filesystems available on NAS, you can follow the instructions in "Configuring a File System on a Service Node for use with a Network File System (NFS) Server" on page 80. In this example, you need to replace `service0-ib1` with the `ib1` InfiniBand host name for the NAS server and you need to know where on the NAS server the home filesystem is mounted to craft the `sgi-fstab` script properly.

- For information on centrally managed user accounts, see "Configuring a Service Node as a Network Information Service (NIS) Server" on page 96. It describes NIS master set up. In this design, the master server residing on the service node provides the filesystem and the NIS slaves reside on the RLCs. If you have more than one home server, you need to export all home filesystems on all home servers to the server acting as the NIS master. You also need to export the filesystems to the NIS master using the `no_root_squash export` flag.

Configuring a Service Node as a Network Information Service (NIS) Server

You can enable a NIS server on one of the service nodes on your SGI ICE X system. Make sure you consider the following when you configure NIS:

- You can configure a service node to be a NIS master server, and you can configure the rack leader controllers (RLCs) as the NIS slave servers

  Do not configure the system admin controller (SAC) as the NIS master server. The SAC cannot mount all storage types. When you mount the storage on the NIS master server, you can use NIS to add accounts.

- If multiple service nodes provide home filesystems, the NIS master server should mount all the remote home filesystems. You need to export home filesystems to the NIS master service node with the `no_root_squash export` option. The examples in the following sections assume a single service node with storage and that same node is the NIS master.

- NIS traffic goes over the Ethernet. No NIS traffic goes over the InfiniBand network.

  The compute node NIS traffic goes over the Ethernet, by way of using the `lead-eth` server name in the `yp.conf` file. This design feature prevents NIS traffic from affecting the InfiniBand traffic between the compute nodes.

Determine the following before you begin your NIS configuration:

- Select the service node upon that you want to designate as the NIS master server. You can configure the other service node(s) as NIS clients. The rack leader controllers (RLCs) in your SGI ICE X system can become NIS slave servers.
- Select an NIS domain name. For example: `ice`.

The procedures assume a SLES 11 operating system. The following topics explain how to configure a service node as a NIS master server:

- "Configuring a Network Information Service (NIS) Master Server and One or More NIS Slave Servers" on page 97
- "Configuring a Network Information Service (NIS) Client on a Service Node" on page 99
- "Configuring a Rack Leader Controller (RLC) as a Network Information Server (NIS) Slave Server and Client (SLES)" on page 100
- "NAS Configuration for Multiple IB Interfaces" on page 103
- "Configuring the Compute Nodes as Network Information Service (NIS) Clients (SLES)" on page 102
- "Creating User Accounts (SLES)" on page 106

If you want to use an existing house network NIS server, see "Configuring Service Nodes and/or Compute Nodes as Network Information Service (NIS) Clients to the House Network's NIS Server" on page 88.

Configuring a Network Information Service (NIS) Master Server and One or More NIS Slave Servers

The following procedure explains how to configure a service node as a NIS master server and one or more rack leader controllers (RLCs) as NIS slave servers. The procedure applies to service nodes that run the SLES 11 operating system and uses the text-based YaST2 interface. The graphical YaST2 interface is slightly different.

Procedure 4-8 To configure a service node as a NIS master server

1. Type the following command to start YaST2:

   # **yast nis_server**

2. Select **Create NIS Master Server**, and select **Next** to continue.

3. Choose an NIS domain name, and type the name into the **NIS Domain Name window**.

   This example uses **ice**.

4. Select **This host is also a NIS client**.

5. Select **Active Slave NIS server exists**.

6. Select **Fast Map distribution**.

7. Select **Allow changes to passwords**.

8. Click **Next** to continue.

   You are now in the **NIS Master Server Slaves Setup**.

   At this point, you can enter the IDs for the RLCs. If you add new RLCs or if you reconfigure existing RLCs, you need to update this list.

9. In the **Edit Slave** screen, select **Add**, and type **r1lead**.

   If you have other RLCs that you want to configure as NIS slave servers, type the IDs for those RLCs, too.

   After you specified all the RLCs you want to configure as NIS slave servers, select **Next** to continue.

10. On the **NIS Server Maps Setup**, select **Next** .

    You can use the default selected maps.

    Do not use the **hosts** map. The **hosts** map is not selected by default. This map can interfere with SGI ICE X system operations.

11. On the **NIS Server Query Hosts Setup** screen, select **Finish**.

    You can use the default settings. You can, however, adjust the settings for security purposes.

At this point, the NIS master is configured. Assuming you checked the **This host is also a NIS client box**, the service node will be configured as a NIS client to itself and start `yp ypbind` for you.

Configuring a Network Information Service (NIS) Client on a Service Node

The following procedure explains how to configure your other service nodes to be broadcast-binding NIS clients. Do not configure the NIS client on the same service node that you configured as the NIS master server. For information about how to configure the NIS master server on a service node, see "Configuring a Network Information Service (NIS) Master Server and One or More NIS Slave Servers" on page 97.

The procedure applies only to service nodes that host SLES11, and the procedure uses the YaST2 interface.

Procedure 4-9 To configure a service node as a NIS client

1. Type the following command to enable `ypbind`:

   # **`chkconfig ypbind on`**

2. Type the following command to set the default domain:

   `echo "`*`NIS_domain_name`*`" > /etc/defaultdomain`

   For *NIS_domain_name*, type the domain name you created in Procedure 4-8, step 3 on page 98.

   For example:

   # **`echo "ice" > /etc/defaultdomain`**

3. In order to ensure that no NIS traffic goes over the IB network, SGI does **not** recommend using NIS broadcast binding on service nodes. You can list a few rack leader controllers (RLCs) the in `/etc/yp.conf` file on non-NIS-master service nodes. The following is an example `/etc/yp.conf` file. Add or remove RLCs as appropriate. Having more entries in the list allows for some redundancy. If `r1lead` is hit by excessive traffic or goes down, `ypbind` can use the next server in the list as its NIS server. SGI does not suggest listing other service nodes in `yp.conf` file because all resolvable names for service nodes on service

nodes use IP addresses that go over the InfiniBand network. For performance reasons, it is better to keep NIS traffic off of the InfiniBand network.

```
ypserver r1lead
ypserver r2lead
```

4. Start the `ypbind` service, as follows:

```
# rcypbind start
```

The service node is now bound.

5. Add the NIS include statement to the end of the password and group files, as follows:

```
# echo "+:::" >> /etc/group
# echo "+::::::" >> /etc/passwd
# echo "+" >> /etc/shadow
```

Configuring a Rack Leader Controller (RLC) as a Network Information Server (NIS) Slave Server and Client (SLES)

This section provides instructions for setting up rack leader controllers (RLCs) as NIS slave servers. It is possible to make all these adjustments to the RLC image in `/var/lib/systemimager/images`. Currently, SGI does not recommend using this approach.

Procedure 4-10 To configure an RLC as a NIS slave server

1. Through an `ssh` connection, log into the system admin controller (SAC) as the root user.
2. Type the following command to verify that the InfiniBand interfaces are configured and operational:

   If the system generates the following message, check to be sure that the InfiniBand network is operational.

   ```
   can't enumerate maps from service0
   ```

3. Type the following commands:

```
admin:~ # cexec --head --all chkconfig ypserv on
admin:~ # cexec --head --all chkconfig ypbind on
```

```
admin:~ # cexec --head --all chkconfig portmap on
admin:~ # cexec --head --all chkconfig nscd on
admin:~ # cexec --head --all rcportmap start
```

4. Type the following commands:

```
admin:~ # cexec --head --all "echo NIS_domain_name > /etc/defaultdomain"
admin:~ # cexec --head --all "ypdomainname NIS_domain_name"
```

   For *NIS_domain_name*, specify the NIS domain name at your site.

   For example, if the NIS domain name at your site is `ice`, type the following commands:

```
admin:~ # cexec --head --all "echo ice > /etc/defaultdomain"
admin:~ # cexec --head --all "ypdomainname ice"
```

5. Type the following commands:

```
admin:~ # cexec --head --all "echo ypserver node_ID > /etc/yp.conf"
admin:~ # cexec --head --all /usr/lib/yp/ypinit -s node_ID
```

   For *node_ID*, specify the service node ID that you configured as the NIS master server.

   For example, if `service0` is the NIS master server at your site, type the following commands:

```
admin:~ # cexec --head --all "echo ypserver service0 > /etc/yp.conf"
admin:~ # cexec --head --all /usr/lib/yp/ypinit -s service0
```

6. Type the following commands:

```
admin:~ # cexec --head --all rcportmap start
admin:~ # cexec --head --all rcypserv start
admin:~ # cexec --head --all rcypbind start
admin:~ # cexec --head --all rcnscd start
```

Configuring the Compute Nodes as Network Information Service (NIS) Clients (SLES)

You can configure NIS on the clients to use a server list that only contains the their rack leader controller (RLC). All operations are performed from the system admin controller (SAC).

The following procedure explains how to configure the compute nodes (blades) as NIS clients.

Procedure 4-11 To configure the compute nodes as NIS clients

1. Through an `ssh` connection, log into the system admin controller (SAC) as the root user.
2. Create a compute node image clone. SGI recommends that you always work with a clone of the compute node images. For information on how to clone the compute node image, see the *SGI ICE X Administration Guide*.
3. Type the following command to specify that the compute nodes use the cloned image/kernel pair:

```
admin:~ # cimage --set compute-sles11-clone 2.6.16.46-0.12-smp "r*i*n*"
```

4. Type the following command to configure the NIS domain:

```
admin:~ # echo "NIS_domain_name" > /var/lib/systemimager/images/compute-sles11-clone/etc/defaultdomain
```

   For *NIS_domain_name*, specify the NIS domain name at your site.

   For example:

```
admin:~ # echo "ice" > /var/lib/systemimager/images/compute-sles11-clone/etc/defaultdomain
```

5. Type the following command to enable the compute nodes to get NIS services from their RLC (fix the domain name as appropriate):

```
admin:~ # echo "ypserver lead-eth" > /var/lib/systemimager/images/compute-sles11-clone/etc/yp.conf
```

6. Type the following command to enable the `ypbind` service:

```
admin:~# chroot /var/lib/systemimager/images/compute-sles11-clone chkconfig ypbind on
```

7. Type the following commands to configure the password, shadow, and group files with NIS includes:

```
admin:~# echo "+:::" >> /var/lib/systemimager/images/compute-sles11-clone/etc/group
admin:~# echo "+::::::" >> /var/lib/systemimager/images/compute-sles11-clone/etc/passwd
```

```
admin:~# echo "+" >> /var/lib/systemimager/images/compute-sles11-clone/etc/shadow
```

8. Type the following command to push out the updates:

```
admin:~ # cimage --push-rack compute-sles11-clone "r*"
```

NAS Configuration for Multiple IB Interfaces

You can attach storage devices to a service node on an SGI ICE X system. The NAS cube needs to be configured such that each InfiniBand fabric interface is in a separate subnetwork. The following procedure logically separates the interfaces and attaches them to the same physical network. The procedure configures the large physical network into four smaller subnets. Each subnet becomes capable of containing all the nodes, including the service nodes. The subnets you configure are as follows:

- `10.149.0.0/18`
- `10.149.64.0/18`
- `10.149.128.0/18`
- `10.149.192.0/18`

This procedure assumes the following:

- The `-ib1` InfiniBand fabric for the compute nodes has addresses assigned in the `10.149.0.0/16` network.
- The lowest address that the cluster management software uses is `10.149.0.1` and the highest is `10.149.1.3` (already assigned to the NAS cube).

After the discovery of the storage node has happened, SGI personnel will need to log onto the NAS box and change the network settings to use the smaller subnets, and then define the other three adapters with the same offset within the subnet; for example: Initial configuration of the storage node had set `ib0` fabric's IP to `10.149.1.3 netmask 255.255.0.0`. After the addresses are changed, `ib0=10.149.1.3:255.255.192.0, ib1=10.149.65.3:255.255.192.0`, `ib2=10.149.129.3:255.255.192.0, ib3=10.149.193.3:255.255.192.0`. The NAS cube should now have all four adapter connections connected to the fabric with IP addresses which can be pinged from the service node.

Note: The service nodes and the rack leads will remain in the 10.149.0.0/16 subnet.

Procedure 4-12 To configure NAS

1. Through an `ssh` connection, log into the system admin controller (SAC) as the root user.
2. Use a text editor to open file `/opt/sgi/share/per-host-customization/global/sgi-setup-ib-configs`.

   The next few steps in this procedure modify the file significantly. As a precaution, you can copy the file to a backup location before you begin to edit.
3. Search for `iruslot=$1`.
4. After the line that contains `iruslot=$1`, add the following lines:

```
# Compute NAS interface to use
IRU_NODE=‘basename ${iruslot}‘
RACK=‘cminfo --rack‘
RACK=$(( ${RACK} - 1 ))
IRU=‘echo ${IRU_NODE} | sed -e s/i// -e s/n.*//‘
NODE=‘echo ${IRU_NODE} | sed -e s/.*n//‘
POSITION=$(( ${IRU} * 16 + ${NODE} ))
POSITION=$(( ${RACK} * 64 + ${POSITION} ))
NAS_IF=$(( ${POSITION} % 4 ))
NAS_IPS[0]="10.149.1.3"
NAS_IPS[1]="10.149.65.3"
NAS_IPS[2]="10.149.129.3"
NAS_IPS[3]="10.149.193.3"
```

5. Search for `$iruslot/etc/opt/sgi/cminfo`.
6. After the line that contains `$iruslot/etc/opt/sgi/cminfo`, add the following lines:

```
IB_1_OCT12=‘echo ${IB_1_IP} | awk -F "." ’{ print $1 "." $2 }’‘
IB_1_OCT3=‘echo ${IB_1_IP} | awk -F "." ’{ print $3 }’‘
IB_1_OCT4=‘echo ${IB_1_IP} | awk -F "." ’{ print $4 }’‘
IB_1_OCT3=$(( ${IB_1_OCT3} + ${NAS_IF} * 64 ))
IB_1_NAS_IP="${IB_1_OCT12}.${IB_1_OCT3}.${IB_1_OCT4}"
```

7. Search for `IPADDR=’${IB_1_IP}’`, and replace it with `IPADDR=’${IB_1_NAS_IP}’`.
8. Search for `NETMASK=’${IB_1_NETMASK}’`, and replace it with `NETMASK=’255.255.192.0’`.

9. Go to the end of the file, and add the following lines:

```
# ib-1-vlan config
cat << EOF >$iruslot/etc/sysconfig/network/ifcfg-vlan1
# ifcfg config file for vlan ib1
BOOTPROTO='static'
BROADCAST=''
ETHTOOL_OPTIONS=''
IPADDR='${IB_1_IP}'
MTU=''
NETMASK='255.255.192.0'
NETWORK=''
REMOTE_IPADDR=''
STARTMODE='auto'
USERCONTROL='no'
ETHERDEVICE='ib1'
EOF
if [ $NAS_IF -eq 0 ]; then
    rm $iruslot/etc/sysconfig/network/ifcfg-vlan1
fi
```

10. Save and close the file.

11. Use a text editor to open file `/opt/sgi/share/per-host-customization/global/sgi-fstab`.

    The next few steps in this procedure modify the file significantly. As a precaution, you can copy the file to a backup location before you begin to edit.

12. Your goal is to modify file `sgi-fstab` for the compute blades by adding lines similar to the lines you added to file `/opt/sgi/share/per-host-customization/global/sgi-setup-ib-configs`.

    Perform the following steps:

1. Add a `# Compute NAS interface to use` section into this file. For information about this section, see Procedure 4-12, step 4 on page 104.

2. Add lines similar to the following to specify mount points:

```
# SGI NAS Server Mounts
${NAS_IPS[${NAS_IF}]}:/mnt/data/scratch     /scratch nfs    defaults 0 0
```

Creating User Accounts (SLES)

The example used in this section assumes that the home directory is mounted on the NIS Master service and that the NIS master is able to create directories and files on it as root. The following example use command line commands. You could also create accounts using YaST.

The following procedure explains how to create user accounts on an SGI ICE X system.

Procedure 4-13 Creating User Accounts on a NIS Server

1. Through an `ssh` connection, log in to the NIS master service node as the root user.
2. Use the `useradd`(8) command to add the new user and create a home directory for the new user.

   For example:

   # **useradd -c "Joe User" -m -d /home/juser juser**
3. Use the `passwd`(1) command to create a password for the new user.

   For example:

   # **passwd juser**
4. Type the following command to push the new account to the NIS servers:

   # **cd /var/yp && make**

Installing MPI on a Running SGI ICE X System

This section describes how to install MPI on an SGI ICE X system that has already been installed. The instructions in this section update existing images instead of creating new ones. It should be noted that integrating MPI before cluster deployment is easier.

SGI supplied media, such as SGI® MPI and SGI® Accelerate™ CDs, have embedded in them suggested package lists for each node type. The `crepo` command, used in the following example, makes use of these lists and indeed recomputes the lists when new media is added and then selected.

File names in this example are just illustrations.

Register SGI MPI and SGI Accelerate with SMC, as follows:

```
# crepo --add accelerate-1.0-cd1-media-rhel6-x86_64.iso
# crepo --add mpi-1.0-cd1-media-rhel6-x86_64.iso
```

Update the `crepo` selected repositories so that all repositories associated with the software distribution (distro) you are installing are present. For example, if you want MPI to work on RHEL 6, you might do something like this:

Show what is currently selected (the asterisks to the left):

```
# crepo --show
* SGI-Management-Center-1.5-rhel6 : /tftpboot/sgi/SGI-Management-Center-1.5-rhel6
* SGI-Foundation-Software-2.5-rhel6 : /tftpboot/sgi/SGI-Foundation-Software-2.5-rhel6
* SGI-XFS-XVM-2.5-for-RHEL-rhel6 : /tftpboot/sgi/SGI-XFS-XVM-2.5-for-RHEL-rhel6
* SGI-Accelerate-1.3-rhel6 : /tftpboot/sgi/SGI-Accelerate-1.3-rhel6
* SGI-Tempo-2.5-rhel6 : /tftpboot/sgi/SGI-Tempo-2.5-rhel6
* SGI-MPI-1.3-rhel6 : /tftpboot/sgi/SGI-MPI-1.3-rhel6
* Red-Hat-Enterprise-Linux-6.2 : /tftpboot/distro/rhel6.2
```

Unselect unrelated repositories:

```
# crepo --unselect SGI-Tempo-2.5-rhel6
Updating: /etc/opt/sgi/rpmlists/generated-compute-rhel.6.2.rpmlist
Updating: /etc/opt/sgi/rpmlists/generated-service-rhel6.2.rpmlist
# crepo --unselect SGI-Foundation-Software-2.5-rhel6
Updating: /etc/opt/sgi/rpmlists/generated-compute-rhel6.2.rpmlist
Updating: /etc/opt/sgi/rpmlists/generated-service-rhel6.2.rpmlist
# crepo --unselect Red-Hat-Enterprise-Linux-6.2
Removing: /etc/opt/sgi/rpmlists/generated-compute-rhel6.2.rpmlist
Removing: /etc/opt/sgi/rpmlists/generated-service-rhel6.2.rpmlist
```

Select RHEL 6 related repositories:

```
# crepo --select Red-Hat-Enterprise-Linux-6.2
Updating: /etc/opt/sgi/rpmlists/generated-compute-rhel6.2.rpmlist
Updating: /etc/opt/sgi/rpmlists/generated-lead-rhel6.2.rpmlist
Updating: /etc/opt/sgi/rpmlists/generated-service-rhel6.2.rpmlist
# crepo --select SGI-Foundation-Software-2.5-rhel6
Updating: /etc/opt/sgi/rpmlists/generated-compute-rhel6.2.rpmlist
Updating: /etc/opt/sgi/rpmlists/generated-lead-rhel6.2.rpmlist
Updating: /etc/opt/sgi/rpmlists/generated-service-rhel6.2.rpmlist
```

```
# crepo --select SGI-XFS-XVM-2.5-for-RHEL-rhel6
Updating: /etc/opt/sgi/rpmlists/generated-compute-rhel6.2.rpmlist
Updating: /etc/opt/sgi/rpmlists/generated-lead-rhel6.2.rpmlist
Updating: /etc/opt/sgi/rpmlists/generated-service-rhel6.2.rpmlist
```

After performing the steps, above, the proper repositories are registered and selected so you can operate on them by default. Since you are using an already deployed system, you need to update existing images and potentially existing service nodes themselves. This example uses SGI suggested/ generated rpmlists. If you have custom rpmlists, you need to manually reconcile the two lists for each node type. The list fragments in `/var/opt/sgi/sgi-repodata/` may help you.

Note: The commands in this topic include a continuation character (\).

For a service node image, type the following:

```
# cinstallman --refresh-image --image service-rhel6.2 \
--rpmlist /etc/opt/sgi/rpmlists/generated-service-rhel6.2.rpmlist
```

For a compute node image, type the following:

```
# cinstallman --refresh-image --image compute-rhel6.2 \
--rpmlist /etc/opt/sgi/rpmlists/generated-compute-rhel6.2.rpmlist
```

Finally, you need to push the updated compute image to the rack leader controllers (RLCs).

Note: If the compute nodes are booted on the image and are using NFS for roots, you need to shut the compute nodes down before being able to run this command.

```
# cimage --push-rack compute-rhel6.2 r"*"
```

To make sure the compute nodes your are operating on have the associated compute image you just updated, perform a command similar to the following”

```
# cimage --set compute-rhel6.2 2.6.32-71.el6.x86_64 "*"
```

You can find the available images and kernels using the `cimage --list-images` command.

If you have booted service/login nodes, you likely want to refresh those running nodes also. (You could also reinstall them, as well). Here is a refresh example:

```
# cinstallman --refresh-node --node service0 \
--rpmlist /etc/opt/sgi/rpmlists/generated-service-rhel6.2.rpmlist
```

Now reset or bring up the nodes (depends on the state you left them). If you want to bring up all nodes, this command will not disrupt nodes already operating:

```
# cpower --system --up
```

Troubleshooting Configuration Changes

If a configuration change does not affect the SGI ICE X system in the intended manner, try one of the following approaches:

- Edit the node image on the system admin controller (SAC). For example, you can reconfigure the service node image on the SAC and reimage the service nodes with that new image.
- Edit the node customization scripts. For example, the compute node update scripts reside on the SAC in the `/opt/sgi/share/per-host-customization/global` directory.

Troubleshooting

This chapter provides answers to some common problems users encounter when installing an SGI ICE X system and includes diagnosis and troubleshooting information. It covers the following topics:

- "Initial Installation Settings" on page 112
- "System Discovery Overview" on page 112
- "Compute Nodes Are Taking Too Long To Boot" on page 115
- "Verify the Bonding Mode on the Rack Leader Controller (RLC)" on page 116
- "`cimage --push-rack` Pushes Too Many (or Too Few) Expansions" on page 119
- "Cannot `ping` the CMCs from the Rack Leader Controller (RLC)" on page 120
- "`r1lead` Configured with `vlan1`/`vlan2` and Not `vlan101`" on page 122
- "How to Make the `blademond` Daemon Start Over from Scratch" on page 122
- "Log Files" on page 123
- "CMC `slot_map` / `blademond` Debugging Hints" on page 123
- "`ssh` Commands to Compute Nodes: ssh Key Failures / Known Hosts" on page 125
- "Compute Node Hosts Seem to Actually be BMCs" on page 125
- "Resolving CMC Slot Map Ordering Issues" on page 125
- "In `tmpfs` Mode, File Has Date in the Future Warnings" on page 126
- "Ensuring Hardware Clock Has the Correct Time" on page 126
- "Configure Switches for a Rack Leader Controller (RLC)" on page 127
- "Switch Wiring Rules" on page 129
- "System Admin Controller (SAC) `eth2` Link in the Bond is Down" on page 130
- "No InfiniBand Interfaces on Rack Leader Controller (RLC), Service, or Compute Node Images" on page 131

- "Troubleshooting a Network Address Translation (NAT) Configuration" on page 132

Initial Installation Settings

This section describes some values you should set or verify when configuring the system for the first time before any system nodes are discovered:

- Discover management switches first. It is very important for systems with top level management switches that these switches be discovered before any other nodes. As other nodes are discovered, the already-discovered management switches are configured.
- Use the `configure-cluster` GUI to enable the switch management network.

  The `cmcdetectd` daemon sets switch management network value when it detects SGI ICE X components. Note that this only works when the top level switches have never before been configured for SGI ICE X. It is a good practice to set this value to yes using the `configure-cluster` GUI to enable the switch management network.
- Use the `configure-cluster` GUI **Configure Default Max Rack IRU Setting** option to set the default number of individual rack units (IRUs), supported by a rack leader controller (RLC). Set this value to the number of CMCs that will be served by each RLC. The default is 8.
- Use the `configure-cluster` GUI **Configure Redundant Management Network** option to turn on the redundant management network (RMN) system. If this system has switch stacks and each node has two Ethernet connections up to the stack, and both CMC-0 and CMC-1 ports on the CMC are connected, it is an RMN system.
- Use the `configure-cluster` GUI **Configure MySQL Replication (optional)** to enable MySQL Replication. This is particularly important for large systems.

System Discovery Overview

This section describes software related to system discovery.

`configure-cluster` Command

The `configure-cluster` command starts the cluster configuration tool. This menu-based tool enables you to configure certain global variables, for example:

- The redundant management network
- The switch management network
- The default max rack IRU setting
- The default `blademond` rescan interval

For more information about how to run the cluster configuration tool, see the *SGI ICE X Installation and Configuration Guide.*

Warning: These settings in `configure-cluster` represent the global defaults. They are adjustable per-node by parameters to the `discover` command and the `cadmin` command.

`cmcdetected` Daemon

The `cmcdetectd` daemon runs on the system admin controller (SAC). When it sees a chassis management controller (CMC) asking for an IP address, it looks at the client ID of the request. That client ID contains the rack number and slot number. The `cmcdetectd` daemon provides this information to the `switchConfig` application programming interface (API) and configures the top level switch fabric.

The `cmcdetectd` daemon performs the following:

- The `cmcdetectd` daemon really only starts working when at least one management switch is discovered.
- It configures the switch and "moves" the CMCs to the appropriate VLAN.
- If there are two switches in the switch stack, `cmcdetectd` through the `switchConfig` API configures the CMC ports for manual trunking (the CMC-0 and CMC-1 ports on the physical CMC).

- Once moved, the dynamic host configuration protocol (DHCP) requests are directed to the rack VLAN for a given rack and detected by the rack leader controller (RLC) when it is discovered.
- If the `cmcdetectd` daemon detects any SGI ICE X CMCs, it automatically sets the switch management network variable to true.
- If you install a second slot or reinstall a system, the switch is already configured and the `cmcdetectd` daemon does not see the requests any more. It is a good practice to manually configure the switch management network setting using the `configure-cluster` option.

`discover` Command

The `discover` command is used to discover the various non-compute nodes and switches including:

- Top level management switches
- External InfiniBand switches
- Rack leader controllers (RLCs)
- Service nodes

Using options to the `discover` command, you can control the bonding mode (active-backup or 802.3ad link aggregation), enable the switch management network (for SGI ICE X components of the system), Redundant Management Network, and many other things.

Note: Always discover management switches first. The `discover` command works with the `switchconfig` command to set up switches appropriately for RLCs, service nodes, and so on.

The `discover` command facilitates the first installation of a node. You can always reinstall already-discovered nodes with the `cinstallman` command.

`blademond` Daemon

Once the RLC is operating, the CMCs get their IP addresses. The `blademond` daemon creates the initial `dhcpd.conf` file (`ice.conf`) when it is started for the first time. It looks for CMCs, polls them to figure out where the compute nodes are, and

configures the system for these nodes. For new CMCs/IRUs, `blademond` powers them up for the first time.

- The `blademond` daemon notices when blades are removed, swapped, added and updates the system as needed.
- You can configure how often blademond checks/polls the CMCs (see `configure-cluster` **Configure blademond rescan interval** and `cadmin --set-blademond-rescan`.
- You can decide not to run `blademond` as a daemon (use `chkconfig` to turn it off on the Rack Leader Controller (RLC) and RLC image). In that case, you can run it in `--scan-once` mode by hand or via `cron`.

Compute Nodes Are Taking Too Long To Boot

When compute nodes are taking a long time to boot, perform the following:

- See "Verify the Bonding Mode on the Rack Leader Controller (RLC)" on page 116 to verify that the compute nodes have the proper bonding setup.
- Verify the rack leader controller (RLC) has a MegaRAID controller. 144 nodes will not boot well with 106x controllers, for example. You can verify this with `lspci` command.

  To verify the MeagaRAID battery is working and charged, perform the following:

  ```
  # /opt/MegaRAID/MegaCli/MegaCli64 -ShowSummary -a0
  ```

  You should see '`Status : Healthy` ' under 'BBU' (`BBU = Battery Backup Unit`).

  Note: If this is the first time the node has booted up, it takes several hours for the BBU to be charged.

- Verify cache is set to write-back, as follows:

  ```
  # /opt/MegaRAID/MegaCli/MegaCli64 -LDGetProp -Cache -LALL -a0
  ```

  Note: Never force write-back on if bad BBU (`-CachedBadBBU`) as data loss happens with an orderly shutdown that includes a power off.

When you see the output: `Cache Policy:WriteBack`, write-back is enabled.

To enable the write-back policy, perform the following:

```
# /opt/MegaRAID/MegaCli/MegaCli64 -LDSetProp -NoCachedBadBBU -LALL -a0
```

Verify the Bonding Mode on the Rack Leader Controller (RLC)

The redundant management network (RMN) is the default configuration for SGI ICE X systems. To verify the bonding mode, perform the following from the RLC:

```
r1lead:~ # cat /proc/net/bonding/bond0
Ethernet Channel Bonding Driver: v3.5.0 (November 4, 2008)

Bonding Mode: IEEE 802.3ad Dynamic link aggregation
Transmit Hash Policy: layer2+3 (2)
MII Status: up
MII Polling Interval (ms): 100
Up Delay (ms): 0
Down Delay (ms): 0

802.3ad info
LACP rate: slow
Aggregator selection policy (ad_select): stable
Active Aggregator Info:
        Aggregator ID: 1
        Number of ports: 2
        Actor Key: 17
        Partner Key: 4
        Partner Mac Address: b4:0e:dc:37:4f:a7

Slave Interface: eth0
MII Status: up
Link Failure Count: 1
Permanent HW addr: 00:25:90:38:e5:22
Aggregator ID: 1

Slave Interface: eth1
MII Status: up
Link Failure Count: 0
Permanent HW addr: 00:25:90:38:e5:23
Aggregator ID: 1
```

If you see Bonding Mode: IEEE 802.3ad Dynamic link aggregation, RMN is on.

If you see `Bonding Mode: fault-tolerance` (active-backup), it means that the bonding mode and potentially redundant management networking is disabled.

Use the `configure-cluster` GUI **Configure Redundant Management Network** option to turn on the redundant management network (RMN) system for nodes being discovered going forward.

Set the redundant management networking mode on, as follows:

```
# cadmin --enable-redundant-mgmt-network --node r1lead
```

Set the bonding mode per node, as follows:

```
# cadmin --set-mgmt-bonding --node r1lead 802.3ad
```

You need to reboot the system.

The `/proc/net/bonding/bond0` file, should show the bonding mode with link aggregation configured, as follows:

```
Bonding Mode: IEEE 802.3ad Dynamic link aggregation
```

The number of ports should be the following:

```
Number of ports: 2
```

`2` is the correct value for an RMN configuration. If the number is 1, it mean the trunk has not formed. The likely causes for this are, as follows:

- The Ethernet cable is not connected to top level switch. From the RLC, use the `/sbin/ethtool` on `eth0` and `eth1` to verify the link is present, as follows:

```
r1lead:~ # /sbin/ethtool eth0
Settings for eth0:
        Supported ports: [ TP ]
        Supported link modes:   10baseT/Half 10baseT/Full
                                100baseT/Half 100baseT/Full
                                1000baseT/Full
        Supports auto-negotiation: Yes
        Advertised link modes:  10baseT/Half 10baseT/Full
```

```
                                100baseT/Half 100baseT/Full
                                1000baseT/Full
        Advertised auto-negotiation: Yes
        Speed: 1000Mb/s
        Duplex: Full
        Port: Twisted Pair
        PHYAD: 1
        Transceiver: internal
        Auto-negotiation: on
        Supports Wake-on: umbg
        Wake-on: g
        Current message level: 0x00000003 (3)
        Link detected: yes
```

- The Ethernet cable is connected, but linking is wrong. When the `/sbin/ethtool` command output shows the link speed as 100mbit due to a bad cable the trunk leg is rejected.
- The top level Ethernet switch misconfigured: Perhaps the `switchconfig` tool did not get this port configured properly. You can either log in to the switch to try to diagnose, or try the following procedure:

  1. Find the MAC address of the `r1lead` bond interface, as follows:

```
r1lead:~ # ifconfig bond0
bond0     Link encap:Ethernet  HWaddr 00:25:90:38:E5:22
          inet addr:172.23.0.7  Bcast:172.23.255.255  Mask:255.255.0.0
          inet6 addr: fe80::225:90ff:fe38:e522/64 Scope:Link
          UP BROADCAST RUNNING MASTER MULTICAST  MTU:1500  Metric:1
          RX packets:286749167 errors:0 dropped:0 overruns:0 frame:0
          TX packets:328574062 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:38868281915 (37067.6 Mb)  TX bytes:153036792319 (145947.2 Mb)
```

  2. From the system admin controller (SAC), run the `switchconfig list --switches mgmtsw0` command to list the MAC addresses trunks from the switches, as follows:

```
sys-admin:~ # switchconfig list --switches mgmtsw0
Current MAC/port configuration:

Switch Identifier: mgmtsw0   IP Address: 172.23.0.6
```

```
      MAC          Port  Trunk  default-VLAN  allowed-VLANs
-------------------------------------------------------------
00-25-90-3F-16-C4   1/6                   1           1(u)
00-30-48-F7-84-65  1/48                   1           1(u)
00-25-90-38-E5-22   1/5      1            1  1(u),  101(t)
00-25-90-38-E5-23   1/5      1            1  1(u),  101(t)
00-25-90-38-E5-22   1/5      1          101  1(u),  101(t)
00-25-90-38-85-BC   1/7      2            1           1(u)
00-25-90-38-85-BD   1/7      2            1           1(u)
                          ...
```

If the RLC `r1lead` bond interface `MAC` address shows up in the `Port` column and not the `Trunk` column, the switch is not configured correctly.

3. To properly configure the switch, from the SAC node, perform a command similar to the following:

```
# switchconfig set -s mgmtsw0 -v num=1 -v num=101,tag=tagged -b lacp -d 1 -m 00:25:90:38:E5:30
```

This replaces 101 with the proper VLAN number. 101 for rack group 1, 102 for rack group 2, and so on.

4. `ssh` onto the `r1lead` and verify that the RLC shows `Number of ports:  2`.

Note: This procedure only applies to RLCs.

`cimage --push-rack` Pushes Too Many (or Too Few) Expansions

When you perform `cimage --push-rack` (or when `blademond` calls `discovery-rack`), it creates read/write expansions for each compute node.

Use the `configure-cluster` GUI **Configure Default Max Rack IRU Setting** option to set the default number of individual rack units (IRUs), supported by a rack leader controller (RLC). Set this value to the number of CMCs that will be served by each RLC. The default is 8. When you change it, it only impacts node discoveries in the future.

You can change the setting per-node with the `cadmin` command, as follows:

```
sys-admin:~ # cadmin --set-max-rack-irus --node r1lead 8
```

Cannot `ping` the CMCs from the Rack Leader Controller (RLC)

If this is an RLC with a brand new, never-before-discovered top level switch (or set of switches), the `cmcdetectd` daemon will see CMCs asking for IP addresses on the HEAD network. It configures the top level switch(es) so that the CMCs are on the appropriate rack VLAN. Make sure `cmcdetectd` is running, restart if needed.

You can diagnose this some by running the `tcpdump` command looking for DHCP requests. The requests should be seen on the RLC and not the system admin controller (SAC). For example, type the following command from `r1lead`:

```
# /usr/sbin/tcpdump -i bond0 -s600 -nn -vv -e -t -l -p broadcast and src port 68 and dst port 67
tcpdump: listening on bond0, link-type EN10MB (Ethernet), capture size 600 bytes
00:25:90:3f:16:c4 > ff:ff:ff:ff:ff:ff, ethertype IPv4 (0x0800), length 590: (tos 0x0, ttl 64, id 0, offset 0, \
flags [none], proto UDP (17), length 576) 0.0.0.0.68 > 255.255.255.255.67:
[udp sum ok] BOOTP/DHCP, Request from 00:25:90:3f:16:c4, length 548, xid 0x8b8d332a, Flags [none] (0x0000)
          Client-IP 172.24.0.2
          Client-Ethernet-Address 00:25:90:3f:16:c4
          Vendor-rfc1048 Extensions
            Magic Cookie 0x63825363
            DHCP-Message Option 53, length 1: Request
                        ...
```

If the switch was previously discovered but you are reinstalling the system or discovering a new root slot, `cmcdetectd` will not detect any CMC DHCP requests on HEAD. In this case, you need to be sure to run `configure-cluster` and set **Configure Switch Management Network** to yes. Note that changing `configure-cluster` only takes effect for nodes discovered in the future. If you have an existing RLC already discovered, you will need to run a command like the following:

```
# cadmin --enable-switch-mgmt-network --node r1lead
```

After rebooting the RLC, make sure that the `ifconfig` command shows `vlan101` as an interface and not `vlan1` or `vlan2` interfaces, as follows:

```
r1lead:~ # ifconfig
                                    ...
vlan101   Link encap:Ethernet  HWaddr 00:25:90:38:E5:22
          inet addr:192.168.160.1  Bcast:192.168.160.255  Mask:255.255.255.0
          inet6 addr: fe80::225:90ff:fe38:e522/64 Scope:Link
          UP BROADCAST RUNNING MASTER MULTICAST  MTU:1500  Metric:1
          RX packets:290550897 errors:0 dropped:0 overruns:0 frame:0
```

```
          TX packets:268387414 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:30869741447 (29439.6 Mb)  TX bytes:120262245830 (114691.0 Mb)
```

Confirm `dhcpd` is running on the RLC. If `dhcpd` is not running, CMCs will not get their IP addresses. Check for errors starting `dhcpd`. If `blademond` failed to create the `ice.conf dhcpd` configuration file (`/etc/dhcpd.conf.d`) , see "How to Make the `blademond` Daemon Start Over from Scratch" on page 122.

Verify proper CMC configuration. The CMC is configured for its rack number and slot number. If they are not configured correctly, multiple CMCs can be configured the same way resulting in problems. This can also result in the `ice.conf dhcp` configuration file being corrupted. You may need a USB serial cable to fix the CMCs if this is the case.

One troubleshooting approach is to run `tcpdump` on the RLC, as follows:

```
usr/sbin/tcpdump -i bond0 -s600 -nn -vv -e -t -l -p broadcast and src port 68 and dst port 67
```

Watch the DHCP requests over several minutes. If you see the same Client Identifier being requested by more than one MAC address, you are in a situation where the CMCs are not configured correctly.

Verify that the RLC is properly configured in the switch (see "Configure Switches for a Rack Leader Controller (RLC)" on page 127).

See "`r1lead` Configured with `vlan1`/`vlan2` and Not `vlan101`" on page 122.

Confirm the wiring rules. See "Switch Wiring Rules" on page 129.

If you moved some CMCs from one RLC number to another and you already adjusted the rack and slot number in the CMC, The switch likely does not know about the changes. The CMCs are likely "lost" in the wrong VLAN, potentially a VLAN that is no longer in use. For example, if you had the CMCs served by the `r3lead` RLC but decided to decommission `r3lead` and move the CMCs to `r1lead` instead this situation could arise. In this case, the switch must be reconfigured. Use the `switchconfig` command to configure the ports connected to those CMCs for head. The system admin controller (SAC) `cmcdetectd` daemon will move them to the correct ultimate location.

You need to know the MACs of the CMC embeded Linux for this, so perhaps record this when you change the slot/rack number in the CMC. **Hint:** `dbdump` may still have the information depending on how you removed the RLC.

An example command is, as follows:

```
# switchconfig set -v num=1 -b manual -d 1 -m 08:00:69:16:51:49 --switches mgmtsw0
```

If you have more than one management switch, then list them in a comma-separated-list for `--switches`.

In a non-redundant-management configuration (switches not stacked), if the `dhcpd` daemon shows DHCP requests from the CMC but the CMC remains unpingable, it could be that both CMC-0 and CMC-1 are connected and linked. This breaks the wiring rules. When we are **not** wired for redundant management networking, only CMC-0 should be connected.

When **not** wired for redundant management networking (when switches are not stacked), do not connect CMC-1.

`r1lead` Configured with `vlan1`/`vlan2` and Not `vlan101`

See the VLAN and switch management network settings explanation in "Cannot `ping` the CMCs from the Rack Leader Controller (RLC)" on page 120.

How to Make the `blademond` Daemon Start Over from Scratch

From the rack leader controller (RLC), perform the following steps:

1.

   ```
   r1lead:~ # service blademond stop
   ```

2.

   ```
   rm ice.conf dhcpd.conf
   ```

   Remove `etc/dhcpd.conf.d/ice.conf` or `/etc/dhcp/dhcpd.conf.d/ice.conf`

3.

   ```
   r1lead:~ # rm /var/opt/sgi/lib/blademond/slot_map
   ```

4.

   ```
   r1lead:~ # service blademond start
   ```

Log Files

All of the logs in `/var/log` directory. In addition to the `messages` log file and in some cases `dhcpd` file on the rack leader controller (RLC), here are some interesting `/var/log` directory log files:

- `/var/log/discover-rack`

  On the system admin controller (SAC), the `discover-rack` call is facilitated by `blademond` when new nodes are found. This log will often show problems with discovering nodes.

- `var/log/blademond`

  On the RLCs, this shows the `blademond` daemon actions. This includes showing when blade changes are found and it also shows its call to `discover-rack`, and so on. If there are CMC communication issues, they will often be noticed in this log.

- `/var/log/cmcdetectd.log`

  On the SAC, `cmcdetectd` logs its actions as it configures the switches for CMCs in the system. Watch for progress or errors here.

- `/var/log/switchconfig.log`

  On the SAC, there is a `switchconfig` command line tool. This tool is largely used by the `discover` command as nodes are discovered. Its actions are logged in to this log file. If RLC VLANs are not functioning properly, check the `switchconfig` log file.

CMC `slot_map` / `blademond` Debugging Hints

This section describes what to do when the `blademond` daemon cannot find a system blade, as follows:

- Can you `ping` the CMCs? See "Cannot `ping` the CMCs from the Rack Leader Controller (RLC)" on page 120.

- If the CMCs are pingable, verify that they have a valid slot map. If the slot map returned by the CMC is missing entries, then `blademond` cannot function properly. It operates on information passed to it by the CMC. Some commands to run from the rack leader controller (RLC) are, as follows:

  –

  ```
  r1lead:~ # /opt/sgi/lib/dump-cmc-slot-tables
  ```

  This command attempts to dump the slot map from each CMC to your screen.

  –

  ```
  r1lead:~ # echo STATUS | netcat r1i0c 4502
  ```

  This command can query an individual slot map.

  Note: In some software distributions, `netcat` is `nc`.

If the CMCs are pingable and the CMCs have valid slot maps, then you can focus on how `blademond` is functioning.

You can turn on debug mode in the `blademond` daemon by sending it a `SIGUSR1` signal from the RLC, as follows:

```
# kill -USR1 pid
```

To turn debug mode off, send it another `SIGUSR1` signal. You should see a message in the `blademond` log about debug mode being enabled or disabled.

The `blademond` daemon maintains the slot map at `/var/opt/sgi/lib/blademond/slot_map` on the RLCs. This appears as `/var/opt/sgi/lib/blademond/slot_map.`*rack_number* on the system admin controller (SAC).

For a `blademond --help` statement, `ssh` onto the `r1lead` RLC, as follows:

```
[root@admin ~]# ssh r1lead
Last login: Tue Jan 17 13:21:34 2012 from admin
[root@r1lead ~]#
[root@r1lead ~]# /opt/sgi/lib/blademond --help
Usage: blademond [OPTION] ...

Discover CMCs and blades managed by CMCs.
```

```
Note: This daemon normally takes no arguments.
  --help      Print this usage and exit.
  --debug     Enable debug mode (also can be enabled by setting CM_DEBUG)
  --fakecmc   Development only: Discover fake CMCs instead of real ones
  --scan-once Initialize, scan for blades, set blades up. Do not daemonize.
              Do not keep looping - do one pass and exit.
```

ssh Commands to Compute Nodes: ssh Key Failures / Known Hosts

For information, see "Resolving CMC Slot Map Ordering Issues" on page 125.

Compute Node Hosts Seem to Actually be BMCs

For information, see "Resolving CMC Slot Map Ordering Issues" on page 125.

Resolving CMC Slot Map Ordering Issues

The CMC maintains a cache file that records which MACs are BMC-MACs and which are host-MACs. It uses this information, combined with switch port location information in the embedded Broadcom switch, to generate the slot map used by the `blademond` daemon.

In certain situations, such as, a CMC reflash, may remove the cache file but leave CMC power active. In this situation, the CMC does not know which MACs on a given embedded switch port are host and which are BMC and gets the order randomly incorrect. It then caches the incorrect order. To fix this for each CMC, turn the power off with `pfctl`, zero out the MAC cache file, and reset each CMC. Then have `blademond` start over from scratch (see "How to Make the `blademond` Daemon Start Over from Scratch" on page 122). Perform the following steps:

1. `ssh` as root to the rack leader controller (RLC), as follows:

   ```
   sys-admin:~ # ssh r1lead
   Last login: Thu Jan 26 13:57:53 2012 from admin
   r1lead:~ #
   ```

2. Disable the blademond daemon, as follows:

   ```
   r1lead:~ # service blademond off
   ```

3. Turn off IRU power for each CMC using the cpower command, as follows:

   ```
   # PDSH_SSH_ARGS_APPEND="-F /root/.ssh/cmc_config" pdsh -g cmc pfctl off
   ```

4. Zero out the slot map cache file, as follows:

```
# PDSH_SSH_ARGS_APPEND="-F /root/.ssh/cmc_config" pdsh -g cmc cp /dev/null /work/net/broadcom_mac_addr_c
```

5. Reboot the CMC, as follows:

   ```
   # PDSH_SSH_ARGS_APPEND="-F /root/.ssh/cmc_config" pdsh -g cmc reboot
   ```

6. Restart blademond from scratch, see "How to Make the blademond Daemon Start Over from Scratch" on page 122.

In tmpfs Mode, File Has Date in the Future Warnings

If you boot a compute node with tmpfs, part of the process transfers a root tarball using multicast. This tarball is then expanded. If you see hundreds of "`file X has a time in the future`" messages, it likely means your hardware clock is not set to system time properly (see "Ensuring Hardware Clock Has the Correct Time" on page 126).

Ensuring Hardware Clock Has the Correct Time

Some software distributions do not synchronize the system time to the hardware clock as expected. As a result, the hardware clock may not get synchronized with the system time as it should. At shut down, the system time is copied to the hardware clock, but sometimes this does not happen.

To set all the compute node hardware clocks up properly, perform the following:

- Make sure the system admin controller (SAC) and rack leader controller (RLC) have the correct time
- Make sure the SAC and RLCs are synchronized with ntp. A SAC can show a message like the following:

  ```
  ntpd[20489]: synchronized to 128.162.244.1, stratum 2
  ```

- An RLC might show a message like the following:

  ```
  20 Jan 22:54:14 ntpd[16831]: synchronized to 172.23.0.1, stratum 3
  ```

- Maker sure the compute nodes have the correct time. They use `ntp` broadcast packets but still will display this:

  ```
  20 Jan 23:05:16 ntpd[4925]: synchronized to 192.168.159.1, stratum 4
  ```

  You can also use a command like the following and view the output:

  ```
  sys-admin:~ # pdsh -g leader pdsh -g compute date
  ```

- Issue the following command to set the hardware clock to the system clock, as follows:

  ```
  sys-admin:~ # pdsh -g leader pdsh -g hwclock --systohc
  ```

- You can run the `hwclock` without options to confirm the current hardware clock time, as follows:

  ```
  sys-admin:~ # hwclock
  Thu 26 Jan 2012 10:57:27 PM CST  -0.750431 seconds
  ```

Configure Switches for a Rack Leader Controller (RLC)

Normally, as you discover RLCs, `switchconfig` is called automatically and the switch ports associated with the RLC are configured in the special way needed for RLCs, as follows:

- Default VLAN 1
- Accept rack VLAN packets tagged (rack 1 `vlan` is `vlan101`)
- Link Aggregation is the bonding mode between the two ports associated with the RLC

If an RLC is moved in the switch or if `switchconfig` failed during discovery for some reason, you can run `switchconfig` by hand to configure the switch, as follows:

1. Certain switch wires rules must be followed in switch configuration, see "Switch Wiring Rules" on page 129.
2. Make sure all management switches are reachable from the system admin controller (SAC).

3. Find the MAC addresses associated with the RLC interfaces. You can do this by running the following command on the RLC in question:

```
r1lead:~ # cat /proc/net/bonding/bond0
Ethernet Channel Bonding Driver: v3.5.0 (November 4, 2008)

Bonding Mode: IEEE 802.3ad Dynamic link aggregation
Transmit Hash Policy: layer2+3 (2)
MII Status: up
MII Polling Interval (ms): 100
Up Delay (ms): 0
Down Delay (ms): 0

802.3ad info
LACP rate: slow
Aggregator selection policy (ad_select): stable
Active Aggregator Info:
        Aggregator ID: 1
        Number of ports: 2
        Actor Key: 17
        Partner Key: 4
        Partner Mac Address: b4:0e:dc:37:4f:a7

Slave Interface: eth0
MII Status: up
Link Failure Count: 0
Permanent HW addr: 00:25:90:38:e5:22
Aggregator ID: 1

Slave Interface: eth1
MII Status: up
Link Failure Count: 0
Permanent HW addr: 00:25:90:38:e5:23
Aggregator ID: 1
```

Caution: Because bonded interfaces are in play, you cannot get both MAC addresses from using the `ifconfig` command . The `ifconfig` command will show the same MAC address for `eth0` and `eth1` if redundant management networking is enabled.

4. Determine which management switches are present, as follows:

   ```
   r1lead:~ # cnodes -mgmtsw
   mgmtsw0
   ```

5. When you have the list of management switches and the MAC addresses of the RLCs, run a command similar to the following:

```
# switchconfig set --vlan num=1 --vlan num=101,tag=tagged --bonding=802.3ad --default-vlan 1 /
--macs 00:e0:ed:0a:f2:0d,00:e0:ed:0a:f2:0e --switches mgmtsw0,mgmtsw
```

   This replaces the MACs and management switches with the proper ones. It replaces the `101` with the VLAN for the rack, normally "100 + rack number" so rack 1 is 101, rack 2 102.

Switch Wiring Rules

This section is mainly of interest to SGI ICE X system configurations that have a redundant management network setup (stacked pairs of switches) or larger systems that have switch stacks cascaded from the top level switch.

When discovering cascaded switches, it is impossible to know the connected switch ports of all trunks in advance. So when discovering cascaded switches, you can only start with one cable for discovering, then add the second one later on.

When trunks are configured, it is often hard to find the MAC address of both legs of the trunk. This is because the trunked connection just uses one MAC for the connection. Therefore, you need to rely on rules that infer the second port's connection based on the first port.

Some simple wiring rules are, as follows:

- In a redundant management network (RMN) configuration, when connecting system admin controllers (SACs), rack leader controllers (RLCs), service nodes,

and CMCs, you must always use the same port number for the same node in both switches in the stack. In other words:

  - If you connect `r1lead eth0` to switch A, port 43, then you must connection `r1leadeth1` to switch B, port 43.
  - Likewise, if you connect CMC `r1i0c` CMC-0 port to switch A, port 2, then `r1i0c` CMC-1 port must go to switch B port 2.

- When adding cascaded switch stacks, all switch stacks must cascade from the central switch stack. In other words, there is always only, at most, one switch hop.
- When discovering cascaded switches pairs in an RMN setup, observe the following:
  - If you are connecting switch stack 1, switch A, port 48 to switch stack 2, then you must connect the second trunked connection to stack 2, switch B, port 48.
  - Until the cascaded switch stack is discovered, you must leave one trunk leg unplugged temporarily.
  - The `discover` command will tell you when it is safe to plug in the second leg of the trunk. This avoids circuit loops.

System Admin Controller (SAC) `eth2` Link in the Bond is Down

A problem occasionally occurs, especially in SGI XE270 SACs, where the active-backup or 802.3ad bonded `bond0` interface contains an Ethernet `eth2` interface that is down/not linked. To verify this, perform the following:

- Check the Ethernet port of the add-in card and confirm that it is lit.
- Confirm that the add-in card connection to the management switches is using port "0" with port "1" not connected (so not miswired).
- If you look at `/proc/net/bonding/bond0` file, you can confirm that `eth2` is the link that is down.
- Use the `/sbin/ethtool eth2` command and confirm that the `Link detected:` is no.
- Run the commands `ifconfig up eth3` and the run the `/sbin/ethtool eth3` command to determine if the link detected is `yes`.

In this scenario, it is likely that the `eth2`/`eth3` interfaces have been swapped. Another clue is that if `eth2` (look at `/proc/net/bonding/bond0` since the bond enforces the same MAC address for all bonded members) has a MAC address that is larger than the MAC address of `eth3` (as seen by `ifconfig eth3`).

To correct this situation, edit the `/etc/udev/rules.d/70-persistent-net.rules` file and swap the MACs associated with `eth2` and `eth3` in the file.

When you reboot the system, the SAC comes back up with `eth2` and `eth3` properly ordered.

No InfiniBand Interfaces on Rack Leader Controller (RLC), Service, or Compute Node Images

Note: This section only applies to systems running SLES 11 SP1.

If you find that an RLC, service node, or compute nodes seem to lack an expected InfiniBand `ib0` interface, this is likely caused by Open Fabrics Enterprise Distribution (OFED) packages that are too old.

In addition to the `ib0` network interface not being present, you may observe the following message:

```
Loading kernel module for a network device with CAP_SYS_MODULE (deprecated).
Use CAP_NET_ADMIN and alias netdev-ib0 instead
```

The minimum OFED versions to avoid this problem are, as follows:

- `ofed-kmp-default-1.5.2_2.6.32.46_0.3-0.9.13.1.x86_64.rpm`
- `ofed-1.5.2-0.9.13.1.x86_64.rpm`

SGI also suggests that you use a the following kernel level (or later):
`kernel-default-2.6.32.54-0.3.1.3900.0.PTF.743209.x86_64.rpm`

Find the updated SLES 11 SP1 packages in your local updates mirror, as follows:

```
/data/mirrors/novell/sles/updates/SLES11-SP1-Updates/sle-11-x86_64/rpm/x86_64
```

Troubleshooting a Network Address Translation (NAT) Configuration

Troubleshooting can become very complex. The first steps are to determine that the service node(s) are correctly configured for the house network and can ping the house IP addresses. Good choices are house name servers possibly found in the `/etc/resolv.conf` or `/etc/name.d.conf` files on the system admin controller (SAC). Additionally, the default gateway addresses for the service node may be a good choice. You can use the `netstat -rn` command for this information, as follows:

```
system-1:/ # netstat -rn
Kernel IP routing table
Destination     Gateway         Genmask         Flags   MSS Window  irtt Iface
128.162.244.0   0.0.0.0         255.255.255.0   U         0 0          0 eth0
172.16.0.0      0.0.0.0         255.255.0.0     U         0 0          0 eth1
169.254.0.0     0.0.0.0         255.255.0.0     U         0 0          0 eth0
172.17.0.0      0.0.0.0         255.255.0.0     U         0 0          0 eth1
127.0.0.0       0.0.0.0         255.0.0.0       U         0 0          0 lo
0.0.0.0         128.162.244.1   0.0.0.0         UG        0 0          0 eth0
```

If the `ping` command executed from the service node to the selected IP address gets responses, network monitoring tools such as `tcpdump`(1) should be used. On the service node, monitor the `eth1` interface and simultaneously in a separate session monitor the `ib[01]` interface. You should specify monitoring specific-enough to not have additional noise then attempt execute a `ping` command from the compute node.

Example 5-1 `tcpdump` Command Examples

```
tcpdump -i eth1 ip proto ICMP # Dump ping packets on the public side of service node.
tcpdump -i ib1 ip proto ICMP # Dump ping packets on the IB fabric side of service node.
tcpdump -i eth1 port nfs # Dump NFS traffic on the eth1 side of service node.
tcpdump -i ib1 port nfs # Dump NFS traffic on the eth1 side of service node.
```

If packets do not reach the service nodes respective IB interface, perform the following:

- Check the SAC's compute image configuration of the default route.
- Verify that this image has been pushed to the compute nodes.
- Verify that the compute nodes have booted with this image.

If the packets reach the service nodes IB interface, but do not exit the `eth1` interface, verify the NAT configuration on the service node.

If the packets exit the `eth1` interface, but replies do not return, verify the house network configuration and that IP masquerading is properly configured so that the packets exiting the interface appear to be originating from the service node and not the compute node.

YaST2 Navigation

The following list shows SLES YaST2 navigation key sequences:

Key	Action

`Tab`
`Alt` + `Tab`
`Esc` + `Tab`
`Shift` + `Tab`
	Moves you from label to label or from list to list.

`Ctrl` + `L`	Refreshes the screen.
`Enter`	Starts a module from a selected category, runs an action, or activates a menu item.
`Up arrow`	Changes the category. Selects the next category up.
`Down arrow`	Changes the category. Selects the next category down.
`Right arrow`	Starts a module from the selected category.
`Shift` + `right arrow` `Ctrl` + `A`	Scrolls horizontally to the right. Useful in screens if use of the `left arrow` key would otherwise change the active pane or current selection list.
`Alt` + *letter* `Esc` + *letter*	Selects the label or action that begins with the *letter* you select. Labels and selected fields in the display contain a highlighted *letter*.
Exit	Quits the YaST2 interface.

Virtual Local Area Network (VLAN) Information

This chapter contains the following topics:

- "About VLANs" on page 137
- "VLAN Ethernet Network Configurations" on page 137
- "SGI ICE X IP Address Ranges and VLANs for Management and Application Software" on page 141
- "Component Naming Conventions" on page 143
- "System Control Configuration" on page 145

About VLANs

SGI ICE X hardware components are attached to one or more VLANs. This appendix section contains networking reference material that can be useful if you want to reconfigure or debug an SGI ICE X network problem.

VLAN Ethernet Network Configurations

An SGI ICE X system includes at least two VLANs, one head VLAN and one or more rack VLANs. There is one rack VLAN for each rack in the system.

Figure B-1 on page 138 shows an SGI ICE X system with three VLANs.

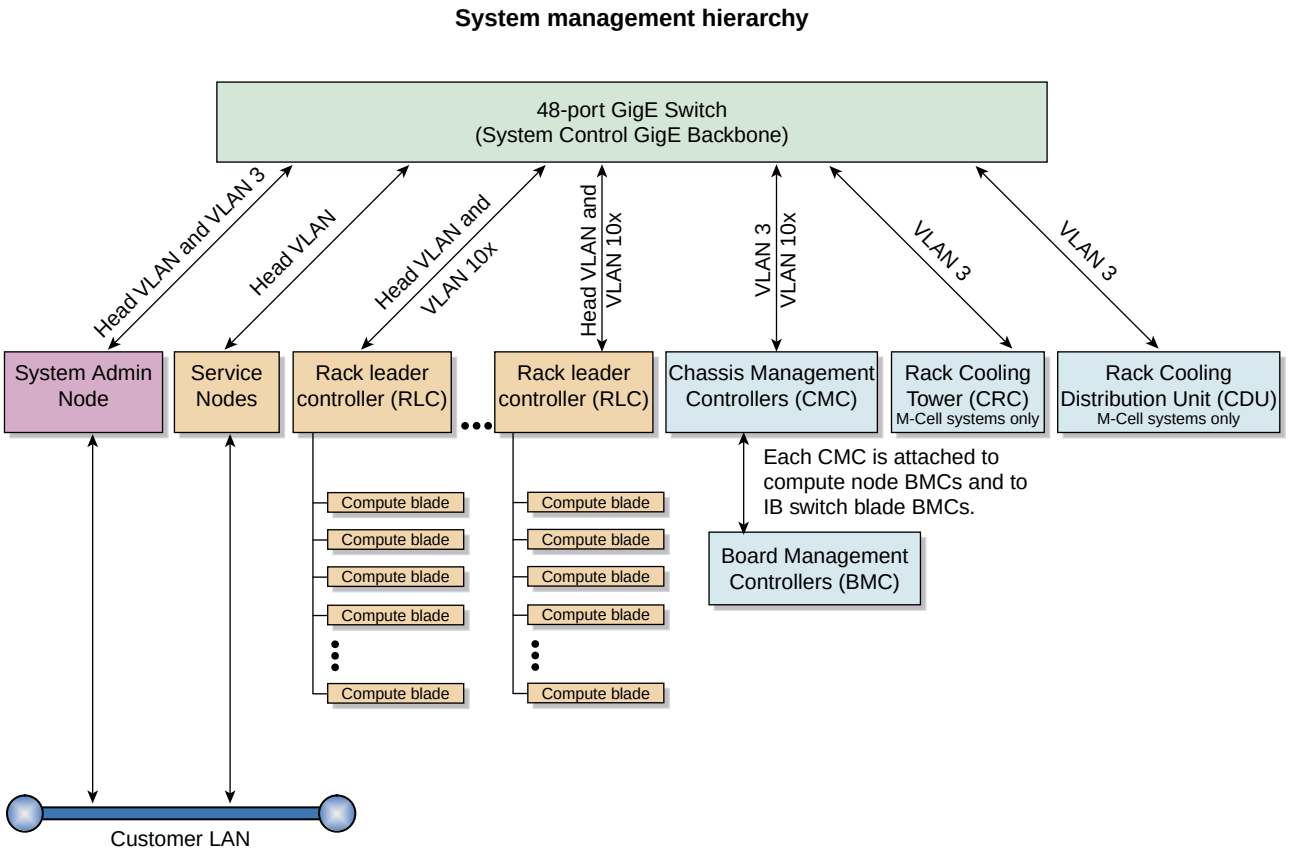

Figure B-1 VLAN Overview

The following topics contain more information about the head and rack VLANs:

- "Head VLAN" on page 138
- "Rack VLANs" on page 139

Head VLAN

The head VLAN is VLAN 1. The head VLAN includes the system admin controller (SAC), all rack leader controllers (RLCs), and all service nodes. The head VLAN is

always configured as untagged. Any untagged packets coming into a SAC, RLC, or service node port are associated with the head VLAN. Internally, the head VLAN is configured as `HEAD_VLAN=1` in the following file:

Figure B-2 on page 139 shows the components that participate in the head network. The figure shows the top level switch.

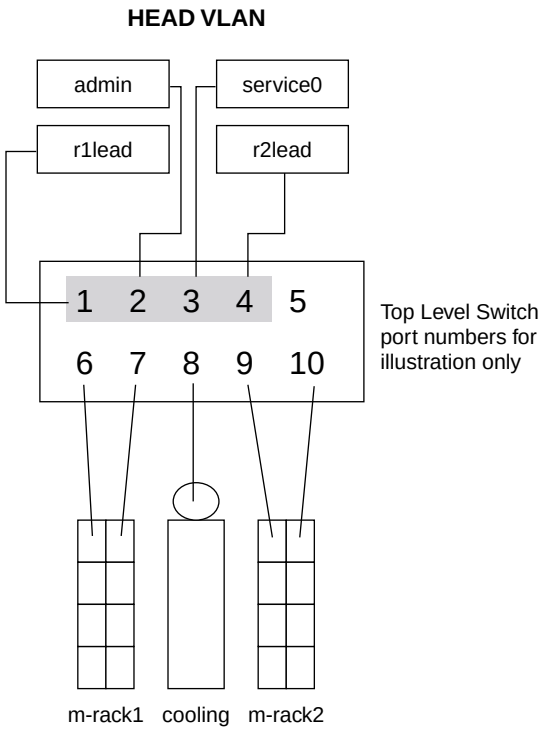

Figure B-2 HEAD VLAN Ethernet Connections

Rack VLANs

There is one rack VLAN for each rack in the system. The VLANs are numbered incrementally. The VLAN for rack 1 is `VLAN_101`. The VLAN for rack 2 is `VLAN_102`. The VLAN for rack 99 is VLAN 199. At the maximum, the VLAN number for rack 1000 is `VLAN_1100`.

The following system components reside on a rack VLAN:

- Rack leader controllers (RLCs). Each RLC resides in both the head VLAN and on its own rack VLAN. The dual residence enables the RLCs to communicate with both the components in the head network and with the compute nodes in their rack.

  The RLC's management-related IP address subnetworks are as follows:

  – One IP address on the head VLAN.

  – The following two IP addresses on the rack VLAN:

    - The BMC network IP address
    - The GBE network IP address

  The RLC Ethernet interface is configured with VLAN tagging. VLAN tagging ties these networks to the rack VLAN.

- Chassis management controllers (CMCs). Each CMC internal switch cascades down from the top level switch. The switch ports for all CMCs in a rack are configured to be on the rack VLAN. Each CMC connects to a port on a top level switch. The port is configured so that all traffic coming in and going to that port travels to the rack VLAN by default. The CMC gets its rack VLAN IP address using DHCP.

- Compute nodes. The backplane connects the compute nodes to the cascaded CMC switch. The compute node's BMC has a shared Ethernet connection with the host interface. Both the compute node and BMC traffic are on the rack VLAN.

  The compute nodes and baseboard management controllers (BMCs) reside on the same rack VLAN. The BMCs have a subnetwork that is separate from the host interfaces.

Figure B-3 on page 141 shows rack VLANs for two racks.

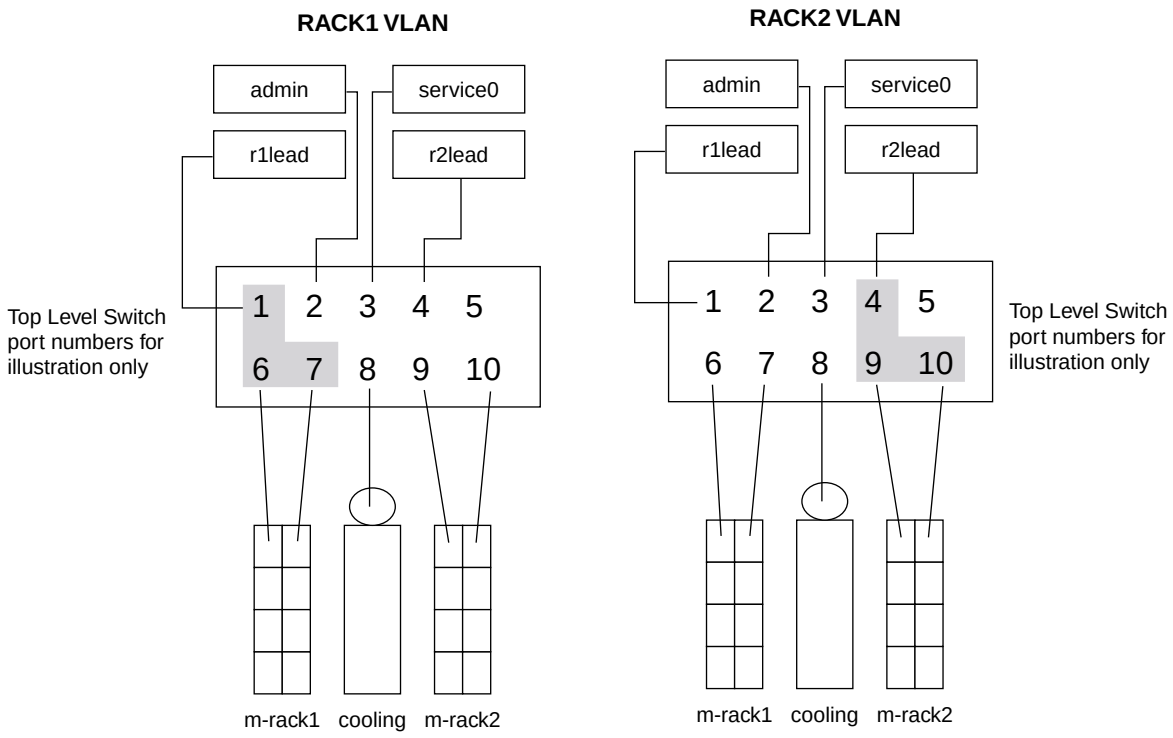

Figure B-3 RACKx VLAN Ethernet Connections

SGI ICE X IP Address Ranges and VLANs for Management and Application Software

Table B-1 on page 141 shows the system-wide SGI ICE X IP address ranges that the cluster management software uses. The HEAD_BMC network is a separate IP subnetwork.

Table B-1 System-wide IP Address Ranges for the Head Network

standard

Subnetwork Name	IP Range	Nodes
HEAD_VLAN=1	172.23.0.0/16	SAC Service nodes RLCs
HEAD_BMC	172.24.0.0/16	SAC BMC Service BMCs RLC BMCs

Table B-2 on page 142 shows the per-rack IP address ranges that the cluster management software uses in the rack VLANs.

Table B-2 Per-rack IP Address Ranges for Cluster Management Software

VLAN Name	Subnetwork Name	IP Range	Components
VLAN_101	gbe	192.168.159.0/24	Rack 1's RLC Rack 1's CMCs Rack 1's compute nodes
VLAN_101	bmc	192.168.160.0/24	Rack 1's RLC's BMC Rack 1's CMCs' BMC Rack 1's compute nodes' BMCs
VLAN_102	gbe	192.168.159.0/24	Rack 2's RLC Rack 2's CMCs Rack 2's compute nodes
VLAN_102	bmc	192.168.160.0/24	Rack 2's RLC's BMC Rack 2's CMCs' BMC Rack 2's compute nodes' BMCs

VLAN Name	Subnetwork Name	IP Range	Components
VLAN_*X*	gbe	192.168.159.0/24	Rack *X*'s RLC Rack *X*'s CMCs Rack *X*'s compute nodes
VLAN_*X*	bmc	192.168.160.0/24	Rack *X*'s RLC's BMC Rack *X*'s CMCs' BMCs Rack *X*'s compute nodes' BMCs

Table B-3 on page 143 shows the system-wide IP address ranges for cluster application software. Only the RLCs that provide InfiniBand subnetwork services need to connect.

Table B-3 Application Software System-wide IP Address Ranges

VLAN Name	Subnetwork Name	IP Range	Nodes
IB0	ib-0	10.148.0.0/16	Service nodes Some RLCs Compute nodes
IB1	ib-0	10.149.0.0/16	Service nodes Some RLCs Compute blades

Component Naming Conventions

The SGI ICE X commands enable you to run commands and perform some procedures on only one component or on a range of similar components. Addressing methods differ depending on the component, the VLAN (or VLANs) in which the component resides, and whether or not the component has an IP address that is externally available.

The topics that follow use the following terms:

- Component. The name of the component that you typically use in speech or in writing. For example: SAC, RLC, and so on.
- Node name. The system-wide unique identifier for the component.

Table B-4 on page 144 explains how to specify components when you run administrative and user commands. *x* is always an integer number.

Table B-4 Naming Conventions

Component	Node name	Examples

SAC	Site-defined hostname	`icex1`
		`mysiteicex`
		`sleet`

RLC	`r`*x*`lead`	`r0lead`, the RLC on the first rack `r1lead`, the RLC on the second rack
RLC BMC	`r`*x*`lead-bmc`	`r1lead-bmc`, the BMC on the second rack
Service node	`service`*x*	`service0`, the first service node `service3`, the fourth service node
Service node BMC	`service`*x*`-bmc`	`service1-bmc`, the BMC on the second service node
InfiniBand switch	`ibswitch`*x*`-bmc`	`ibswitch1-bmc`, the BMC on the second InfiniBand switch
Compute node	`r`*x*`i`*x*`n`*x*	`r1i3n10`, the compute node on the second rack, in the fourth IRU, in position 10
Compute node BMC	`r`*x*`i`*x*`n`*x*`-bmc`	`r1i3n10-bmc`, the BMC on the second rack, in the fourth IRU, in position 10
CMC	`r`*x*`i`*x*`c`	`r1i1c`, the CMC for the second rack, in the second IRU

System Control Configuration

The system control network for an SGI ICE X system can be configured in one of the following ways:

- A redundant management network configuration. This is the default. In a redundant management network configuration, the number of GigE switches in the system control network is doubled. A redundant management network also includes the following:
  - The GigE switches are stacked (using stacking cables).
  - The links from the CMCs are doubled.
  - Links from the SAC, RLCs, and most service nodes are doubled. BMC connections are not doubled. Certain failures can cause temporary inaccessibility to the BMCs, but the host interfaces remain accessible.

  Figure B-4 on page 146 shows the switches in a redundant management network configuration.
- A nonredundant management network configuration. In the nonredundant configuration, a single GigE fabric has a single connection to the SAC, RLCs, and CMCs. Figure B-5 on page 147 shows the switches in a nonredundant management network configuration.

Figure B-4 on page 146 shows a redundant management network cascaded switch configuration.

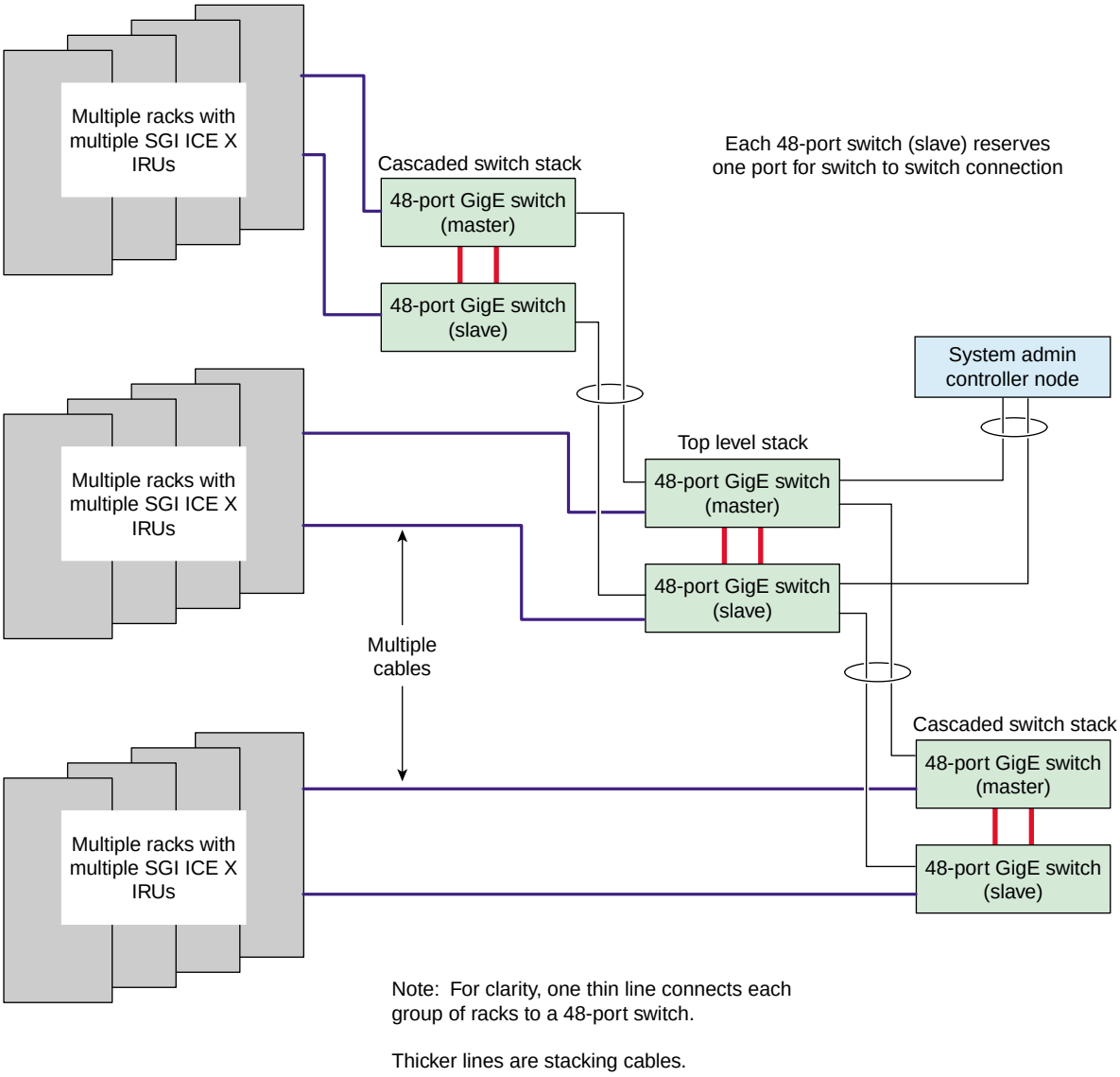

Figure B-4 Redundant Cascaded Switch Configuration

Figure B-5 on page 147 shows a nonredundant management network cascaded switch configuration.

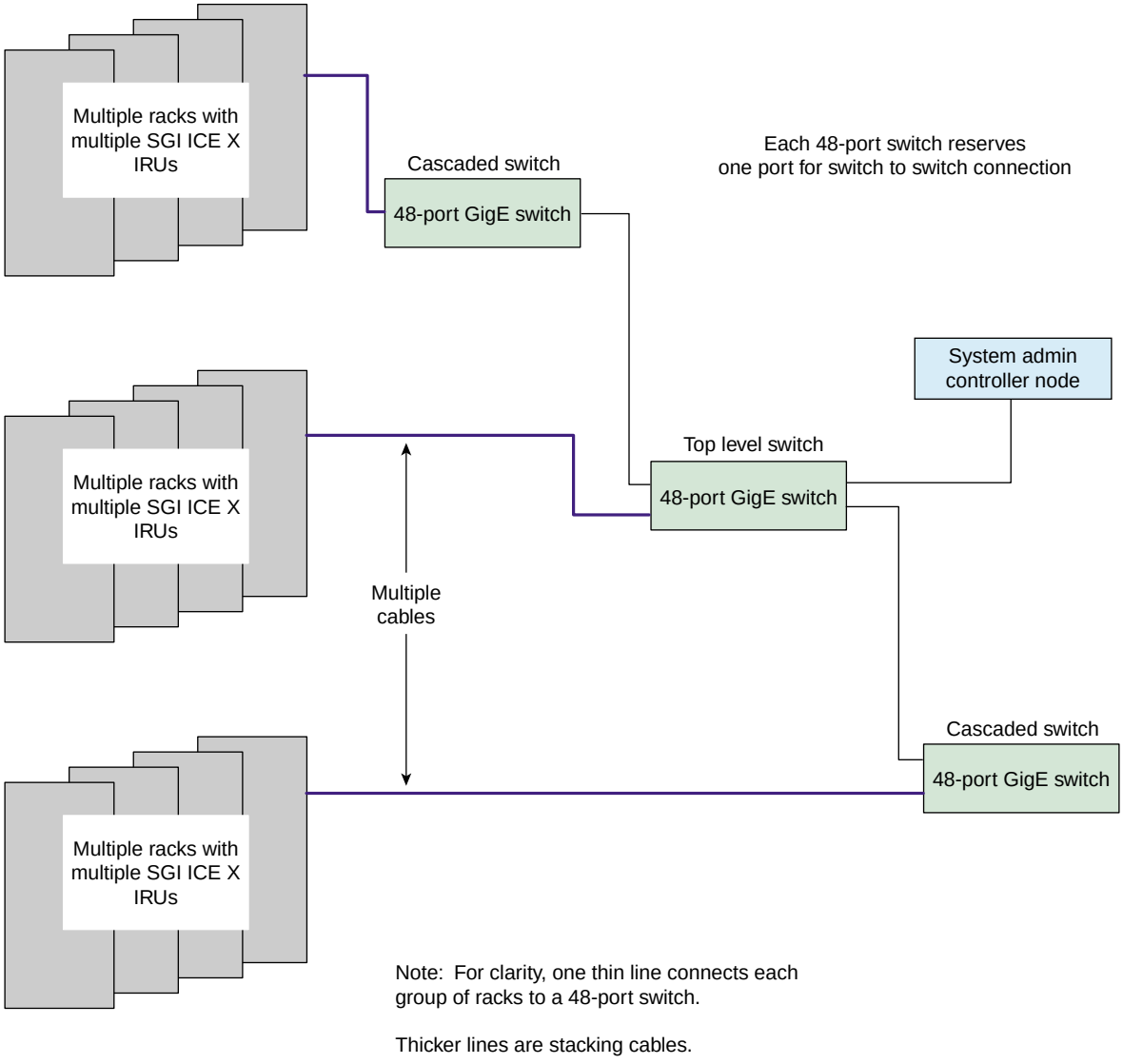

Figure B-5 Nonredundant Cascaded Switch Network Configuration

For diagrams that show both redundant and nonredundant management network wiring, see the chassis manager interconnect diagrams in the *SGI ICE X System Hardware User Guide*.

MCell Network IP Addresses

If you need to troubleshoot the MCell cooling equipment on an SGI ICE X system, you need to know the IP addresses of the cooling rack controllers (CRCs) and cooling distribution units (CDUs) so that you can type a `ping`(8) command to the component. Each piece of equipment bears a label with its equipment number.

For CRCs, the IP address is 172.26.128.*number*.

For CDUs, the IP address is 172.26.144.*number*.

Table C-1 on page 149 shows the *number* at the end of the IP address for CRCs and CDUs.

Table C-1 MCell Network Associations

Rack	Cooling Rack Controllers (CRCs)	Cooling Distribution Unit (CDUs)
1	1	1
2	1	1
3	2	1
4	2	1
5	3	2
6	3	2
7	4	2
8	4	2
9	5	3
10	5	3
11	6	3
12	6	3
13	7	4
14	7	4

Rack	Cooling Rack Controllers (CRCs)	Cooling Distribution Unit (CDUs)
15	8	4
16	8	4
17	9	5
18	9	5
19	10	5
20	10	5
21	11	6
22	11	6
23	12	6
24	12	6
25	13	7
26	13	7
27	14	7
28	14	7
29	15	8
30	15	8
31	16	8
32	16	8
33	17	9
34	17	9
35	18	9
36	18	9
37	19	10
38	19	10
39	20	10
40	20	10

Rack	Cooling Rack Controllers (CRCs)	Cooling Distribution Unit (CDUs)
41	21	11
42	21	11
43	22	11
44	22	11
45	23	12
46	23	12
47	24	12
48	24	12
49	25	13
50	25	13
51	26	13
52	26	13
53	27	14
54	27	14
55	28	14
56	28	14
57	29	15
58	29	15
59	30	15
60	30	15
61	31	16
62	31	16
63	32	16
64	32	16
65	33	17
66	33	17

Rack	Cooling Rack Controllers (CRCs)	Cooling Distribution Unit (CDUs)
67	34	17
68	34	17
69	35	18
70	35	18
71	36	18
72	36	18
73	37	19
74	37	19
75	38	19
76	38	19
77	39	20
78	39	20
79	40	20
80	40	20
81	41	21
82	41	21
83	42	21
84	42	21
85	43	22
86	43	22
87	44	22
88	44	22
89	45	23
90	45	23
91	46	23
92	46	23

Rack	Cooling Rack Controllers (CRCs)	Cooling Distribution Unit (CDUs)
93	47	24
94	47	24
95	48	24
96	48	24
97	49	25
98	49	25
99	50	25
100	50	25

Index

C

Configure backup DNS server, 68
configuring the service node
  for NAT, 76
  for NIS for the house network, 88
creating user accounts, 106

D

DHCP option code, 65
dhcp options
  changing, 65
disabling InfiniBand switch monitoring, 69

H

home directories on NAS, 96

I

InfiniBand configuration, 69
  disabling InfiniBand switch monitoring, 69
initial configuration of a RHEL 6 SAC, 37

N

NAS home directories, 96
NAT
  configuring the service node, 76
network interface naming conventions, 137
networks
  network interface naming conventions, 137
NIS
  service node configuration for the house network, 88

O

overview, 1

S

setting up a NIS Server, 96
setting up an NFS home server on a service node, 80
system management software, 1
system overview, 1

T

troubleshooting
  frequently asked questions, 111
  initial system setup, 111
troubleshooting service node configuration for NAT, 132

U

user accounts
  creating, 106