SGI® Management Center™ Installation and Configuration Guide for Clusters

007–6359–001

The SGI Tempo systems management software stack 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

restrictions: 1. Neither the authors nor Purdue University are responsible for any consequences of the use of this software. 2. The origin of this software must not be misrepresented, either by explicit claim or by omission. Credit to the authors and Purdue University must appear in documentation and sources. 3. Altered versions must be plainly marked as such, and must not be misrepresented as being the original software. 4. This notice may not be removed or altered.

Copyright (c) 1990 The Ohio State University. All rights reserved. Redistribution and use in source and binary forms are permitted provided that: (1) source distributions retain this entire copyright notice and comment, and (2) distributions including binaries display the following acknowledgment: ''This product includes software developed by The Ohio State University and its contributors'' in the documentation or other materials provided with the distribution and in all advertising materials mentioning features or use of this software. Neither the name of the University 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 ''AS IS'' AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, WITHOUT LIMITATION, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. 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 that copyright notice and this permission notice appear in supporting documentation. This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.

pysqlite:

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 that copyright notice and this permission notice appear in supporting documentation.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.

LIMITED RIGHTS LEGEND

The software described in this document is "commercial computer software" provided with restricted rights (except as to included open/free source) as specified in the FAR 52.227-19 and/or the DFAR 227.7202, or successive sections. Use beyond license provisions is a violation of worldwide intellectual property laws, treaties and conventions. This document is provided with limited rights as defined in 52.227-14.

TRADEMARKS AND ATTRIBUTIONS

Altix, ICE, Performance Co-Pilot, Rackable, SGI, the SGI logo, and Supportfolio are trademarks or registered trademarks of Silicon Graphics International Corp. or its subsidiaries in the United States and other countries.

Altair is a registered trademark and PBS Professional is a trademark of Altair Engineering, Inc. Intel, Itanium, Phi, and Xeon are trademarks or registered trademarks of Intel Corporation. InfiniBand is a trademark of the InfiniBand Trade Association. InfiniScale is a registered trademark of Mellanox Technologies. Linux is a registered trademark of Linus Torvalds. LSI Logic and MegaRAID are registered trademarks of the LSI Logic Corporation. CentOS, Red Hat, and Red Hat Enterprise Linux are registered trademarks of Red Hat, Inc., in the United States and other countries. SLES, SUSE, and YAST are registered trademarks of SUSE LLC in the United States and other countries.

All other trademarks mentioned herein are the property of their respective owners.

New Features

This new manual describes the SGI Management Center 3.0 release, which SGI supports on SGI® ICE™ X cluster platforms and on SGI® Rackable® cluster platforms.

Record of Revision

Version	Description
001	November 2014 Original publication. This revision supports the SGI Management Center 3.0 release and the SGI Foundation Software 2.11 release.

Contents

About This Guide

This guide is a reference document for people who install and configure SGI® ICE™ cluster computer systems and SGI Rackable® cluster computer systems. SGI Rackable clusters are sometimes called *flat cluster* computer systems because of their nonhierarchical structure. This manual describes how to perform general system discovery, installation, configuration, and operations.

Related Publications

The SGI Foundation Software release notes and the SGI Performance Suite release notes contain information about the specific software packages provided in those products. The release notes also list SGI publications that provide information about the products. The release notes are available in the following locations:

- Online at Supportfolio. After you log into Supportfolio, you can access the release notes. The SGI Foundation Software release notes are posted to the following website:

  https://support.sgi.com/content_request/194480/index.html

  The SGI Performance Suite release notes are posted to the following website:

  https://support.sgi.com/content_request/786853/index.html

  Note: You must sign into Supportfolio, at https://support.sgi.com/login, in order for the preceding links to work.

- On the product media. The release notes reside in a text file in the `/docs` directory on the product media. For example, `/docs/SGI-MPI-1.`*x*`-readme.txt`.
- On the system. After installation, the release notes and other product documentation reside in the `/usr/share/doc/packages/`*product* directory.

All SGI publications are available on the Technical Publications Library at the following website:

http://docs.sgi.com

The following documentation might be useful to you:

- *SGI Management Center Administration Guide for Clusters*, publication 007–6358–*xxx*

  This manual explains how to use manage an SGI ICE cluster or an SGI Rackable cluster.

  Note: The following documentation is obsolete as of this release:

  – *SGI ICE X Installation and Configuration Guide*, publication 007–5917–*xxx*
  – *SGI ICE X Administration Guide*, publication 007–5918–*xxx*
  – *SGI Management Center (SMC) System Administrator Guide*, publication 007–5642–*xxx*
  – *SGI Management Center (SMC) Installation and Configuration*, publication 007–5643–*xxx*

- *Message Passing Toolkit (MPT) User's Guide*

  Describes industry-standard message passing protocol optimized for SGI computers. This manual describes how to tune the run-time environment to improve the performance of an MPI message passing application on SGI computers. None of these ways involve application code changes.

- *MPInside Reference Guide*

  Documents the SGI MPInside MPI profiling tool.

- SGI hardware documentation.

  SGI creates hardware manuals that are specific to each product line. The hardware documentation typically includes a system architecture overview and describes the major components. It also provides the standard procedures for powering on and powering off the system, basic troubleshooting information, and important safety and regulatory specifications.

  The following procedure explains how to retrieve a list of hardware manuals for your system.

Procedure 0-1 To retrieve hardware documentation

1. Type the following URL into the address bar of your browser:

   `docs.sgi.com`

2. In the search box on the Techpubs Library, narrow your search as follows:

   – In the **search** field, type the model of your SGI system.

For example, type one of the following: **"UV 2000"**, **"ICE X"**, **Rackable**.

Remember to enclose hardware model names in quotation marks (" ") if the hardware model name includes a space character.

- Check **Search only titles**.
- Check **Show only 1 hit/book**.
- Click **search**.

- In addition to SGI documentation, the following documentation from other sources might interest you:
  - SUSE documentation for SLES 11
  - Red Hat documentation for Red Hat Linux Enterprise Server 6 (RHEL 6) and CentOS 6.5
  - 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 the following website:

  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.

- 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 either 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:

  http://www.sgi.com/support/supportcenters.html

SGI values your comments and will respond to them promptly.

System Software Overview

This chapter includes the following topics:

About SGI Cluster Computer Systems

Figure 1-1 on page 2 shows an SGI Rackable cluster and an SGI ICE X cluster. Each type of cluster includes an admin node and flat compute nodes. As the figure shows, the admin node and the flat compute nodes attach directly to the management network. Admin nodes are sometimes referred to as *system admin controller* (SAC) nodes.

In an SGI ICE X cluster, the SGI ICE compute nodes are configured in a hierarchical way, under a rack leader controller (RLC). In the SGI ICE X cluster, it is the RLC that attaches to the management network, not the compute nodes.

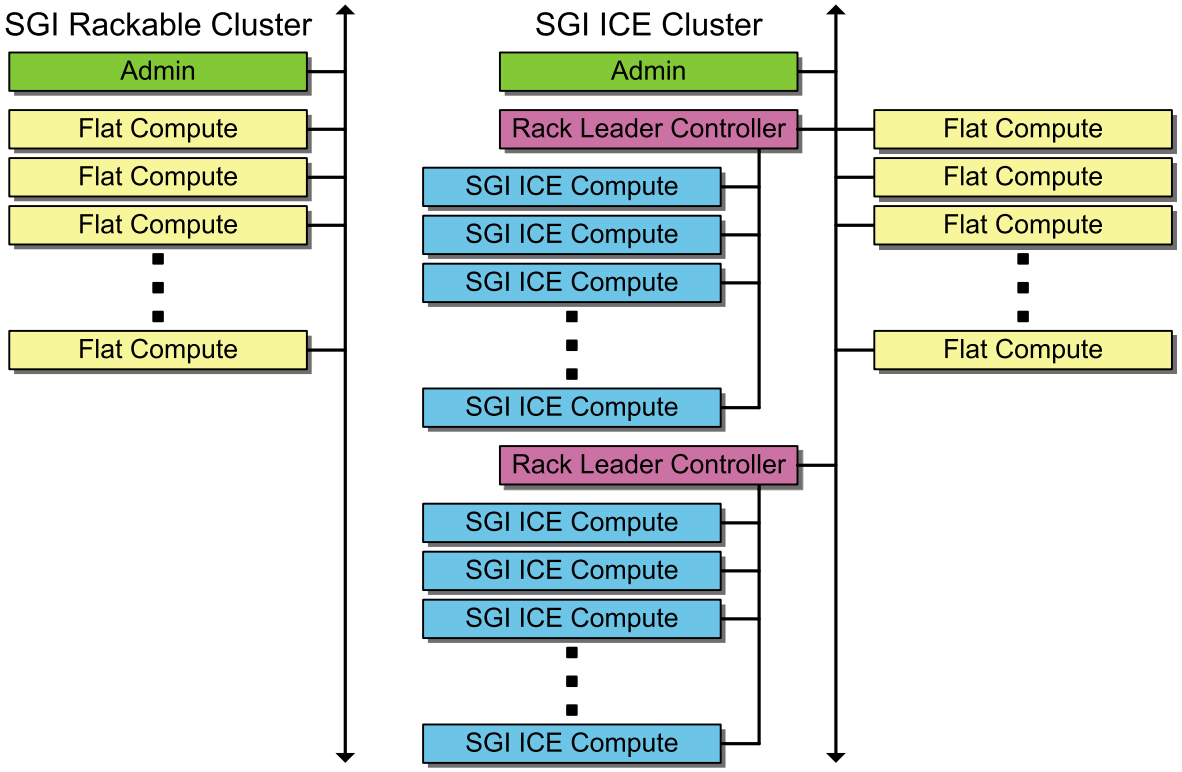

Figure 1-1 SGI Rackable Clusters and SGI ICE Clusters

The following topics describe the SGI cluster systems:

- "About SGI Rackable Clusters" on page 2
- "About SGI ICE X Clusters" on page 4
- "About High Availability Nodes in SGI Clusters" on page 4

About SGI Rackable Clusters

The nodes in an SGI Rackable cluster have the following roles:

- The admin node is the cluster's administrative node. This is the node from which you install software and manage the cluster. SGI Management Center (SMC)

software resides on the admin node. SMC enables you to install, provision, configure, and manage the SGI Rackable cluster computing system.

Each cluster has admin node. The admin node hosts the original, factory-installed copies of the software images for each component. System administrators log into the admin node to run system management commands, to modify component images, and to perform system-wide operations.

The SMC software distribution includes the master system image for the admin node. During the installation and configuration process, the installation software creates the master system images for the other components in the cluster. As you customize the system for your site, you modify the component-specific system images on the admin node and push the updated images to the other nodes.

- The compute nodes in an SGI Rackable cluster are called *flat compute nodes* because they are not configured in a hierarchical structure. An admin node can manage thousands of compute nodes, depending on the cluster's workload. The compute nodes all receive a hostname and an IP address during the configuration process. You can configure flat compute nodes with one or more of the following types of user 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 workload schedulers such as Altair's PBS Professional, Adaptive Computing's Moab, SLURM, or TORQUE.
  - I/O gateway. On a small system, you can combine the I/O gateway, login services, and batch scheduling on the same compute node.

    The I/O gateway services connect the cluster to your site network. You can configure one or more of the following protocols on the node: network file system (NFS), network address translation (NAT), or network information service (NIS).
  - Storage. A compute node with storage 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.

SGI recommends the following login practices:

- Only the system administrator should be able to log into the admin node. SGI recommends that sites prohibit end-user access to the admin node.
- User services can be configured on the compute nodes, and end users can have access to these compute nodes.

About SGI ICE X Clusters

On an SGI ICE X cluster, the admin node and the flat compute nodes have the same characteristics that they have in an SGI Rackable cluster. That is, an admin node can support flat compute nodes configured for services such as logging in, batch computing, I/O, gateway, OSS, MDS, or storage. SGI recommends that in an SGI ICE cluster, you install user services on the flat compute nodes. For more information about the flat compute nodes, see the following:

"About SGI Rackable Clusters" on page 2

The hierarchical structure of the RLCs and the SGI ICE compute nodes is unique to the SGI ICE X cluster. The hierarchical design enables these computing systems to be provisioned quickly. Master software images for each type of node in the cluster reside on the admin node. When an SGI ICE X system is configured, the admin node pushes the software images to the flat compute nodes and to the RLCs. Each RLC pushes the compute node images to the SGI ICE X compute nodes that reside in its rack.

An SGI ICE X admin node can support many RLCs, each of which can manage hundreds of SGI ICE compute nodes. The characteristics of the RLCs and SGI ICE compute nodes are as follows:

- The RLC's role is to manage a set of SGI ICE compute nodes in a rack. SGI ICE clusters have at least one RLC.
- The SGI ICE compute nodes are simplified, typically diskless compute nodes that reside in a rack with an RLC. These compute nodes require an RLC for services, infrastructure, and support. An RLC can manage up to 288 SCI ICE compute nodes in a rack. The exact number of SGI ICE compute nodes that an RLC can manage depends on the specific model number of your SGI ICE cluster.

About High Availability Nodes in SGI Clusters

If you have an SGI Rackable or SGI ICE cluster, you can configure the admin node for high availability (HA) operations. An HA admin node consists of two physical nodes dedicated to the admin node role and one virtual machine (VM). The VM hosts the

functioning admin node, and the VM resides on one host at a time. The host upon which the VM resides is the *active node*. The other node is the *passive node*.

If you have an SGI ICE X cluster, you can also configure the RLC for HA operations. In this case, two physical nodes and one virtual machine are dedicated to the RLC role, and the RLC VM resides on one of the RLC nodes at a time.

SGI Cluster System Node Images

Figure 1-2 on page 6 shows two example cluster systems. The figure on the left is an SGI Rackable cluster, and the figure on the right is an SGI ICE cluster that also includes some flat compute nodes.

Figure 1-2 Cluster Nodes and Software Image Names

Figure 1-2 on page 6 shows the image names for the software images on each node. As the example shows, you can configure nodes with different operating system images within the same cluster. On an SGI ICE cluster, the admin node, RLC nodes, and SGI ICE compute nodes must be installed with the same operating system. The image names shown in the figure are the default names for the factory-installed system images. The system images for the cluster nodes are unique to each type of node.

Table 1-1 on page 7 lists the nodes in each type of SGI cluster and shows the images that reside on each node. The table shows the default, factory-given name for each image.

Table 1-1 Cluster Nodes and Software Image Names

SGI Rackable Nodes	SGI ICE Nodes	Node System Image Name
Admin node	Admin node	—Not applicable— Installation of the admin node is facilitated by the admin install DVD (`sgi-mgmtsuite-install`).
Flat compute nodes	Flat compute nodes	*os_name* For example, `centos6.5`
—Not applicable—	Rack leader controller (RLC)	`lead-`*os_name* For example, `lead-sles11sp3`
—Not applicable—	SGI ICE compute nodes	`ice-`*os_name* For example, `ice-rhel6.6`

If you modify the image to include site-specific software, it is typical to copy the image, update the copy, and give the modified image a new name. Copying is also referred to as *cloning* in the cluster documentation. By using a new image name for your changes, you always able to refer back to the original image.

Whenever you add or modify the software on a node, you can use the SGI Management Center software's version control software to manage multiple versions of each node's software. The version control system facilitates the following:

- Storage. You can have many versions of each individual software image, and each version is easily retrieved.

- Experimentation. Each software image is tagged with a version number, so you can easily enable and disable specific versions of the software images.

Operating System Support

The SGI cluster computer systems support the following operating systems:

- Red Hat Enterprise Linux (RHEL)
- SLES
- CentOS

Note: In SGI documentation, you can assume that feature descriptions for RHEL platforms also pertain to CentOS platforms unless otherwise noted.

SGI Cluster Networks

The following topics explain the SGI cluster networks:

- "SGI Rackable Networks" on page 8
- "SGI ICE X Networks" on page 12

SGI Rackable Networks

Figure 1-3 on page 9 is a logical representation of the SGI ICE Ethernet networks.

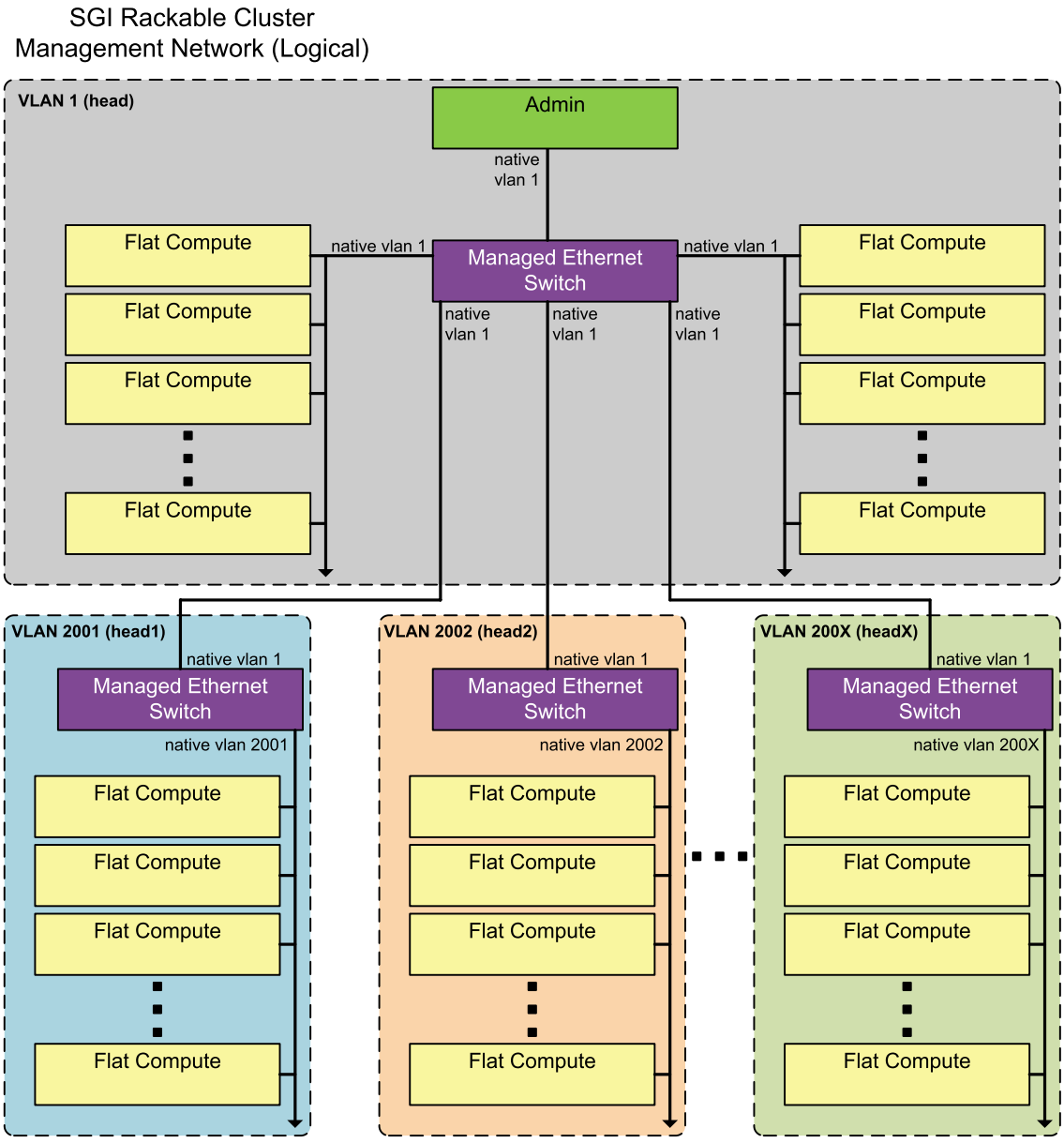

Figure 1-3 SGI Rackable Networks

Figure 1-4 on page 11 shows the logical networks of an SGI Rackable cluster.

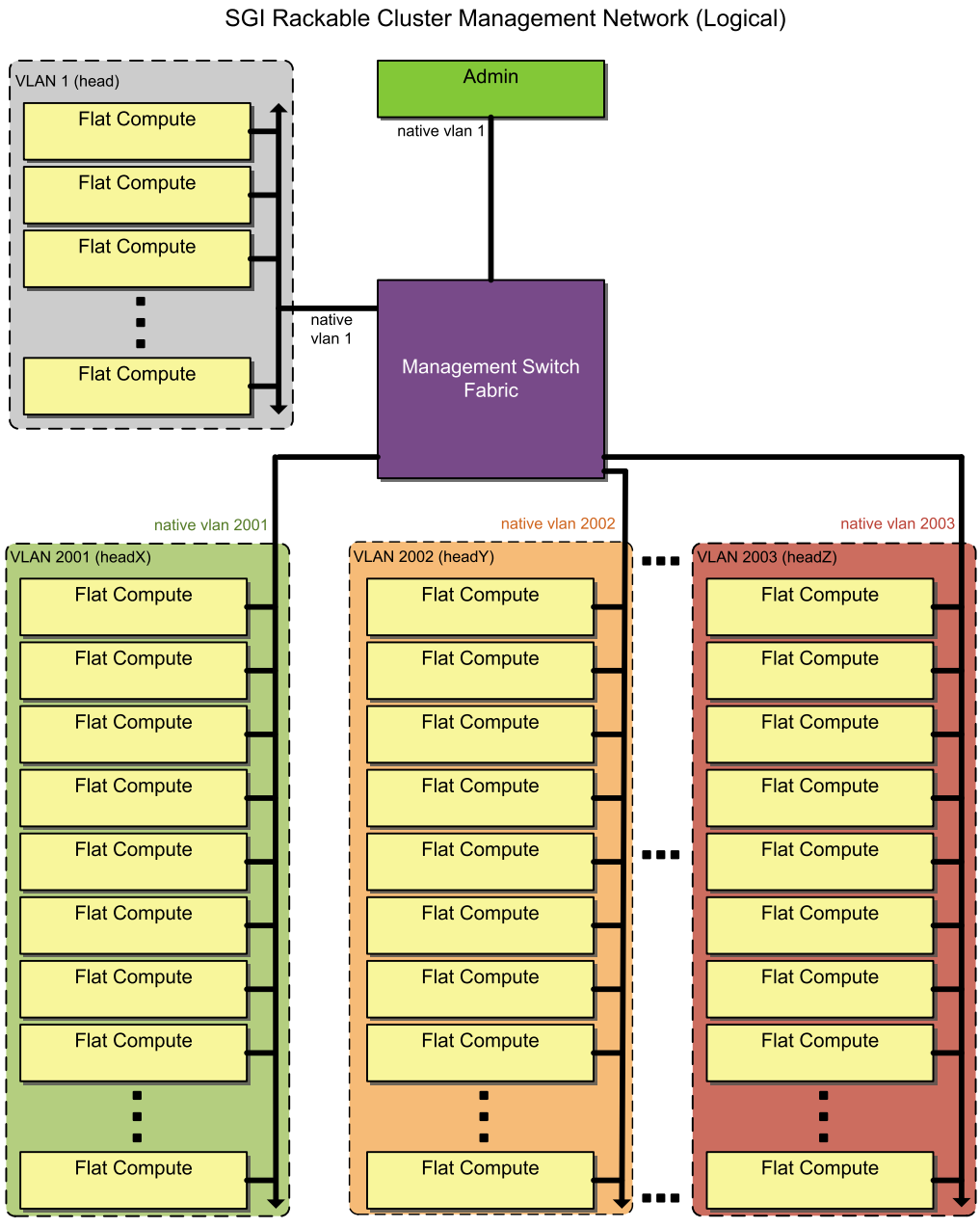

Figure 1-4 SGI Rackable Cluster Logical Representation

SGI ICE X Networks

The system components in an SGI ICE X cluster are attached to one or more of the following two internal networks:

- The data network.
- The management network.

The data network is designed for high-performance computing and bandwidth-intensive applications. This network is based on InfiniBand (IB) technology, and it facilitates communication to all SGI ICE compute nodes from the flat compute nodes. It connects the following:

- The SGI ICE compute nodes to each other. The InfiniBand network connects all the SGI ICE compute nodes (blades) to each other. The SGI ICE compute node InfiniBand connections are not part of the head network.
- The flat compute nodes to the SGI ICE compute nodes. One or two separate InfiniBand networks (or fabrics) 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 InfiniBand interface, as follows:
  - `ib0`, which is used typically for Message Passing Interface (MPI) communication.
  - `ib1`, which is typically used for storage traffic.

The management network, also known as the head network, is designed for monitoring, provisioning, and other functions not covered by the data network. This Ethernet network is designed for communication between the admin node, the RLCs, and the flat compute nodes. These components communicate to each other directly within the head network. The head network connects the following nodes directly into the Ethernet switches:

- Admin Node
- RLCs
- Flat compute nodes
- Additional Ethernet Switches

The head network also includes several additional virtual local area networks (VLANs). Figure 1-5 on page 14 is a logical representation of the SGI ICE X Ethernet

networks that shows how SGI ICE X components are logically separated by these VLAN boundaries.

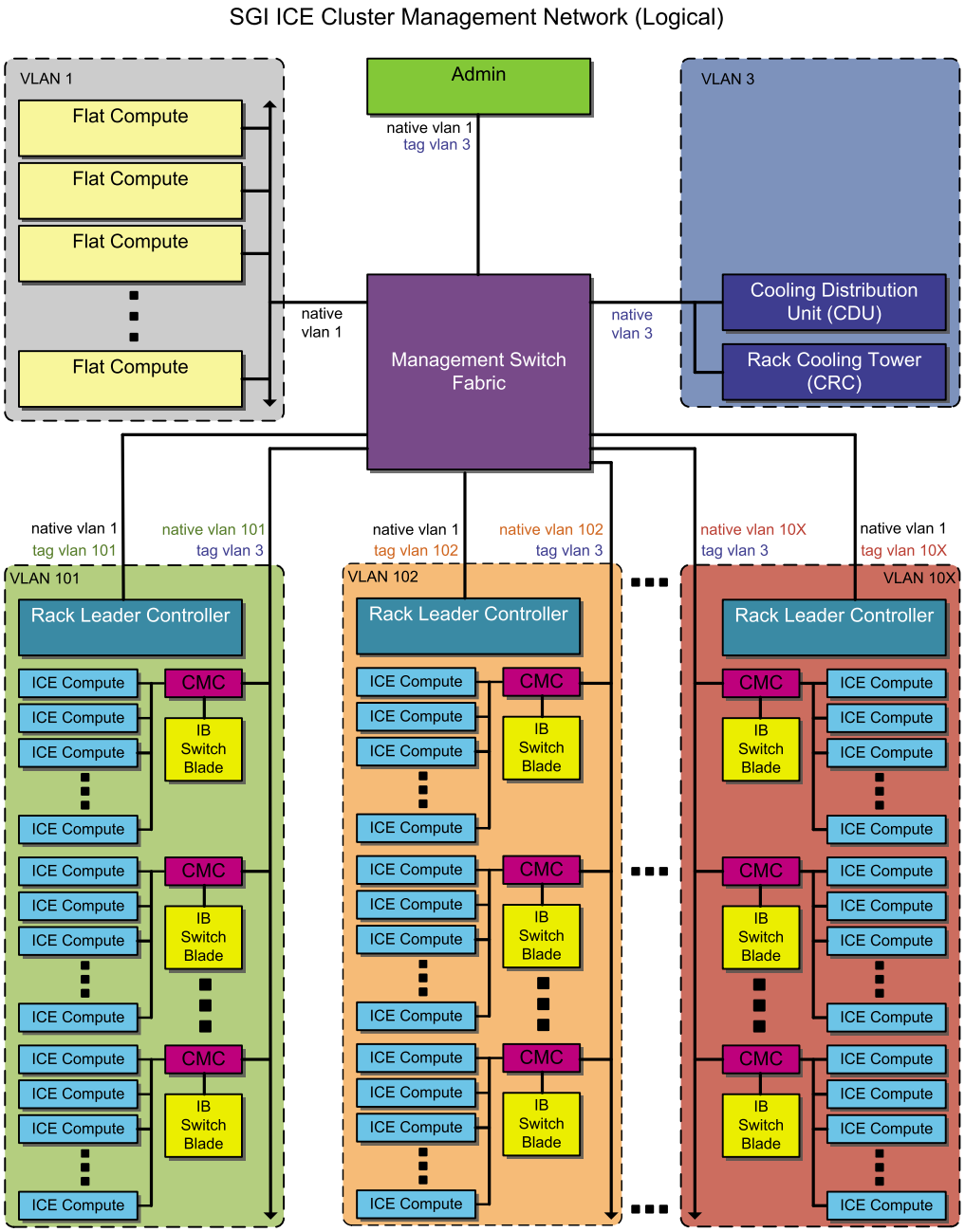

Figure 1-5 SGI ICE Networks

On an SGI ICE X cluster, several virtual local area networks (VLANs) are configured. The following list shows the typical VLAN names and numbers:

- Head network. VLAN tag 1.
- Cooling network. VLAN tag 3. Clusters with MCells only.
- Rack 1 network. VLAN tag 101.
- Rack 2 network. VLAN tag 102.
- Rack *x* network. VLAN tag 10*x*.

Note: The head VLAN network must always be VLAN 1. Do not attempt to change VLAN 1's number. You can change the other VLAN numbers and names.

In Figure 1-5 on page 14, the head network is VLAN 1. The ports connected to the admin node and the flat compute nodes are in VLAN 1 natively.

The Ethernet Switches are configured with a VLAN for each RLC. This VLAN segregates management traffic so that communication between the SGI ICE compute nodes and their corresponding RLC is contained within that VLAN. Physically, the SGI ICE compute nodes are connected to a chassis management controller (CMC) and do not directly connect to the Ethernet switch. Instead, the CMCs connect directly to the Ethernet Switch. Only the RLC can communicate with the SGI ICE compute nodes and CMCs in its own logical rack.

Users can log into the admin node and into the flat compute nodes directly. If access to the RLC is required, users will need to log directly into the admin node and then use the `ssh`(1) command to log into an RLC. The typical VLAN mapping for the Ethernet switches on each node is as follows:

Node type	**VLANs**
Admin	Native VLAN 1
	Tagged VLAN 3
Flat compute	Native VLAN 1
RLC	Native VLAN 1
	Tagged VLAN 10*x*. VLAN created for each RLC.

CMCs	Native VLAN 10*x*. VLAN created for their corresponding RLC. Tagged VLAN 3.
Cooling equipment cooling distribution units (CDUs) and cooling rack controllers (CRCs)	Native VLAN 3. Clusters with MCells only.

Figure 1-6 on page 17 shows the physical networks of an SGI ICE cluster.

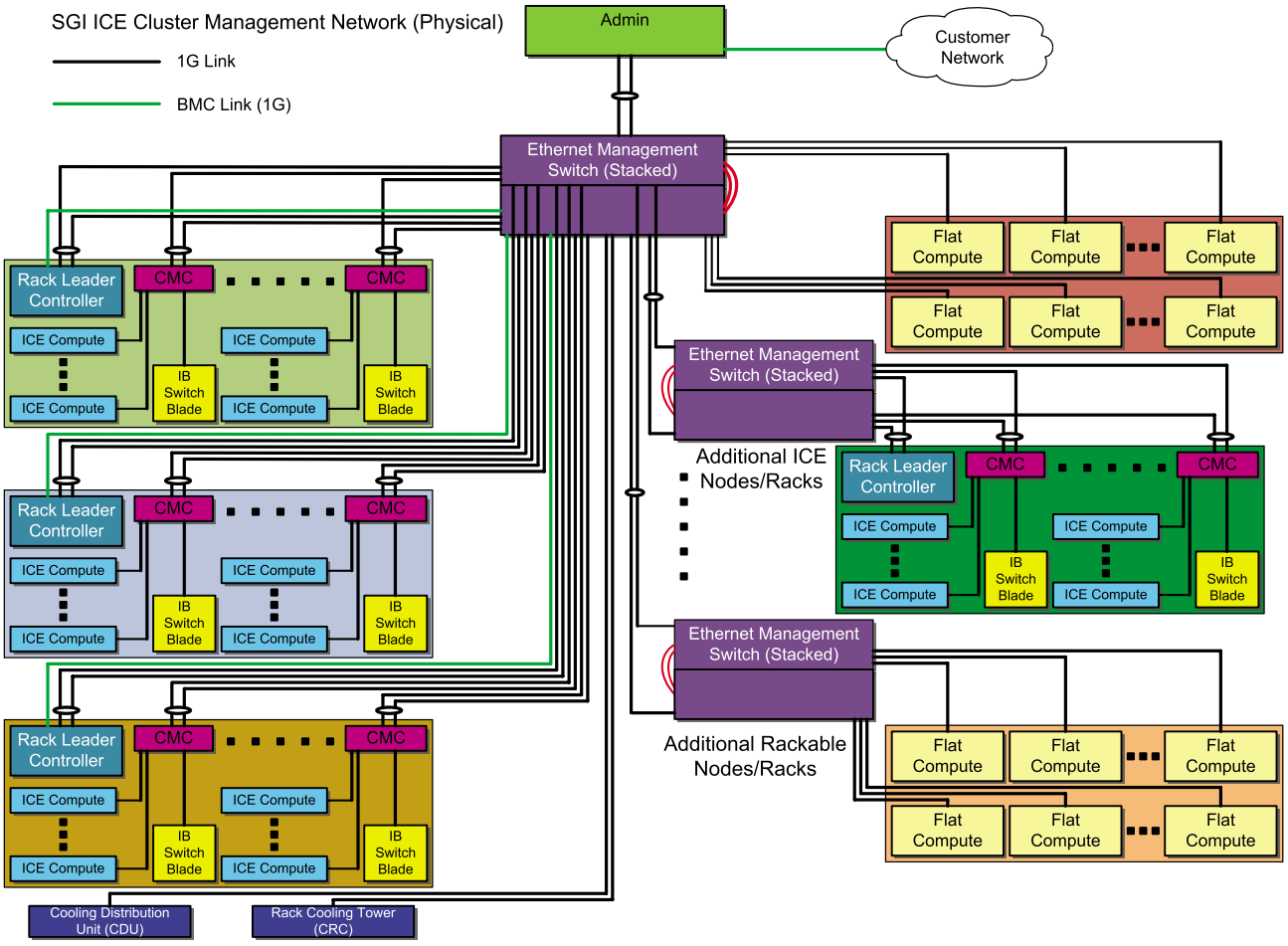

Figure 1-6 SGI ICE Cluster Physical Representation

Customizing a Factory-installed SGI Cluster

This chapter contains the following topics:

About Customizing a Factory-installed Cluster

Your SGI cluster was configured and tested 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 admin 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), CentOS, or SLES. Slot 2 is blank.

  The cluster supports a maximum of five slots. If you need more than the factory-configured 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 Cluster System" on page 33.
- A serial-over-LAN connection on the admin node. This connection allows you to use the admin node as the system console. You can access the console by using the IPMItool's serial-over-LAN function.

If you want to attach the cluster to your network and retain the factory-installed configuration, complete the following procedures:

- "Obtaining Information" on page 20

Obtaining Information

Your configuration session can proceed more quickly if you gather some information before you start. When you perform the configuration, you update the factory-installed, system-wide root password and the time zone. In addition, you provide information about your site network for the admin node'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 Cluster
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 _____

Cluster subdomain name _____

2. Proceed to the following:

"Changing the Password and Specifying Network Information" on page 21

Changing the Password and Specifying Network Information

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

Procedure 2-2 To change the password and attach a factory-installed cluster to your site network

1. Use the console attached to the admin node, and log into the admin node as the root user.
2. Type the following command, and follow the prompts, to change the root password on the admin node and on all other nodes:

   # **cpasswd**

   If you need help, type the following command:

   # **cpasswd --h**

   If you do not have the current password, you can obtain the factory-installed password from your SGI representative.

   For example:

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

3. Change the system time zone.

   This step is different, depending on the cluster's operating system, as follows:

- On RHEL platforms, 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.

- On SLES platforms, complete the following steps:
  - Type the following command to start YAST:

    # **yast**

  - Select **System** > **Date and Time**.
  - On the **Clock and Time Zone** page, select the correct setting.
  - Click **Change**
  - Click **Back**
  - Exit YAST.

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 ifcfg-eth0:
   - On RHEL platforms, the path is as follows:

     /etc/sysconfig/network-scripts/ifcfg-eth0

   - On SLES platforms, the path is as follows:

     /etc/sysconfig/network/ifcfg-eth0

5. Edit file ifcfg-eth0 as follows:
   - On RHEL platforms, add the IPADDR and NETMASK lines, and then add values appropriate for your site network. Also add a line that includes ONBOOT=yes.

For example:

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

- On SLES platforms, add the `IPADDR` and `NETMASK` lines, and then add values appropriate for your site network. Also add a line that includes `STARTMODE='onboot'`.

  For example:

  ```
  IPADDR='128.162.244.88'
  NETMASK='255.255.255.0'
  STARTMODE='onboot'
  ```

6. Save and close file `ifcfg-eth0`.
7. Create the networking configuration file.
   - On RHEL platforms, complete the following steps:
     - Use a text editor to create the following file:

       `/etc/sysconfig/network`

     - Add the following three lines to file `/etc/sysconfig/network`:

       NETWORKING=yes
       HOSTNAME=*admin_node_hostname*
       GATEWAY=*gateway_IP_address*

       For *admin_node_hostname*, type the hostname you want to assign to the admin node.

       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
       ```

     - Save and close file `/etc/sysconfig/network`.

- On SLES platforms, complete the following steps:
  - Create the following file:

    `/etc/sysconfig/network/routes`

- Add the following line to file `/etc/sysconfig/network/routes`:

  `default` *gateway* `- -`

  For *gateway*, type the IP address for the site gateway server.

For example:

```
default 100.100.100.101 - -
```

Note that SMC manages and rewrites everything below the `default` *gateway* line.

8. Use a text editor to open file `/etc/hosts`.
9. Add a line in the following format to file `/etc/hosts`:

   admin_node_IP admin_node_FQDN admin_node_hostname

   The variables in the preceding line are as follows:

   - For *admin_node_IP*, type the IP address of the admin node.
   - For *admin_node_FQDN*, type the fully qualified domain name (FQDN) of the admin node.
   - For *admin_node_hostname*, type the hostname of the admin node.

   For example, add the following line:

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

10. Save and close file `/etc/hosts`.
11. Type the following command to set the admin node's hostname:

    # **hostname** ***admin_node_hostname***

    For *admin_node_hostname*, type the hostname of the admin node. Make sure to type the hostname, which is the short name. Do not type the admin node's FQDN, which is the longer name.

For example:

```
# hostname acme-admin
```

12. Proceed to the following:

"Completing the Customization" on page 25

Completing the Customization

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

Procedure 2-3 To customize the cluster database

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

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

   For general information about the cluster configuration tool, see "Configuring the Cluster" on page 58.

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 to 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 YAST menu. Follow the prompts in the YAST 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. Use the `cadmin` command, in the following format, 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. Use the `cadmin` command, in the following format, to change the subdomain name for the cluster:

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

   For *cluster_name*, specify the name of the system. For example, `ice.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 name of the admin node:

   # **`cadmin --show-hostname --node admin`**

8. Use the `cadmin` command, in the following format, to customize the admin node name:

   # **`cadmin --set-hostname --node admin`** ***`new_admin_node_name`***

   For *new_admin_node_name*, specify the name you want to use for the cluster's admin node. For example, specify the hostname.

9. Type the following command to list the system images:

   # **`cinstallman --show-images`**
   The following list shows the types of nodes that each cluster includes:

SGI Rackable clusters	**SGI ICE X clusters**
One admin node	One admin node
Several flat compute nodes	Several flat compute nodes
	One rack leader controller (RLC)
	Several SGI ICE compute nodes

   The characteristics of each node are as follows:

   - The admin node is the administrative node. You log into this node to install additional software, create user accounts, and so on.

- The flat compute nodes include disks. They are used for computation or user services, such as workload managers, network address translation (NAT) services, and so on.
- The rack leader controllers (RLCs) manage the SGI ICE compute nodes. For each rack of SGI ICE compute nodes, there is one RLC.
- Several diskless SGI ICE compute nodes. These nodes are also called *blades*. The SGI ICE compute nodes are housed in racks, and each rack has one RLC.

The admin node 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 admin node and the production images on the nodes. For more information about the system images and how they install, see "About Performing a New Installation and Configuring the Software on an SGI Cluster" on page 34.

The following example shows how to list the paths to the node images on a RHEL 6.5 admin node:

```
# cinstallman --show-images
Image Name                              BT VCS Compat_Distro
lead-rhel6.5                            0  1   rhel6
rhel6.5                                 0  1   rhel6
ice-rhel6.5                             0  1   rhel6
```

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

10. Use the cp(1) command, in the following format, 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 cp(1) command for each master node system image. On an SGI ICE X cluster that includes both SGI ICE compute nodes and flat compute nodes, you need three cp(1) commands, one for each image type. On an SGI Rackable cluster, you need one cp(1) command.

Example 1. On an SGI ICE X cluster with RHEL 6.5, type the following commands, one for each master image:

```
# cp /etc/localtime /var/lib/systemimager/images/ice-compute-rhel6.5/etc
# cp /etc/localtime /var/lib/systemimager/images/rhel6.5/etc
# cp /etc/localtime /var/lib/systemimager/images/lead-rhel6.5/etc
```

Example 2. On an SGI Rackable SLES 11SP3 cluster, type the following command to copy the compute node's master image:

```
# cp /etc/localtime /var/lib/systemimager/images/sles11sp3/etc
```

11. Update the RLC image and flat compute node images with your site's time zone information.

    On an SGI ICE X cluster, type the following command:

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

    On an SGI cluster that includes flat compute nodes, type the following command:

    ```
    # pdcp -g compute /etc/localtime /etc/localtime
    ```

12. Proceed to one of the following:

    - If you have an SGI ICE X cluster, proceed to the following:

      "(Conditional) Pushing Changes to SGI ICE Compute Nodes" on page 28

    - If you have an SGI Rackable cluster, proceed to the following:

      "Configuring Additional Features" on page 30

(Conditional) Pushing Changes to SGI ICE Compute Nodes

Cluster admin nodes can host multiple forms of system images for each of the rack leader controller (RLC) nodes, SGI ICE compute nodes, and flat compute nodes. For example, you can have some production images and some test images, and you can push the images to the nodes as needed. "Completing the Customization" on page 25 explains how to update the master images for the RLC nodes and the compute nodes. In addition, that procedure explains how to push the updated the node images to the cluster's nodes.

In this topic's procedure, you push the updated SGI ICE compute node images to the SGI ICE compute nodes. The push action enables the SGI ICE compute nodes to run with the updated networking information that you configured in "Completing the Customization" on page 25.

Procedures that describe system operations, later on in this manual, refer back to this procedure because you need to update system images and push out new images as part of several system administration tasks.

The following procedure explains how to push SGI ICE compute node system images to compute nodes.

Procedure 2-4 To push software system images

1. (Conditional) Stop the SGI ICE compute nodes

   Complete this step if you want to boot from NFS roots. Do not complete this step if you want to boot from `tmpfs` roots.

   Type the following command:

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

   The preceding command stops all the SGI ICE compute nodes. Use the preceding command when you need to push an updated image out to all the nodes.

2. (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 SGI ICE compute image to all the SGI ICE 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 SGI ICE compute nodes at a time. If you have fewer than eight 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 SGI ICE compute nodes:

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

     # **`cnodes --leader`**

- Type the following command one or more times 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
```

3. Use the `cimage` command, in the following format, to push the changes:

   `cimage --push-rack ice-`*image_name rack*

   For *image_name*, specify the name of the SGI ICE compute node image that you updated.

   For *rack*, specify the nodes. To specify all SGI ICE compute nodes, specify `r\*` or `r*i*n*`. To specify only selected nodes, specify `r`*x*`i`*x*`n`*x*, and substitute specific integer numbers for the *x* characters.

   For example, the following command pushes the time zone changes (from "Completing the Customization" on page 25) to all the SGI ICE compute racks:

   ```
   # cimage --push-rack ice-rhel6.5 r\*
   ```

4. Type the following command to power-up the SGI ICE compute nodes:

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

5. Proceed to the following:

   "Configuring Additional Features" on page 30

Configuring Additional Features

The following procedure explains where you can obtain information about how to configure additional features.

Procedure 2-5 To configure additional features

1. Configure additional features.

   For example, the hardware event tracker (HET) is configured by default, but SGI recommends that you configure the email address to which HET sends critical event notifications. Other features, such as CPU frequency scaling might benefit your installation. For information about these additional features, see the following:

   Chapter 4, "Configuring Additional Features" on page 123

2. (Optional) Configure optional features.

   The clusters support several optional features, for example, networking features such as network address translation (NAT). For information about how to configure optional features, see the *SGI Management Center Administration Guide for Clusters*.

Installing and Configuring an SGI Cluster System

This chapter contains the following topics:

About Performing a New Installation and Configuring the Software on an SGI Cluster

SGI installs operating system software on each cluster 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 35 depicts the software installation process.

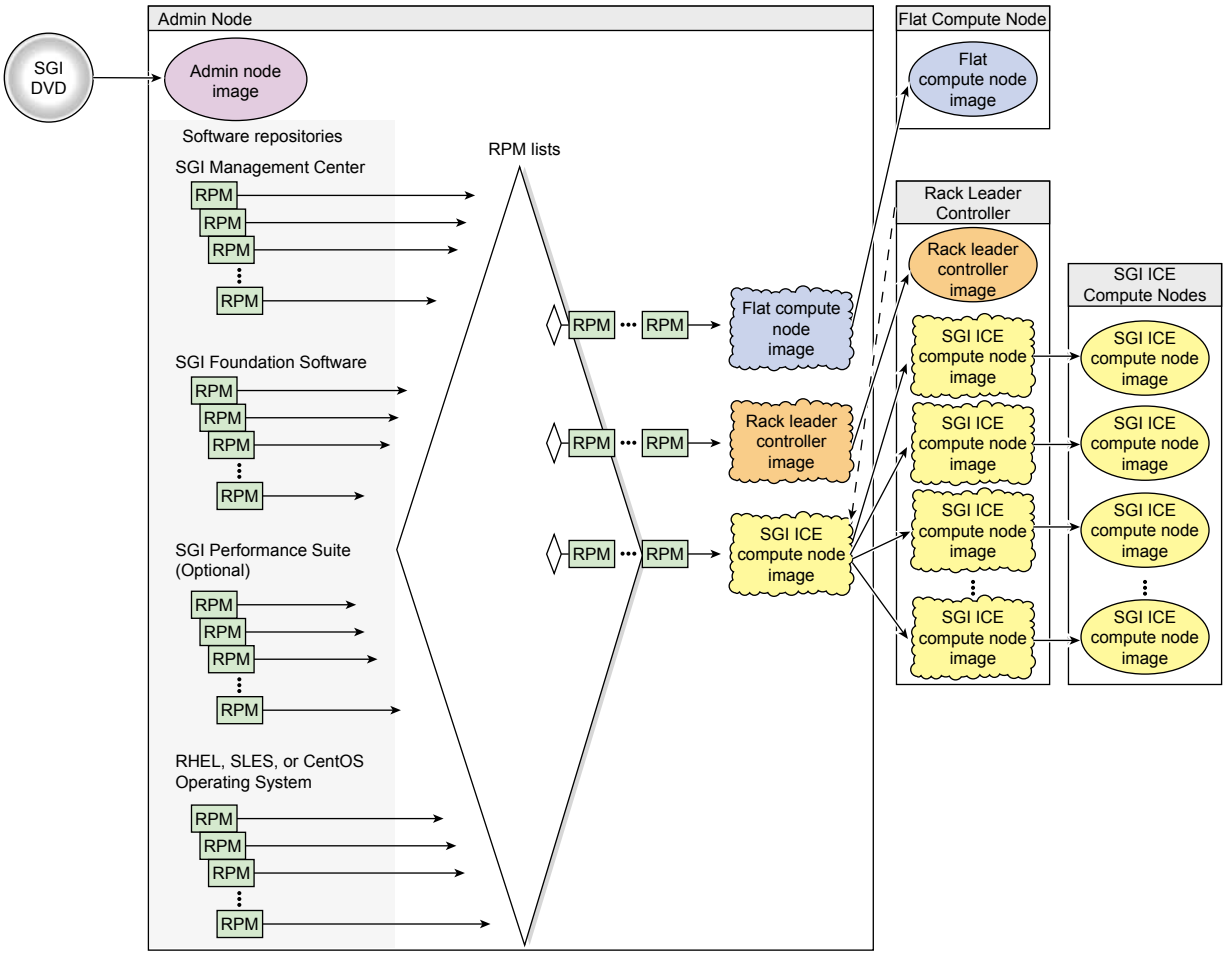

Figure 3-1 SGI Management Center Software Installation Process

Table 3-1 on page 36 shows the installation and configuration procedures to follow if you want to install the a cluster system from scratch. The cluster installation process is the same for SGI ICE X clusters and SGI Rackable clusters. In the case of the SGI Rackable clusters, the SMC software omits the steps that install images on RLCs and on SGI ICE compute nodes. In this case, you reinstall the operating system on the nodes and configure everything yourself.

Table 3-1 SGI ICE System Installation and Configuration Process

Step	Task	See
1	Plan the image installation method.	"Planning the Image Installation Method" on page 37
2	Prepare to install the cluster software.	"Preparing to Install Software on a Cluster" on page 39
3	(Conditional) Configure a static address for the baseboard management controller (BMC) on the admin node. 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 Admin Node" on page 41
4	(Optional) Configure a highly available admin node or a highly available rack leader controller (RLC).	"(Optional) Configuring a High Availability Admin Node or a High Availability Rack Leader Controller (RLC)" on page 43
5	Boot the system.	"Booting the System" on page 43
6	Install the operating system on the admin node. You can install the Red Hat Enterprise Server (RHEL), SLES, or CentOS operating system.	"Configuring the Operating System on the Admin Node" on page 48
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 admin node software. • Configure network settings. • Configure the NTP server. • Set up the initial admin node infrastructure. • Configure the house network DNS resolvers.	"Configuring the Cluster" on page 58
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 73

Step	Task	See
9	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 74
10	(Conditional) Download the Intel Manycore Platform Software Stack (MPSS).	"(Conditional) Downloading the Intel Manycore Platform Software Stack (MPSS) Software and Creating Images" on page 76
11	Configure the switches.	"Configuring the Switches" on page 86
12	Use the `discover` command to install and configure software on the rack leader controller and the flat compute nodes.	"Configuring the Cluster With the `discover` Command" on page 106
13	(Optional) Configure a backup domain name service (DNS) server on a flat compute node.	"(Optional) Configuring a Backup Domain Name Service (DNS) Server" on page 113
14	Configure the InfiniBand subnetworks.	"(Conditional) Configuring the InfiniBand Subnetworks" on page 114
15	Configure optional features.	Chapter 4, "Configuring Additional Features" on page 123

Planning the Image Installation Method

The SMC `discover` command installs software images on the nodes and facilitates adding nodes to a cluster. You use the `discover` command during the initial installation, and you can use the `discover` command again later if you want to reconfigure a node's network settings or you want to update the cluster after a hardware equipment change.

SGI supports three different file transfer methods for use during installation. These methods are `rsync` (default), UDPcast, and BitTorrent. The BitTorrent method is supported for legacy clusters.

The fastest image installation method for your cluster depends on the cluster's topology. Before you begin the installation, familiarize yourself with the image transport and installation methods and make sure that your installation plan uses the

method that is most appropriate for your cluster. Your site network configuration can also affect the speed at which the `discover` command can push software to nodes.

The following procedure explains how to determine the image installation method that is most appropriate for your cluster.

Procedure 3-1 To plan the installation method

1. Determine the number and type of nodes that need to be imaged.

   When you run the `discover` command during system installation, only the flat compute nodes and the rack leader controllers (RLCs) receive software images. The SGI ICE compute nodes receive their images directly from their RLC, so you do not need to consider the number of SGI ICE compute nodes in this calculation. Count the number of nodes as follows:

   - If you have five or six nodes, the default transport method, `rsync`, is appropriate. For example, if you have three RLCs and two flat compute nodes, you can use the default method. You do not need to consider the number of SGI ICE compute nodes that are associated with each RLC. You do not need to edit your cluster definition file, nor do you need to plan for any additional command line options for the `discover` command.

     You do not need to complete the rest of this procedure. Proceed to the following:

     "Preparing to Install Software on a Cluster" on page 39

   - If you have more than five or six nodes, consider using the UDPcast transport method. If you have hundreds of flat compute nodes, you most definitely need to consider using UDPcast.

2. To use UDPcast, either plan to provide additional arguments to the `discover` command when you run it (later in the installation process) or edit the cluster definition file at this time.

   If you specify options on the `discover` command line, those options override those that appear in the configuration file. If you prefer to specify the UDPcast transport on the `discover` command line, plan to include the `udpcast` argument.

   For example, if you have three RLCs and 200 flat compute nodes, you can specify the following command:

```
# discover --leaderset 1,3,transport=udpcast --nodeset 1,200,transport=udpcast  --configfile myfile --all
```

If you want to edit the cluster definition file at this time, complete the following steps:

- Obtain a copy of the cluster definition file from your sales representative or generate one by typing the following command:

  `discover --show-configfile >` *filename*

  For *filename*, specify the output file name.

- Open the cluster definition file from within a text editor.
- Search in the file for each block of text that describes a node. Each node block begins with the keyword `temponame=`. For example, the following text block describes one of the RLCs:

```
temponame=r1lead, mgmt_bmc_net_name=head-bmc, mgmt_bmc_net_macs=00:25:90:58:8b:75,
mgmt_net_name=head, mgmt_net_macs=00:25:90:58:8a:94/00:25:90:58:8a:95,  redundant_mgmt_network=yes,
switch_mgmt_network=yes, mic=0, dhcp_bootfile=grub2,  conserver_logging=yes, conserver_ondemand=no,
console_device=ttyS1
```

- At the end of each node definition block, add the following:

  `, transport=udpcast`

- Save and close the file.

3. Proceed to the following:

   "Preparing to Install Software on a Cluster" on page 39

Preparing to Install Software on a Cluster

The following procedure explains the information you need to obtain before you begin working with the cluster. 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 to obtain network information.

   Obtain the information to use when you configure the baseboard management controller (BMC). Your network administrator can provide an IP address, a

hostname, or a fully qualified domain name (FQDN) for each of the following addresses:

- (Optional) The current IP address of the BMC on the admin node. 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.

Obtain the following information to use when you configure the network for the SGI ICE 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)

(Conditional) Obtain information for one or more routed management networks. SGI recommends that you configure one or more routed management networks for every 300–500 flat compute nodes in your cluster. When you have a large number of flat compute nodes, a routed management network reduces the run-rate overhead that is associated with broadcast traffic. Obtain the following information for each routed management network that you want to configure:

- A name for the routed management network. For example, `head2`.
- Subnetwork address.
- Network mask for the subnetwork address.
- BMC subnetwork address.
- Network mask for the BMC subnetwork address.

2. (Optional) Obtain the configuration file for your cluster from your SGI representative.

   The configuration file contains system data, for example, the 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 Admin Node

Perform the procedure in this topic if one of the following is true:

- Your site practices require a static IP address for the BMC.
- You want to configure a high availability admin node. In this case, perform this topic's procedure on the BMCs on each of the two admin nodes.

When you set the IP address for the BMC on the admin node, you ensure access to the admin node when the site DHCP server is inaccessible.

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 admin node.

Procedure 3-4 Method 2 — To change the IP address from the admin node.

1. Log into the admin node 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 different from the IP address that is set currently.

   Type `ipmitool` commands in the following format:

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

   The arguments are as follows:

Argument	Specification
ip_addr	The IP address you want to assign to the BMC.
netmask	The netmask you want to assign to the BMC.
gateway	The gateway you want to assign to the BMC.

   For example, you can type the following commands to set the IP address 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
   ```

6. Proceed to one of the following:

- If you want to configure a high availability admin node, proceed to the following:

  "(Optional) Configuring a High Availability Admin Node or a High Availability Rack Leader Controller (RLC)" on page 43

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

  "Booting the System" on page 43

(Optional) Configuring a High Availability Admin Node or a High Availability Rack Leader Controller (RLC)

SGI supports the ability to configure the admin node and rack leader controllers (RLCs) as highly available nodes in an SGI ICE X cluster. If you want to enable high availability (HA) on the admin node or on the RLCs, contact your SGI representative.

Booting the System

You can configure the cluster to boot from one, two (default), three, four, or five partitions, or *slots*. This feature enables you to configure either a single-boot cluster or a multiple-boot cluster. A multiple-boot cluster has two or more partitions, so it has more than one root directory (`/`) and more than one boot directory (`/boot`). In a cluster, 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 installer creates the same disk layout on all nodes.

On a multiple-slot cluster, the nodes all have the same disk layout. When you insert an operating system installation disk and power-on the admin node, you can select a boot method from the GNU GRUB menu. If you select **Install: Wipe Our and Start Over: Prompted**, which is the default, the installer creates two slots and writes the initial installation to slot 1. After the system is installed, you cannot change 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. On a multislot cluster, you can boot the cluster with the operating system of your choice. This ability 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 multiple-boot systems:

Multiple-boot	Single-boot
You can install different operating systems, or different operating system versions, into different slots. Note that if you have an SGI ICE X cluster, the admin node and the RLCs must have the same operating system installed.	You can install only one operating system for the entire cluster.
RLCs and flat compute nodes boot from their own disk. Data is retained in the master boot record (MBR).	RLCs and flat compute nodes boot from the boot partition in the slot that is currently configured as the boot slot. Only the admin node retains data in the MBR.
RLC and flat compute node software is reinstalled from the admin node.	Software on the RLCs and flat compute 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.

If all slots on your cluster are running either SGI Tempo 2.9.0 (or later) or SMC 3.0 (or later), then your cluster uses the partition layout designed for the SGI Tempo 2.9.0 and later releases. If you upgraded your cluster, it is possible that you have the legacy partition layout on one or more slots. For information about partitions, including those for legacy partitions, see the following:

Appendix D, "Partition Layout Information" on page 195

Note: SMC supports both EFI BIOS and legacy x86_64 BIOS. If you are not sure which BIOS your cluster supports, contact your sales representative.

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

Procedure 3-5 To boot the system

1. Power-on the admin node.

   As Figure 3-2 on page 45 shows, the power-on button is on the right of the admin node.

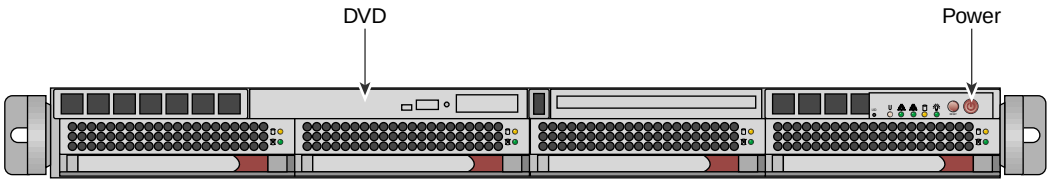

Figure 3-2 Admin Node Power On Button and DVD Drive

2. (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-2.6.16.56-0.12-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. By default, this is set to `ttyS1,38400n8`. You might have `ttys2`, for example.

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

3. Insert the SGI Admin Node Autoinstallation DVD into the DVD drive on the admin node.

4. Use the arrow keys to select one of the boot options, press Enter, and monitor the installation.

   On the GNU GRUB boot menu, the options are as follows:

   - **Display Instructions**

     Select this option if you want information about custom boot parameters. This option displays information and returns to the boot menu.

Each boot option has a set of default behaviors. In addition to the default behaviors, you can specify custom boot parameters if you select one of the following options:

- **Install: Install to Designated Slot**
- **Install: Wipe Out and Start Over: Prompted**
- **Install: Custom, type 'e' to edit kernel parameters**

If you think you might want to specify one or more custom boot parameters, for example, `console=`, select the **Display Instructions** option and familiarize yourself with these parameters before you select an actionable parameter.

- **Install: Install to Designated Slot**

  Select this option if you have an open slot on your cluster, and you want to install an operating system in the slot. If you select this option, only the open slot is affected. All other slots remain as configured.

- **Install: Wipe Out and Start Over: Prompted**

  Select this option if you want to reinstall the cluster. This options destroys all information currently on the cluster. The installer partitions the admin node with 2 slots, and the installer writes the initial installation to slot 1. For example, for an initial installation, select this option.

- **Rescue: Prompted**

  Select this option to create a troubleshooting environment.

- **Install: Custom, type 'e' to edit kernel parameters**

  Select this option if you want to perform a custom installation. For example, because the default number of slots is 2, if you want to install 1, 3, 4, or 5 slots, select this menu option.

All the options launch you into an installation dialog, and at the end of the dialog, the final question asks you to confirm your choices. In this way, you have the chance to cancel your choices and return to the GNU GRUB boot menu to start over. The following are some of the installation dialog prompts that appear when you select a boot option:

- **Enter number of slots to allow space for: (1–5):**

  Type **1**, **2**, **3**, **4**, or **5**, and press Enter.

This dialog question appears only if you select **Install: Wipe Out and Start Over: Prompted** from the GNU GRUB menu.

- **Enter which slot to install to:**

  Type **1**, **2**, **3**, **4**, or **5**, and press Enter.

  This dialog question appears only if you select **Install: Install to Designated Slot** from the GNU GRUB menu.

- **Destructively bypass sanity checks? (y/n):**

  If you type **y** and press Enter, the installer proceeds without checking to see if there is any data in the partition.

  If you type **n** and press Enter, the installer checks to see if there is data in the partition before proceeding.

- **Additional parameters (like console=, etc):**

  If you want to specify any additional boot parameters, type them in a comma-separated list and press Enter.

  For information about the boot parameters that are available, select Display Instructions from the GNU GRUB menu and press Enter.

- **OK to proceed? (y/n):**

  If you type **y** and press Enter, the boot proceeds.

  If you type **n** and press enter, the menu returns you to the main GNU GRUB menu.

The installation can take several minutes.

5. Remove the operating system installation DVD.
6. At the # prompt, type **reboot**.

   This is the first boot from the admin node's hard disk.
7. (Optional) Suppress log messages.

If you want to suppress the admin node's log message output to the screen during the boot, 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.

8. Proceed to the following:

"Configuring the Operating System on the Admin Node" on page 48

Configuring the Operating System on the Admin Node

The cluster systems support the Red Hat Enterprise Linux (RHEL) operating system, CentOS, and the SLES and operating system. Use one of the following procedures to install your operating system software on the admin node:

- "Configuring Red Hat Enterprise Linux (RHEL) on the Admin Node" on page 48
- "Configuring SLES on the Admin Node" on page 53

Note: In SGI documentation, you can assume that feature descriptions for RHEL platforms also pertain to CentOS platforms unless otherwise noted.

Configuring Red Hat Enterprise Linux (RHEL) on the Admin Node

This section describes how to configure Red Hat Enterprise Linux 6 on the admin node.

Procedure 3-6 To install RHEL 6 on an SGI ICE admin node

1. Use one of the following methods to connect to the admin node:
   - Through the intelligent platform management interface (IMPI) tool
   - Through the console attached to the cluster
   - 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 site 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=admin_node_hostname
   GATEWAY=gateway_IP_address
   ```

   For *admin_node_hostname*, type the hostname you want to assign to the admin node.

   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`:

   admin_node_IP admin_node_FQDN admin_node_hostname

   The variables in the preceding line are as follows:

   - For *admin_node_IP*, type the IP address of the admin node.

- For *admin_node_FQDN*, type the fully qualified domain name (FQDN) of the admin node.
- For *admin_node_hostname*, type the hostname of the admin node.

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 admin node hostname:

    # **hostname *admin_node_hostname***

    For *admin_node_hostname*, type the hostname of the admin node.

    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 `nscd`(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 admin node's current time zone information:

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

The previous output shows the admin node set to US Central time. If the output you see is not correct for this cluster, perform the following steps:

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

```
# cd /usr/share/zoneinfo
```

- Select a file from that directory that describes the time zone for the admin node.
- Type the following 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 admin node's time zone to US Pacific time:

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

- 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:

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

    - At the end of the file, add a line 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.

    - Type the following command to restart the NTP server:

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

18. (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.

19. (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-2.6.16.56-0.12-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. By default, this is set to `ttyS1,38400n8`. You might have `ttys2`, for example.

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

20. Proceed to the following:

    "Configuring the Cluster" on page 58

Configuring SLES on the Admin Node

The SLES YAST interface enables you to install the SLES operating system on a cluster. To navigate the YAST modules, use key combinations such as the following:

- The Tab key moves the cursor forward, and the Shift + Tab keys move the cursor backward.
- The arrow keys move the cursor up, down, left, and right.
- To use shortcuts, press the Alt key + the highlighted letter.
- Press Enter to complete or confirm an action.
- Press Ctrl + L to refresh the screen.

For more information about navigation, see Appendix A, "YAST Navigation" on page 171.

The following procedure explains how to use YAST to install SLES 11 on a cluster.

Procedure 3-7 To install SLES 11 on an SGI ICE admin node

1. Connect to the admin node by one of the following methods:
   - Through the intelligent platform management interface (IMPI) tool
   - Through the console attached to the cluster
   - 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 cluster.
   - 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 admin node's house/public network interface.

   Figure 3-3 on page 55 shows the **Network Card Setup** screen.

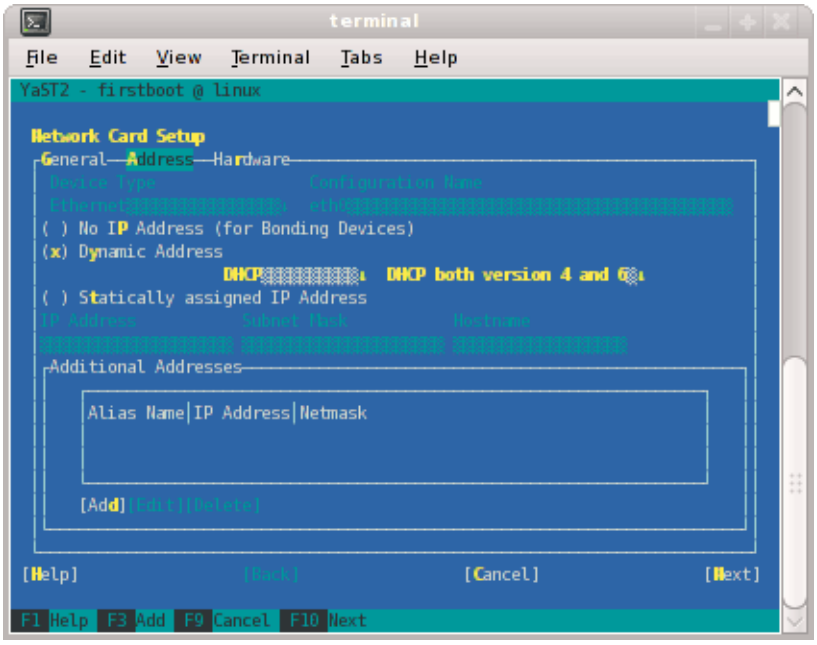

Figure 3-3 Network Card Setup Screen

Complete the following steps:

- Select **Statically Assigned IP Address**. SGI recommends a static IP address, not DHCP, for the admin node.
- 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 system. These nodes are as follows:

- admin node
- Flat compute nodes
- Rack leader controller (RLC) (Optional)
- SGI ICE compute nodes (blades) (Optional)

- 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 admin node, open file `/etc/hosts` within a text editor, and verify that the admin node's fully qualified domain name (FQDN) and hostname are entered correctly.

    For example, the following `/etc/hosts` file entry contains the correct data in the three required fields and is correct for an admin node with an IP address of `100.100.100.100`, an FQDN of `mysystem-admin.mydomain.com`, and a hostname of `mysystem-admin`:

    ```
    100.100.100.100     mysystem-admin.mydomain.com     mysystem-admin
    ```

Make sure that the `/etc/hosts` file on the admin node contains the required information. If it does not, edit the `/etc/hosts` file to contain the three required fields as the preceding example shows.

17. Confirm that the system is working as expected.

    If necessary, restart YAST to correct settings.

18. (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-2.6.16.56-0.12-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. By default, this is set to `ttyS1,38400n8`. You might have `ttys2`, for example.

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

19. Proceed to the following:

    "Configuring the Cluster" on page 58

Configuring the Cluster

Configuring the cluster includes the following actions:

- Creating repositories for software installation files and updates.
- Installing the admin node's cluster software.
- Configuring the cluster subdomain and examine other network settings. The cluster subdomain is likely to be different from the `eth0` domain on the admin node itself.

- Configuring the NTP server.
- Installing the cluster's software infrastructure. This step can take 30 minutes.
- Configuring the house network's DNS resolvers.

You can configure the cluster in one of the following ways:

- With the cluster configuration tool.

  The cluster configuration tool is a menu-driven tool that enables you to supply information about your cluster. You can use the tool to configure, or reconfigure, your cluster. The procedure in this topic explains how to use the cluster configuration tool to complete the general, required configuration steps. If your cluster 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.

- With a cluster definition file.

  As an alternative, you can specify a cluster definition file as an argument to the `configure-cluster` command. When you do this, the `configure-cluster` command reads the options in the cluster definition file and implements them in the cluster. The cluster definition file supplies the information that you would typically define by using the menus in the cluster configuration tool.

  To create a cluster definition file that you can examine, type the following command:

  ```
  discover --show-configfile > name
  ```

  For *name*, specify a file name of your choice.

The following procedure explains how to use either the cluster configuration tool or the cluster definition file to configure the cluster:

Procedure 3-8 To configure the cluster

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 admin node 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 admin node. Avoid running the `configure-cluster` command from a serial console.

3. Use either Method 1 or Method 2 to configure the cluster.

   Method 1 — Using the Cluster Configuration Tool — is as follows:

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

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

   - Proceed to the following step:

     Procedure 3-8, step 4 on page 61

   Method 2 — Using the Cluster Definition File — is as follows:

   - Type `crepo` commands to create repositories for each of the following software's initial installation packages and for updates:

     – The operating system software, either RHEL, SLES, or CentOS

     – SGI Foundation Software

     – SGI Management Center

     – (Optional) SGI Performance Suite

     Use the `crepo` command in the following format:

     `crepo --add` *rpm_repo_directory* `--custom` *rpm_repo_name*

     The variables in this command are as follows:

     – For *rpm_repo_directory*, specify the full path to the directory that contains the RPM files.

       If you have hard media mounted in the admin node's DVD drive, specify the path to that media. If you have the software for the operating system and the SGI packages in an ISO file on your network, specify the path to the files on your network.

     – For *rpm_repo_name*, create a name for the image. You can specify the same name for both *rpm_repo_directory* and *rpm_repo_name*. After the image is

built, the `cinstallman --show-images` command returns the *rpm_repo_name* in the `Image Name` column of its output.

For example, type the following commands:

```
# crepo --add /tmp/sles11sp3 --custom spes11sp3
# crepo --add /tmp/sfs --custom sfs
# crepo --add /tmp/smc --custom smc
# crepo --add /tmp/sps --custom sps
```

- Type the following command to define the cluster according to the content in the cluster definition file:

  ```
  # /opt/sgi/sbin/configure-cluster --configfile path
  ```

  For *path*, specify the path to the configuration file.

- Proceed to the following step:

  Procedure 3-8, step 35 on page 71

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

   Figure 3-4 on page 62 shows the initial window.

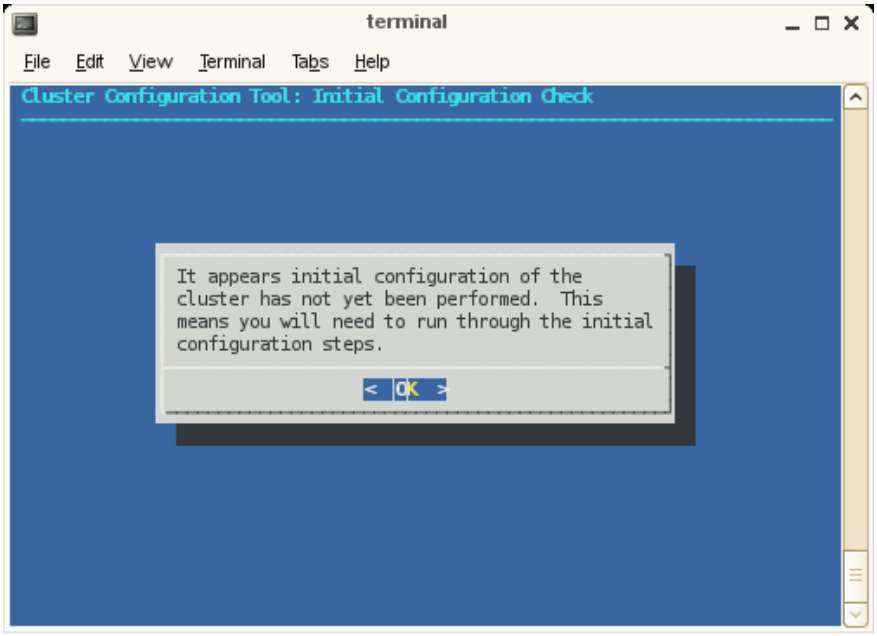

Figure 3-4 Initial Configuration Check Screen

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

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

   Figure 3-5 on page 63 shows the window.

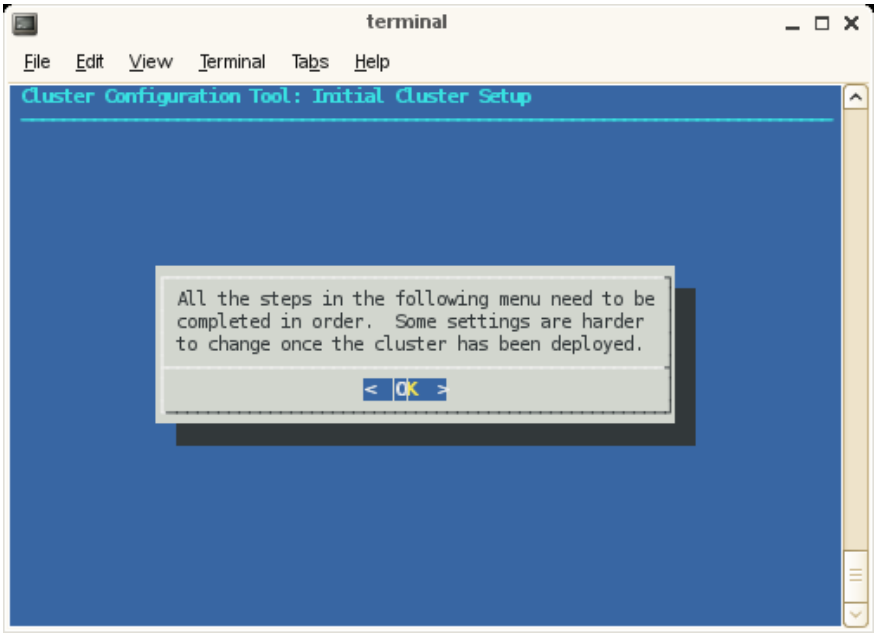

Figure 3-5 Initial Cluster Setup Screen with the initial screen

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

   Figure 3-6 on page 64 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.

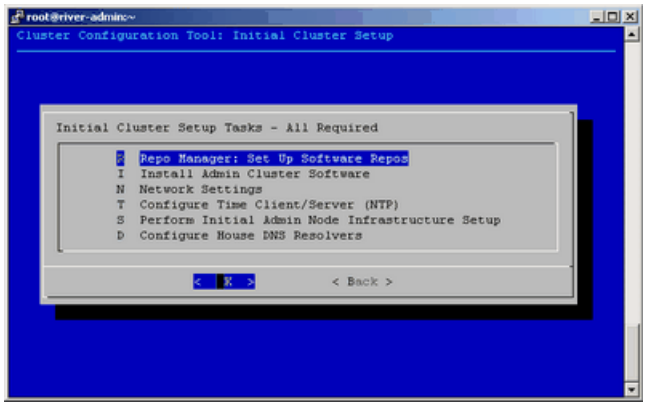

Figure 3-6 Initial Cluster Setup Tasks Screen

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 Software
- SGI Management Center
- (Optional) SGI Performance Suite

The menu system prompts you to insert hard media or specify a path for some of the preceding software, so locate your system disks before you proceed.

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:

- Insert a DVD.
- Select **Mount inserted DVD**.
- On the **Media registered successfully with crepo ...** screen, select **OK**, and eject the DVD.
- 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.

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

- Select **Use custom path/URL**.
- 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.
- On the **Media registered successfully with crepo ...** screen, select **OK**.
- 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. Repeat the following steps until all software is installed:

    • Procedure 3-8, step 10 on page 64

    • Procedure 3-8, step 11 on page 64

    If you plan to configure SGI MPT and run SGI MPT programs, make sure to install SGI-Accelerate and SGI-MPI from the SGI Performance Suite.

13. 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.

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

15. (Conditional) Create a routed management network.

    Complete this step if you have at least 300–500 flat compute nodes in your cluster. If you have more than 500 flat compute nodes, consider creating more than one routed management network.

    Complete the following steps:

    - On the **Cluster Network Settings** screen, select **A Add Subnet**, and select **OK**
    - On the **Select network type** screen, press the space bar to move the asterisk (*) up to the first line. This action selects the upper line, and the line now looks like this:

      ```
      (*) 1 mgmt/mgmt-bmc
      ```

    - Select **OK**.
    - On the **Insert network name, subnet, netmask, bmc subnet and bmc network** screen, type in the information to define the routed management network. Use the arrow keys to move from field to field on this screen. The information you need to enter is as follows:

Field name	Information
name	A unique name for this network. For example, head2.
subnet	The IP address for the nodes on the routed management network.
netmask	The network mask for the nodes on routed management network.
bmc subnet	The IP address for the node BMCs on the routed management network.
bmc netmask	The network mask for the node BMCs on the routed management network.

    - On the **Network *name* ...** screen, verify that this is the information you specified for the routed management network, and select **OK**.

- On the **Network *name*-bmc ...** screen, verify that this is the information you specified for the node BMC network, and select **OK**.

16. On the **Cluster Network Settings** screen, select **S List and Adjust Subnet Addresses**, and select **OK**.
17. On the **Warning: Changing the subnet IP addresses ...** screen, click **OK**.
18. Review the settings on the **Subnet Network Addresses** screen, and modify these settings only if absolutely necessary.

    Figure 3-7 on page 67 shows the **Subnet Network Addresses** screen. This screen displays the default networks and netmasks that reside within the cluster.

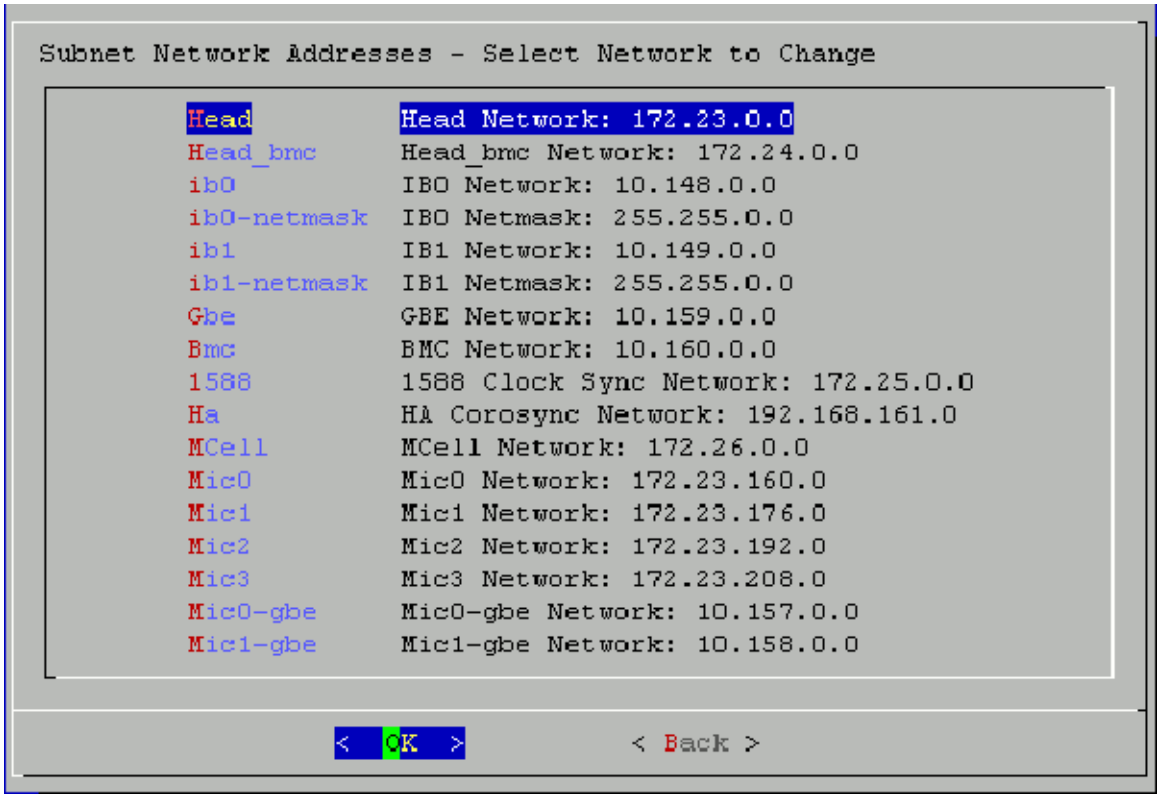

Figure 3-7 Subnet Network Addresses Screen

If you accept the defaults, select **OK**.

If you do not accept the defaults, you can change the network settings. For example, it is possible that your site has existing networks or conflicting network requirements. For additional information about the IP address ranges, see Appendix B, "Subnetwork Information" on page 173. Complete the following steps if you need to change the network settings:

- Highlight the setting you want to change, and select **OK**.
- Type in a new IP address, and select **OK**.
- Press Enter.

On the **Update Subnet Addresses** screen, the **Head Network** field shows the admin node's IP address. SGI recommends that you do not change the IP address of the admin node 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**.

19. On the **Cluster Network Settings** screen, select **D Configure Cluster Domain Name**, and select **OK**.
20. 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`.
21. On the **Cluster Network Settings** screen, select **Back**.
22. On the **Initial Cluster Setup** screen, select **T Configure Time Client/Server (NTP)**, and select **OK**.
23. Configure your NTP server.

    On the subsequent screens, you set the admin node as the time server to the cluster. For this step, the installer screens differ on RHEL platforms and SLES 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**.

24. On the **Initial Cluster Setup Tasks** menu, select **S Perform Initial Admin Node Infrastructure Setup**, and select **OK**.
25. On the **A script will now perform the initial cluster ...** screen, select **OK**.

    This step runs a series of scripts that configure the admin node. The scripts also create the root images for the RLCs, SGI ICE compute nodes, and flat 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
    ```

26. On the **Initial Cluster Setup Complete** window, select **OK**.
27. On the **One or more ISOs were embedded on the admin install DVD and copied to ...**, screen, select **OK**.

    Depending on what you have installed, this screen might not appear.

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

    Figure 3-8 on page 70 shows the **Configure House DNS Resolvers** screen.

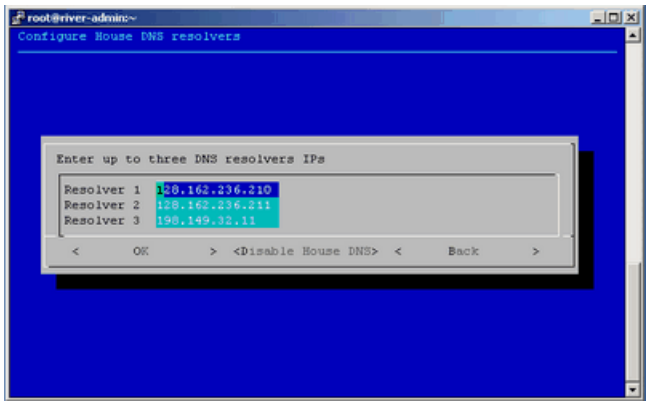

Figure 3-8 Configure House DNS Resolvers Screen

The system autopopulates the values on the **Configure House DNS Resolvers** screen to match the DNS specifications on the admin node. The DNS resolvers you specify here enable the flat compute nodes to resolve host names on your network. You can set the DNS resolvers to the same name servers used on the admin node 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, select **OK**, and then select **Yes**.
- To disable house network resolvers, select **Disable House DNS**.

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

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

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

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

    The switch management network enables the Ethernet switch to control all VLANs and trunking.

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

Figure 3-9 on page 71 shows the selection pop-up window:

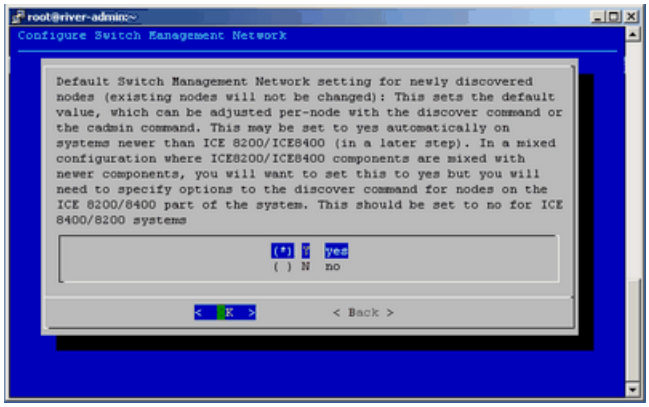

Figure 3-9 Configure Switch Management Network screen

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

    Perform this step if your SGI ICE system contains MCells.

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

    Perform this step if your SGI ICE system contains MCells.

34. Select **Quit**.

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

    Example: The following output is generated on an SGI ICE X cluster 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
  cluster_domain          : smc.americas.sgi.com
  tempo_dhcp_option       : 149
  head_vlan               : 1
  mcell_vlan              : 3
```

```
rack_vlan_start         : 101
rack_vlan_end           : 1100
mgmt_vlan_start         : 2001
mgmt_vlan_end           : 2500
redundant_mgmt_network  : yes
switch_mgmt_network     : yes
mcell_network           : yes
discover_skip_switchconfig : no
max_rack_irus           : 4
mic                     : 0
blademond_scan_interval : 120
dhcp_bootfile           : grub2
udpcast_min_receivers   : 1
udpcast_min_wait        : 10
udpcast_max_wait        : 10
udpcast_max_bitrate     : 900m
udpcast_rexmit_hello_interval : 0
udpcast_mcast_rdv_addr  : 224.0.0.1
my_sql_replication      : yes
conserver_logging       : yes
conserver_ondemand      : no
edns_udp_size           : 512
replication_file        : mysql-bin.000005
replication_position    : 9103
```

Note: On an SGI Rackable cluster, the `cattr` output is similar to the preceding example output, but the output contains fewer fields.

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
```

36. Proceed to one of the following:

    - To configure one or more external Domain Name Service (DNS) servers, proceed to "(Conditional) Configuring External Domain Name Service (DNS) Servers" on page 73.

- To synchronize the software repository, install updates, and clone the images, proceed to "Synchronizing the Software Repository, Installing Software Updates, and Cloning the Images" on page 74.

(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 cluster. A later procedure explains how to configure NAT as a service on a flat compute node. 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 SGI ICE X compute nodes (blades) in the cluster resolve through external DNS servers. The SGI ICE X 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 attempt to configure this feature after you run the `discover` command, the IP addresses assigned previously on the configured 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 SGI ICE compute nodes, and all flat compute nodes.

2. Through an `ssh` connection, log into the admin node as the root user.

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

   # **`/opt/sgi/sbin/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 "Synchronizing the Software Repository, Installing Software Updates, and Cloning the Images" on page 74.

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 cluster has a connection to the internet. If you need to perform this procedure on a secure cluster, 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 3-10 To update the software

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

   ```
   # 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 VCS Compat_Distro
ice-rhel6.5              1  1   rhel6
rhel6.5                  0  1   rhel6
lead-rhel6.5             0  1   rhel6
```

   The preceding output includes a line for the MPSS images if you have downloaded MPSS software for Intel Many Integrated Core Architecture (Intel MIC Architecture) based products.

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-rhel6.5`.

   For *image*, specify a file name for the copied file (the clone). For example: `lead-rhel6.5.backup`

   For example:

```
# cinstallman --create-image --clone --source ice-compute-rhel6.5 --image ice-compute-rhel6.5.backup
# cinstallman --create-image --clone --source rhel6.5 --image rhel6.5.backup
# cinstallman --create-image --clone --source lead-rhel6.5 --image lead-rhel6.5.backup
```

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

   For each *image*, specify the software package you want to install on each type of node.

For example, to install the packages shown in Procedure 3-10, step 4 on page 75, type the following commands:

```
# cinstallman --update-image --image ice-compute-rhel6.5
# cinstallman --update-image --image rhel6.5
# cinstallman --update-image --image lead-rhel6.5
```

7. Proceed to one of the following:

   - If your cluster contains Intel® Many Integrated Core Architecture (Intel MIC Architecture) devices, proceed to "(Conditional) Downloading the Intel Manycore Platform Software Stack (MPSS) Software and Creating Images" on page 76.
   - If your cluster does not contain MIC devices, proceed to "Configuring the Switches" on page 86.

(Conditional) Downloading the Intel Manycore Platform Software Stack (MPSS) Software and Creating Images

Perform the procedures in this topic if nodes in your cluster are equipped with Intel Many Integrated Core Architecture (Intel MIC Architecture) based products. The Intel Many Integrated Core (MIC) devices are part of the Intel Manycore Platform Software Stack that runs on the Intel Xeon Phi Coprocessors found on SGI ICE compute nodes and flat compute nodes.

Intel Corporation provides software for its Intel MIC architecture products, and you need to download this software for use on your SGI cluster. The MPSS package that you download contains the software packages for the MIC devices on the SGI cluster nodes. The procedures in this topic explain how to download the RPMs from the Intel Corporation website and how to create images for the nodes that are equipped with MIC devices. Your system might have MIC devices on SGI ICE compute nodes, on the flat compute nodes, or both. Complete the procedures that are appropriate for your hardware configuration.

The following procedures explain how to obtain and deploy the MPSS software from Intel Corporation:

- "Downloading the MPSS Software From the Intel Corporation" on page 77
- "Creating Images for the SGI ICE Compute Nodes That Include MIC Devices" on page 77

- "Creating Images for the Flat Compute Nodes That Include MIC Devices" on page 82

Downloading the MPSS Software From the Intel Corporation

Complete the procedure in this topic if you have any MIC devices on your cluster.

The following procedure explains how to download the MPSS software from Intel Corporation.

Procedure 3-11 To download the MPSS package

1. Open a browser, and navigate to the following website:

   http://software.intel.com/mic-developer

2. Click the **Tools & Downloads** tab.

3. Click the **Software Drivers: Intel Manycore Platform Software Stack (Intel MPSS)** link.

   Follow the instructions on the website to download the Linux software version for your operating system platform. The download comes in the form of a `tar`(1) file.

4. Use the instructions from Intel to build the RPM files that you need.

   A later procedure explains how to transfer these files to the cluster and build new images.

5. Proceed to one of the following topics:

   - "Creating Images for the SGI ICE Compute Nodes That Include MIC Devices" on page 77
   - "Creating Images for the Flat Compute Nodes That Include MIC Devices" on page 82

   Plan to perform both of the preceding procedures if you have MIC devices on both SGI ICE compute nodes and on flat compute nodes.

Creating Images for the SGI ICE Compute Nodes That Include MIC Devices

The following procedure explains how to create SGI ICE compute node images that include MIC device software.

Procedure 3-12 To create compute node images for SGI ICE compute nodes with MIC devices

1. On the admin node, use the `mkdir`(1) command, in the following format, to create a directory for the RPM repository:

   `mkdir -p /tftpboot/intel/`*`mpss_repository_directory`*

   For *mpss_repository_directory*, type a name for the directory that is to contain the MPSS repository. For convenience, make sure to include an identifier for the MPSS release level you downloaded.

   For example:

   # **`mkdir -p /tftpboot/intel/mpss_u3-2.1.6720-19`**

2. Use operating system commands to copy the RPM files you downloaded to the `/tftpboot/intel/`*mpss_repository_directory* directory on the admin node.

   For example, use `cp`(1), `ftp`(1), `rsync`(1), `scp`(1), or another method.

3. Use the `crepo` command, in the following format, to specify a custom repository for the MPSS RPMs:

   `crepo --add` *rpm_repo_directory* `-custom` *rpm_repo_name*

   The variables in this command are as follows:

   - For *rpm_repo_directory*, specify the full path to the directory that contains the RPM files.
   - For *rpm_repo_name*, create a name for the image. You can specify the same name for both *rpm_repo_directory* and *rpm_repo_name*. After the image is built, the `cinstallman --show-images` command returns this label in the **Image Name** column of its output.

   For example:

```
# crepo --add /tftpboot/intel/mpss_u3-2.1.6720-19 --custom mpss_u3-2.1.6720-19
```

4. Type the following command to confirm that the MPSS image is in the correct repository:

   # **`crepo --show`**

5. Use the `crepo` command, in the following format, to add the custom repository to the generated RPM list in `/etc/opt/sgi/rpmlists`:

   `crepo --select` *rpm_repo_name*

   For *rpm_repo_name*, create a name for the image. You can specify the same name for both *rpm_repo_directory* and *rpm_repo_name*. After the image is built, the `cinstallman --show-images` command returns this label in the **Image Name** column of its output. This is the same *rpm_repo_name* that you specified in the following step:

   Procedure 3-12, step 3 on page 78

   For example, the following command adds the custom repository and displays the content of the respository:

```
# crepo --select mpss_u3-2.1.6720-19
Selecting: mpss_u3-2.1.6720-19
Updating: /etc/opt/sgi/rpmlists/generated-ice-rhel6.5.rpmlist
Updating: /etc/opt/sgi/rpmlists/generated-lead-rhel6.5.rpmlist
Updating: /etc/opt/sgi/rpmlists/generated-rhel6.5.rpmlist
```

6. Type the following command to confirm that you selected the new repository that contains the MPSS RPMs:

```
# crepo --show
* mpss_u3-2.1.6720-19 : /tftpboot/intel/mpss_u3-2.1.6720-19
* Red-Hat-Enterprise-Linux-6.5 : /tftpboot/distro/rhel6.5
* SGI-MPI-1.9-rhel6 : /tftpboot/sgi/SGI-MPI-1.9-rhel6
* SGI-Management-Center-3.0-rhel6 : /tftpboot/sgi/SGI-Management-Center-3.0-rhel6
* SGI-Foundation-Software-2.11-rhel6 : /tftpboot/sgi/SGI-Foundation-Software-2.11-rhel6
```

   The asterisk character (*) in column 1 indicates an image that is selected.

7. Type the following command to display the images that are available for installation on the compute nodes:

```
# cinstallman --show-images
Image Name              BT VCS Compat_Distro
lead-rhel6.5            0  1   rhel6
rhel6.5                 0  1   rhel6
ice-rhel6.5             1  1   rhel6
```

8. Use the `cinstallman` command, in the following format, to clone the current operating system image:

   `cinstallman --create-image --clone --source` *current_image* `--image` *new_image*

   The variables in this command are as follows:

   - For *current_image*, type the name of the operating system image you want to use that is on the system right now. Choose one that appears in the output from the from the `cinstallman --show-images` command in the **Image Name** list. For example, `ice-rhel6.5`.
   - For *new_image*, type a new name for the operating system image that is to include the MPSS file RPMs. SGI recommends that you include information in the new name that can enable you to identify this image as the operating system image that includes MPSS software. For example, `ice-compute-rhel6.5-mic-6720-19` identifies the new image as a RHEL image that contains a revision of the MPSS MIC software.

   For example:

```
# cinstallman --create-image --clone --source ice-rhel6.5 --image ice-rhel6.5-mic-6720-19
```

9. Type the following command to display the images and confirm that the new image appears in the list:

```
# cinstallman --show-images
Image Name                BT VCS Compat_Distro
lead-rhel6.5              0  1   rhel6
rhel6.5                   0  1   rhel6
ice-compute-rhel6.5       1  1   rhel6
ice-compute-rhel6.5-mic-6720-19  1 1   rhel6
```

10. Use the `cinstallman` command, in the following format, to install the `sgi-mic-compute` package and the MPSS RPMs:

    `cinstallman --yum-image --image` *image_name* `install sgi-mic-compute`

    For *image_name*, specify the *new_image* that you created in the following step:

    Procedure 3-12, step 8 on page 80

    For example:

```
# cinstallman --yum-image --image ice-rhel6.5-mic-6720-19 install sgi-mic-compute
```

11. (Conditional) Enable the SLES operating system to load the MPSS package.

    Perform this step if you are installing the MPSS packages on a SLES platform.

    Complete the following steps:

    - Open file `/etc/modprobe.d/unsupported-modules` with a text editor.
    - Add the following line at the end of the file:

      ```
      allow_unsupported_modules 1
      ```

    - Save and close the file.

12. Type the following command to display all the images:

    ```
    # cimage --list-images
    image: ice-rhel6.5
         kernel: 2.6.32-358.el6.x86_64
    image: ice-rhel6.5-mic-6720-19
         kernel: 2.6.32-358.el6.x86_64
    ```

    The preceding output shows the newly installed image, `ice-rhel6.5-mic-6720-19`.

13. Use the `cimage` command, in the following format, to set the default image for the compute nodes:

    `cimage --set-default --`*file_system* `ice-rhel6.5-mic-6720-19` *kernel*

    The variables in this command are as follows:

    - For *file_system*, type either `nfs` or `tmpfs`, according to your site practice.
    - For *kernel*, type the kernel associated with the new image, as shown in the output from the `cimage --list-images` command in the following step:

      Procedure 3-12, step 12 on page 81

    For example:

```
# cimage --set-default --tmpfs --ice-rhel6.5-mic-6720-19 2.6.32-358.el6.x86_64
```

14. Type the following command to confirm that the new image is the default image:

    ```
    # cimage --show-default
    ice-rhel6.5-stout7-mic-6720-15 2.6.32-358.el6.x86_64 tmpfs
    ```

15. Proceed to one of the following:

    - If your cluster contains MIC devices on the flat compute nodes, proceed to "Creating Images for the Flat Compute Nodes That Include MIC Devices" on page 82.
    - If your cluster does not contain MIC devices on the flat compute nodes, proceed to "Configuring the Switches" on page 86.

Creating Images for the Flat Compute Nodes That Include MIC Devices

The following procedure explains how to create images for flat compute nodes that include MIC device software.

Procedure 3-13 To create software images for flat compute nodes with MIC devices

1. On the admin node, use the `mkdir`(1) command, in the following format, to create a directory for the RPM repository:

   `mkdir -p /tftpboot/intel/`*`mpss_repository_directory`*

   For *mpss_repository_directory*, type a name for the directory that is to contain the MPSS repository. For convenience, make sure to include an identifier for the MPSS release level you downloaded.

   For example:

   # **`mkdir -p /tftpboot/intel/mpss_u3-2.1.6720-19`**

2. Use operating system commands to copy the RPM files you downloaded to the `/tftpboot/intel/`*`mpss_repository_directory`* directory on the admin node.

   For example, use `cp`(1), `ftp`(1), `rsync`(1), `scp`(1), or another method.

3. Use the `crepo` command, in the following format, to specify a custom repository for the MPSS RPMs:

   `crepo --add` *`rpm_repo_directory`* `-custom` *`rpm_repo_name`*

   The variables in this command are as follows:

   - For *rpm_repo_directory*, specify the full path to the directory that contains the RPM files.
   - For *rpm_repo_name*, create a name for the image. You can specify the same name for both *rpm_repo_directory* and *rpm_repo_name*. After the image is built,

the `cinstallman --show-images` command returns this label in the **Image Name** column of its output.

For example:

```
# crepo --add /tftpboot/intel/mpss_u3-2.1.6720-19 --custom mpss_u3-2.1.6720-19
```

4. Type the following command to confirm that the MPSS image is in the correct repository:

   ```
   # crepo --show
   ```

5. Use the `crepo` command, in the following format, to add the custom repository to the generated RPM list in `/etc/opt/sgi/rpmlists`:

   `crepo --select` *rpm_repo_name*

   For *rpm_repo_name*, create a name for the image. You can specify the same name for both *rpm_repo_directory* and *rpm_repo_name*. After the image is built, the `cinstallman --show-images` command returns this label in the **Image Name** column of its output. This is the same *rpm_repo_name* that you specified in the following step:

   Procedure 3-13, step 3 on page 82

   For example, the following command adds the custom repository and displays the content of the respository:

```
# crepo --select mpss_u3-2.1.6720-19
Selecting: mpss_u3-2.1.6720-19
Updating: /etc/opt/sgi/rpmlists/generated-ice-rhel6.5.rpmlist
Updating: /etc/opt/sgi/rpmlists/generated-lead-rhel6.5.rpmlist
Updating: /etc/opt/sgi/rpmlists/generated-rhel6.5.rpmlist
```

6. Type the following command to confirm that you selected the new repository that contains the MPSS RPMs:

```
# crepo --show
* mpss_u3-2.1.6720-19 : /tftpboot/intel/mpss_u3-2.1.6720-19
* Red-Hat-Enterprise-Linux-6.5 : /tftpboot/distro/rhel6.5
* SGI-MPI-1.9-rhel6 : /tftpboot/sgi/SGI-MPI-1.9-rhel6
* SGI-Management-Center-3.0-rhel6 : /tftpboot/sgi/SGI-Management-Center-3.0-rhel6
* SGI-Foundation-Software-2.11-rhel6 : /tftpboot/sgi/SGI-Foundation-Software-2.11-rhel6
```

The asterisk character (*) in column 1 indicates an image that is selected.

7. Type the following command to display the images that are available for installation on the flat compute nodes:

```
# cinstallman --show-images
Image Name               BT VCS Compat_Distro
lead-rhel6.5             0  1   rhel6
rhel6.5                  0  1   rhel6
ice-rhel6.5              1  1   rhel6
```

8. Use the `cinstallman` command, in the following format, to clone the current operating system image:

   `cinstallman --create-image --clone --source` *current_image* `--image` *new_image*

   The variables in this command are as follows:

   - For *current_image*, type the name of the operating system image you want to use. Choose one that appears in the output from the `cinstallman --show-images` command in the **Image Name** list. For example, `rhel6.5`.
   - For *new_image*, type a new name for the operating system image that is to include the MPSS file RPMs. SGI recommends that you include information in the new name that can enable you to identify this image as the operating system image that includes MPSS software. For example, `rhel6.5-mic-6720-19` identifies the new image as a RHEL image that contains a revision of the MPSS MIC software.

   For example:

```
# cinstallman --create-image --clone --source rhel6.5 --image rhel6.5-mic-6720-19
```

9. Type the following command to display the images and confirm that the new image appears in the list:

```
# cinstallman --show-images
Image Name               BT VCS Compat_Distro
lead-rhel6.5             0  1   rhel6.5
rhel6.5                  0  1   rhel6.5
ice-rhel6.5              1  1   rhel6.5
rhel6.5-mic-6720-19      1  1   rhel6.5
```

10. Use the `cinstallman` command, in the following format, to install the `sgi-mic-service` package and the MPSS RPMs:

    `cinstallman --yum-image --image` *image_name* `install sgi-mic-service`

    For *image_name*, specify the *new_image* that you created in the following step:

    Procedure 3-13, step 8 on page 84

    For example:

```
# cinstallman --yum-image --image rhel6.5-mic-6720-19 install sgi-mic-service
```

11. (Conditional) Enable the SLES operating system to load the MPSS package.

    Perform this step if you are installing the MPSS packages on a SLES platform.

    Complete the following steps:

    - Open file `/etc/modprobe.d/unsupported-modules` with a text editor.
    - Add the following line at the end of the file:

      ```
      allow_unsupported_modules 1
      ```

    - Save and close the file.

12. Type the following command to display all the images:

    ```
    # cimage --list-images
    image: ice-rhel6.5
         kernel: 2.6.32-358.el6.x86_64
    image: ice-rhel6.5-mic-6720-19
         kernel: 2.6.32-358.el6.x86_64
    image: rhel6.5-mic-6720-19
         kernel: 2.6.32-358.el6.x86_64
    ```

    The preceding output shows the newly installed image, `rhel6.5-mic-6720-19`.

13. Use the `cimage` command, in the following format, to set the default image for the flat compute nodes:

    `cimage --set-default --`*file_system* `rhel6.5-mic-6720-19` *kernel*

    The variables in this command are as follows:

    - For *file_system*, type either `nfs` or `tmpfs`, according to your site practice.

- For *kernel*, type the kernel associated with the new image, as shown in the output from the `cimage --list-images` command in the following step:

  Procedure 3-13, step 12 on page 85

For example:

```
# cimage --set-default --tmpfs --rhel6.5-mic-6720-19 2.6.32-358.el6.x86_64
```

14. Type the following command to confirm that the new image is the default image:

```
# cimage --show-default
rhel6.5-stout7-mic-6720-15 2.6.32-358.el6.x86_64 tmpfs
```

15. Proceed to the following:

"Configuring the Switches" on page 86

Configuring the Switches

SGI clusters have both management switches and InfiniBand (IB) switches. The individual switches are paired into *switch stacks*, and there are two switches per stack. In each stack, the top switch is typically the master switch, and the bottom switch is typically the slave switch. Although a switch stack actually includes two switches, most documentation refers to a switch stack as a *switch*.

An SGI cluster is equipped with the following types of switches:

- Spine switches. A spine switch is the primary management switch or the primary IB switch. There is one primary management switch and one primary IB switch.
- Leaf switches. A leaf switch is a secondary management switch or a secondary IB switch. There can be many leaf switches configured as part of a cluster system.

The `discover` command initializes and configures the system components for the cluster. The switch configuration procedures explain how to use the `discover` command to configure the cluster's management switches first. After you configure the management switches, if you have MCells, you configure the cooling equipment on the MCell network's switch ports.

The cluster configuration requires that the same IP address be assigned to the cluster's head gateway and to the first management switch, usually `mgmtswitch0`. The procedures in this topic assume that you want to use the default IP address,

which is 172.23.0.254, for both components, but the procedures include example commands that show how to configure an alternate IP address.

Proceed to the following to familiarize yourself with the cluster definition file's purpose:

"About the Cluster Definition File" on page 87

About the Cluster Definition File

A cluster definition file specifies media access control (MAC) addresses, IP addresses, node roles, hostnames, and other information for the cluster components. Cluster configuration, and especially switch configuration, can proceed much more quickly if you have a cluster definition file. For new clusters, you can obtain a cluster definition file from your SGI representative. For clusters that are configured, you can type the following command to generate a cluster configuration file:

```
discover --show-configfile > name
```

You can store the cluster definition file in any directory. When you specify the cluster definition file as input to the `discover` command or the cluster configuration tool.

A cluster definition file contains network and node information, and you can specify this file as input to the `discover` command. Among other things, the cluster definition file shows the media access control (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.

Example 1. This example cluster definition file is for an SGI ICE X cluster that includes one SGI ICE compute rack and several flat compute nodes. The following information highlights some characteristics of this cluster:

- The `temponame` field and the `hostname1` field appear in bold print in this example. The `temponame` field can contain the hostname of the node or it can contain a label for the role of the component in the cluster. The `temponame` field is used by SMC internal operations. The `hostname1` field defines the hostname for the component, and it is this hostname that users need to specify when they want to log into the node. For example, if you configure user services on the two of the flat compute nodes in this cluster, users can log into the cluster by logging into `n0` or `n1`.
- The flat compute nodes designated as compute services nodes have hostnames of `n0` and `n1`. The flat compute nodes used for computing have hostnames of `n101`,

n102, n103and so on. The flat compute nodes used for computing are configured on a routed management network called head2.

Note: Not all flat nodes are shown in the example file.

- This cluster uses a routed management network for its flat compute nodes. There is a network for the nodes themselves and a network for the BMCs on the nodes. These networks are called head2 and head2-bmc, and they are associated with mgmtsw1.

The file is as follows:

```
[discover]
temponame=r1lead, mgmt_bmc_net_name=head-bmc, mgmt_bmc_net_macs=00:25:90:58:8b:75,
mgmt_net_name=head, mgmt_net_macs=00:25:90:58:8a:94/00:25:90:58:8a:95,
redundant_mgmt_network=yes, switch_mgmt_network=yes, mic=0, dhcp_bootfile=grub2,
conserver_logging=yes, conserver_ondemand=no, console_device=ttyS1
temponame=service0, mgmt_bmc_net_name=head-bmc, mgmt_bmc_net_macs=00:25:90:58:7d:7f,
mgmt_net_name=head, mgmt_net_macs=00:25:90:58:7d:32/00:25:90:58:7d:33, hostname1=n0,
redundant_mgmt_network=yes, switch_mgmt_network=yes, mic=0, dhcp_bootfile=grub2,
conserver_logging=yes, conserver_ondemand=no, root_type=disk, console_device=ttyS1
temponame=service1, mgmt_bmc_net_name=head-bmc, mgmt_bmc_net_macs=00:25:90:58:96:a2,
mgmt_net_name=head, mgmt_net_macs=00:25:90:58:96:54/00:25:90:58:96:55, hostname1=n1
redundant_mgmt_network=yes, switch_mgmt_network=yes, mic=0, dhcp_bootfile=grub2,
conserver_logging=yes, conserver_ondemand=no, root_type=disk, console_device=ttyS1
temponame=service101, mgmt_bmc_net_name=head2-bmc, mgmt_bmc_net_macs=00:1E:67:2C:53:92,
mgmt_net_name=head2, mgmt_net_macs=00:1E:67:2C:53:8E/00:1e:67:2c:53:8f, hostname1=n101,
redundant_mgmt_network=yes, switch_mgmt_network=yes, mic=0, dhcp_bootfile=ipxe,
conserver_logging=yes, conserver_ondemand=no, root_type=disk, console_device=ttyS0
temponame=service102, mgmt_bmc_net_name=head2-bmc, mgmt_bmc_net_macs=00:1E:67:2C:58:AF,
mgmt_net_name=head2, mgmt_net_macs=00:1E:67:2C:58:AB/00:1e:67:2c:58:ac, hostname1=n102,
redundant_mgmt_network=yes, switch_mgmt_network=yes, mic=0, dhcp_bootfile=ipxe,
conserver_logging=yes, conserver_ondemand=no, root_type=disk, console_device=ttyS0
temponame=service103, mgmt_bmc_net_name=head2-bmc, mgmt_bmc_net_macs=00:1E:67:2C:54:E2,
mgmt_net_name=head2, mgmt_net_macs=00:1E:67:2C:54:DE/00:1e:67:2c:54:df, hostname1=n103,
redundant_mgmt_network=yes, switch_mgmt_network=yes, mic=0, dhcp_bootfile=ipxe,
conserver_logging=yes, conserver_ondemand=no, root_type=disk, console_device=ttyS0
.
.
.
temponame=mgmtsw0, mgmt_net_name=head, mgmt_net_macs=b4:0e:dc:38:6b:17, net=head/head-
```

```
bmc, ice=yes, type=spine
temponame=mgmtsw1, mgmt_net_name=head, mgmt_net_macs=b4:0e:dc:38:6b:18, net=head2/head2-
bmc, ice=no, type=leaf

[dns]
cluster_domain=acme.americas.sgi.com
nameserver1=137.38.225.5
nameserver2=137.38.31.248

[attributes]
dhcp_bootfile=grub2
udpcast_rexmit_hello_interval=0
udpcast_min_receivers=1
head_vlan=1
mcell_network=yes
udpcast_min_wait=10
my_sql_replication=yes
redundant_mgmt_network=yes
max_rack_irus=16
udpcast_max_bitrate=900m
udpcast_max_wait=10
udpcast_mcast_rdv_addr=224.0.0.1
rack_vlan_end=1100
switch_mgmt_network=yes
mcell_vlan=3
mic=0
conserver_logging=yes
rack_vlan_start=101
conserver_ondemand=no
blademond_scan_interval=120

[networks]
name=private, subnet=172.26.0.0, netmask=255.255.0.0
name=public, subnet=137.38.82.0, netmask=255.255.255.0, gateway=137.38.82.254
name=head, type=mgmt, vlan=1, subnet=172.23.0.0, netmask=255.255.0.0, gateway=172.23.255.254
name=head-bmc, type=mgmt-bmc, vlan=1, subnet=172.24.0.0, netmask=255.255.0.0
name=mcell-net, type=cooling, subnet=172.26.0.0, netmask=255.255.0.0
name=ha-net, type=ha, subnet=192.168.161.0, netmask=255.255.255.0
name=ib-0, type=ib, subnet=10.148.0.0, netmask=255.255.0.0
name=ib-1, type=ib, subnet=10.149.0.0, netmask=255.255.0.0
name=gbe, type=lead-mgmt, subnet=10.159.0.0, netmask=255.255.0.0, rack_netmask=255.255.252.0
```

```
name=bmc, type=lead-bmc, subnet=10.160.0.0, netmask=255.255.0.0, rack_netmask=255.255.252.0
name=head2, type=mgmt, vlan=2001, subnet=172.54.0.0, netmask=255.255.0.0, gateway=172.54.255.254
name=head2-bmc, type=mgmt-bmc, vlan=2001, subnet=172.99.0.0, netmask=255.255.0.0,
gateway=172.99.255.254
```

Example 2. This example cluster definition file is for an SGI Rackable cluster with 100 flat compute nodes. For simplicity's sake, the example file shows only two flat compute services nodes and the management switches. The following information highlights some characteristics of this cluster:

- The information in the `temponame` field defines the role for each of the two flat compute nodes in this cluster. The content of the `temponame` field and the `hostname1` field can be identical; in other words, you can use the node's hostname as its `temponame`.

  The content of the `temponame` field for each flat compute node is `service`*n*, where *n* is a number from `1` through `101`.

- The `hostname1` field defines the hostname that users need to specify when they want to log into a node. The text in the `hostname1` field is the text that appears in the output for most SMC commands when the command generates output. In this example, the hostnames for the two nodes shown are `n1` and `n101`. Node `n1` uses the default head management network, `head`. Node `n101` uses a routed management network, `head2`.

- The cluster definition file specifies a multicast installation that uses `udpcast` transport for the flat compute nodes, `service1` and `service101`.

- The top-level switch, `mgmtsw0`, is defined as `spine` switch and serves the `head` network. Switch `mgmtsw1` is defined as a `leaf` switch and serves the routed management network, `head2`.

- The definition for both switches includes `ice=no` because this cluster has no SGI ICE X components.

The file is as follows:

```
[discover]
temponame=service1, mgmt_bmc_net_name=head-bmc, mgmt_bmc_net_macs=00:25:90:1A:6D:3E,
mgmt_net_name=head, mgmt_net_macs=00:25:90:1A:AC:D0/00:25:90:1a:ac:d1, hostname1=n1,
redundant_mgmt_network=yes, switch_mgmt_network=yes, mic=0, dhcp_bootfile=grub2,
conserver_logging=yes, conserver_ondemand=no, root_type=disk,
console_device=ttyS1,transport=udpcast
```

```
...

temponame=service101, mgmt_bmc_net_name=head2-bmc, mgmt_bmc_net_macs=00:1E:67:2C:53:92,
mgmt_net_name=head2, mgmt_net_macs=00:1E:67:2C:53:8E, hostname1=n101,
redundant_mgmt_network=yes, switch_mgmt_network=yes, mic=0, dhcp_bootfile=ipxe,
conserver_logging=yes, conserver_ondemand=no, root_type=disk,
console_device=ttyS0,transport=udpcast
temponame=mgmtsw0, mgmt_net_name=head,mgmt_net_macs=00:26:F3:C3:7A:40, net=head/head-
bmc, ice=no, type=spine
temponame=mgmtsw1, mgmt_net_name=head,mgmt_net_macs=00:04:96:97:C0:78,
net=head2/head2-bmc, ice=no, type=leaf

[dns]
cluster_domain=smc-default.americas.sgi.com
nameserver1=128.162.236.210
nameserver2=128.162.236.211
nameserver3=198.149.32.11

[attributes]
dhcp_bootfile=grub2
udpcast_min_receivers=1
head_vlan=1
mcell_network=yes
udpcast_min_wait=10
my_sql_replication=yes
redundant_mgmt_network=yes
max_rack_irus=16
udpcast_max_bitrate=900m
udpcast_max_wait=10
rack_vlan_end=1100
switch_mgmt_network=yes
mcell_vlan=3
mic=0
conserver_logging=yes
rack_vlan_start=101
conserver_ondemand=no
blademond_scan_interval=120

[networks]
name=private, subnet=172.26.0.0, netmask=255.255.0.0
name=public, subnet=128.162.243.0, netmask=255.255.255.0, gateway=128.162.243.1
```

```
name=head, type=mgmt, vlan=1, subnet=172.23.0.0, netmask=255.255.0.0, gateway=172.23.255.254
name=head-bmc, type=mgmt-bmc, vlan=1, subnet=172.24.0.0, netmask=255.255.0.0
name=mcell-net, type=cooling, subnet=172.26.0.0, netmask=255.255.0.0
name=ha-net, type=ha, subnet=192.168.161.0, netmask=255.255.255.0
name=ib-0, type=ib, subnet=10.148.0.0, netmask=255.255.0.0
name=ib-1, type=ib, subnet=10.149.0.0, netmask=255.255.0.0
name=gbe, type=lead-mgmt, subnet=10.159.0.0, netmask=255.255.0.0, rack_netmask=255.255.252.0
name=bmc, type=lead-bmc, subnet=10.160.0.0, netmask=255.255.0.0, rack_netmask=255.255.252.0
name=head2, type=mgmt, vlan=2001, subnet=172.98.0.0, netmask=255.255.0.0,
gateway=172.98.255.254
name=head2-bmc, type=mgmt-bmc, vlan=2001, subnet=172.99.0.0, netmask=255.255.0.0,
gateway=172.99.255.254
```

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

Procedure	Circumstance
"Verifying the Switch Cabling" on page 92	All switch configuration circumstances. Perform this procedure regardless of the presence of a cluster definition file or MCells.
"Configuring Management Switches With a Cluster Definition File" on page 95	If you have a cluster definition file.
"Configuring Management Switches Without a Cluster Definition File" on page 99	If you do not have a cluster definition file.
"(Conditional) Configuring the Cooling Racks and Cooling Distribution Units (CDUs) on the MCell Network's Switch Ports" on page 103	(Conditional) If you have MCells. This extra procedure configures the MCell switches separately from the rest of the cluster switches.

Verifying the Switch Cabling

The following figures show example switch cabling. Depending on the switch configuration procedure you use, you might have to plug cables out from and into switch ports during the configuration process. Regardless of the procedure you use, the cables on your switches eventually become cabled as shown in the following figures.

Figure 3-10 on page 93 shows a switch stack with two switches. In this switch stack, the two switches constitute the spine switch stack. One is the master switch and the other is the slave switch.

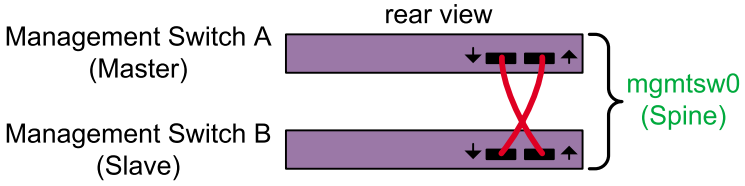

Figure 3-10 Spine Switch Stack With Two Switches

Figure 3-11 on page 93 shows a switch stack with multiple switches. The first two switches constitute the spine switch stack, and the other switches constitute the secondary switch stack.

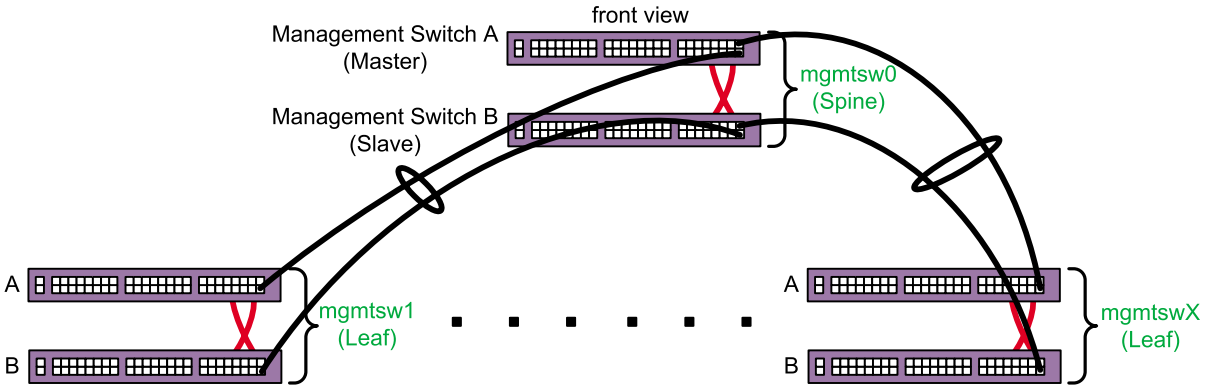

Figure 3-11 Switch Stack With Multiple Switches

The following procedure explains how to inspect your switches and prepare for the configuration procedure.

Procedure 3-14 To verify your switches

1. Visually inspect your system.

Note the types of switches you have and their identifiers. At a minimum, you have one spine switch stack. The admin node connects to the master switch in the spine switch stack. You might have additional leaf switch stacks.

Within each stack, each switch is labeled MSW*xx*. In the spine switch stack, the master switch is labeled MSW0A, and the slave switch is labeled MSW0B. In the first leaf switch stack, the master switch is labeled MSW1A, and the slave switch is labeled MSW1B. The A and B on the switch labels identify the master switch and slave switch in the stack. The `switchconfig set` command operates on a switch stack (not just one switch), so you need to note only the characters on the label that precede the A and B when you provide information to the `switchconfig set` command. Use the following table to determine the value you need to use for *switch* on the `switchconfig` commands:

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

2. Make sure that only the admin node is plugged in and that all rack circuit breakers are powered off.

   If you have a cluster definition file, connect all other nodes and switches to a power source, but do not power them on. That is, make sure that all chassis management controllers (CMCs) on the rack leader controllers (RLCs), all RLCs, all SGI ICE compute nodes, all flat compute nodes, all switches, and so on, are not powered on.

   If you do not have a cluster definition file, unplug all nodes and switches other than the admin node. The admin node should be the only component that is plugged in.

3. (Conditional) Remove the cascade cables that connect the slaves switches together.

   Perform this step only if you have two or more switch stacks.

   Locate the cascading cables that connect the switch stacks to each other, and unplug the cascading cable end from the lower switch in the neighboring switch stack. On the front of the switch stack, unplug the cascading cables that plug into the ports on the slave switches. When you unplug one end of each cable, you prevent a switching loop.

Note: Do not unplug the stacking cables in the rear of the switch stack. The installation procedure instructs you to plug or unplug only the cascading cables in the front of the switch stack.

   For example, if you have one leaf switch stack, locate the cascading cable that runs from the slave switch on the spine switch to the slave switch on the leaf switch. Unplug that cable from the port on the slave switch.

4. Proceed as follows:

   - If you have a cluster definition file, proceed to the following:

     "Configuring Management Switches With a Cluster Definition File" on page 95

   - If you do not have a cluster definition file, proceed to the following:

     "Configuring Management Switches Without a Cluster Definition File" on page 99

Configuring Management Switches With a Cluster Definition File

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

Procedure 3-15 To configure switches — with a cluster definition file

1. Through an `ssh` connection, log in as root to the admin node, and write the cluster definition file to a location on the admin node.

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

2. Plug in all the switches.

3. (Conditional) Specify a site-specific IP address for the head gateway.

Perform this step if you need to set a site-specific, nondefault IP address for the spine switch stack (head gateway).

By default, SGI sets the IP address for the spine switch stack (head gateway) to 172.23.0.254. Type the `cadmin` command in the following format to set a site-specific IP address:

`cadmin -set-head-gateway` *IP_addr*

For *IP_addr*, specify your site-specific IP address.

4. Use the `discover` command, in the following format, to configure the spine switch stack (the switches labeled `MSW0A` and `MSW0B`), which is attached to the admin node:

   `discover --mgmtswitch 0 --configfile` *path*

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

   For example:

   # **`discover --mgmtswitch 0 --configfile /var/tmp/config_file`**

   This step assigns an IP address to the spine switch stack. The spine switch stack becomes the head gateway.

5. (Conditional) Type additional `discover` commands for each secondary switch stack.

   Complete this step if you have more than one switch.

   The formats for these additional commands are as follows:

   `discover --mgmtswitch` *num* `--configfile` *path*

   The arguments are as follows:

Argument	Specification
num	The identifier for the switch stack.
path	The full path to the location of the cluster definition file.

   For example:

   # **`discover --mgmtswitch 1 --configfile /var/tmp/mac_file`**

6. (Conditional) On the front of the switch stack, plug the cascading cables into the ports on the slave switches.

   Perform this step only if you have two or more switch stacks.

   This step is the opposite of the following step, in which you unplugged the cascading cables from the ports on the slave switches:

   Procedure 3-14, step 3 on page 95

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

   ```
   # cnodes --mgmt-switch
   ```

8. Use the `ssh`(1) command to open a console window to the admin node.

9. Type the following command, and monitor the power-on process in the admin node's console window:

   ```
   # tail -f /var/log/cmcdetectd.log
   ```

10. Flip the power breakers on the cluster's CMCs, one rack at a time.

    The `cmcdetectd` daemon runs on the admin node. It configures the top level switches so that the CMCs are on the appropriate rack VLAN. After you power on rack one, the `cmcdetectd` daemon detects the rack and adds the rack to the switch. After the CMCs for rack one appear on the switch, power on the CMCs for rack two.

11. Use the `switchconfig` command, in the following format, to list the CMCs associated with each switch:

    `switchconfig list -s` *switch*

    For *switch*, specify the system ID for a switch, for example `mgmtsw0`, `mgmtsw1`, and so on.

    Issue one of these commands for each switch in your configuration and examine the output. If the output includes all the CMCs in your SGI ICE system, you can proceed with your configuration. If the output contains errors or does not include all expected CMCs, contact your SGI representative for troubleshooting information.

12. Save the configuration to the nonvolatile memory (flash) on the switches.

Note: This step is very important. In the event of a power outage or other interruption, the switch stack boots with the saved configuration.

Type the `switchconfig` command in the following format:

```
switchconfig save_running_config -s mgmtsw0[,mgmtsw1,mgmtsw2,...]
```

Include the parameters `mgmtsw1`, `mgmtsw2`, and so on, only if there are switches in addition to the spine switch (`mgmtsw0`).

13. Use the `switchconfig` command, in the following format, to back up the switch configuration to a file on the admin node:

    `switchconfig pull_switch_config -s` *switch_ID* `-f` *file* `[--debug]`

    The arguments are as follows:

Argument	Specification
switch_ID	The switch system ID. For example, this could be `mgmtsw0`, `mgmtsw1`, `mgmtsw2`, and so on. The output from the `cnodes` command includes this information.
file	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.

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

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

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

    Issue one of these commands per switch.

    In the future, if you need to replace the switch, you can save configuration time if you push out this configuration file from the admin node to the new switch.

14. After all management switches have been configured, proceed as follows::

    - If you have MCells, proceed to the following:

"(Conditional) Configuring the Cooling Racks and Cooling Distribution Units (CDUs) on the MCell Network's Switch Ports" on page 103

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

  "Configuring the Cluster With the `discover` Command" on page 106

Configuring Management Switches Without a Cluster Definition File

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

Procedure 3-16 To configure switches — without a cluster definition file

1. Through an `ssh` connection, log in to the admin node as the root user.
2. (Conditional) Specify a site-specific IP address for the head gateway.

   Perform this step if you need to set a site-specific, nondefault IP address for the spine switch stack (head gateway).

   By default, SGI sets the IP address for the spine switch stack (head gateway) to 172.23.0.254. Type the following command to set a site-specific IP address:

   `cadmin -set-head-gateway` *IP_addr*

   For *IP_addr*, specify your site-specific IP address.
3. Type the `discover` command in one of the following formats to configure the spine switch:

   - On SGI ICE X platforms, type the following `discover` command:

     # **`discover --mgmtswitch 0,ice=yes,type=spine`**

   - On SGI Rackable platforms, type the following `discover` command:

     # **`discover --mgmtswitch 0,ice=no,type=spine`**

   Notes:

   - When you specify the `ice=yes` parameter, you indicate to the `discover` command that the cluster includes rack leader controllers and is, therefore, an SGI ICE X cluster.

- The `type=`*type* parameter specifies whether the switch being configured is the spine switch or one of the leaf switches. If you do not specify a `type` parameter, and you are configuring a management switch, the `discover` command uses the link layer discovery protocol (LLDP) to attempt to determine the switch that is directly connected to the admin node.

4. When prompted, connect the spine switch stack to a power source.

   To complete this step, plug in the master switch and then the slave switch so that the entire spine switch stack is powered up. In this way, the master switch boots just a few seconds before the slave switch.

   The `discover` command configures the MAC address of the switch after you connect the spine switch stack to a power source.

5. (Conditional) Plug in the switch when prompted, and type a `discover` command to configure each leaf switch stack.

   Perform this step if you have leaf switch stacks.

   Complete the following steps:

1. Plug in the switch stack when the system prompts you to do so.
2. Type a `discover` command, in the following format, for your platform::

   - On SGI ICE X platforms, type one or more `discover` commands in the following format:

     # **discover --mgmtswitch *num*,ice=yes,type=leaf**

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

     For example:

     # **discover --mgmtswitch 1,ice=yes,type=leaf**

   - On SGI Rackable platforms, type one or more `discover` commands in the following format:

     # **discover --mgmtswitch *num*,ice=no,type=leaf**

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

     For example:

     # **discover --mgmtswitch 1,ice=no,type=leaf**

3. (Conditional) On the front of the switch stack, plug the cascading cables into the ports on the slave switches.

   Perform this step only if you have leaf switch stacks.

   This step is the opposite of the following step, in which you unplugged the cascading cables from the ports on the slave switches:

   Procedure 3-14, step 3 on page 95

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

   # **cnodes --mgmtswitch**

5. Use the `ssh`(1) command to open a console window to the admin node.

6. Type the following command, and monitor the power-on process in the admin node's console window:

   # **tail -f /var/log/cmcdetectd.log**

7. (SGI ICE X clusters only) Flip the power breakers on the cluster's CMCs, one rack at a time.

   The `cmcdetectd` daemon runs on the admin node. It configures the top level switches so that the CMCs are on the appropriate rack VLAN. After you power on rack one, the `cmcdetectd` daemon detects the rack and adds the rack to the switch. After the CMCs for rack one appear on the switch, power on the CMCs for rack two.

8. (SGI ICE X clusters only) Use the `switchconfig` command, in the following format, to list the CMCs associated with each switch:

   `switchconfig list -s` *switch*

   For *switch*, specify the system ID for a switch, for example `mgmtsw0`, `mgmtsw1`, and so on.

   Issue one of these commands for each switch in your configuration and examine the output. If the output includes all the CMCs in your cluster, you can proceed with your configuration. If the output contains errors or does not include all expected CMCs, contact your SGI representative for troubleshooting information.

9. Save the configuration to the nonvolatile memory (flash) on the switches.

Note: This step is very important. In the event of a power outage or other interruption, the switch stack boots with the saved configuration.

Type the `switchconfig` command in the following format:

```
# switchconfig save_running_config -s mgmtsw0[,mgmtsw1,mgmtsw2,...]
```

Include the parameters `mgmtsw1`, `mgmtsw2`, and so on, only if there are switches in addition to the spine switch (`mgmtsw0`).

10. For each switch you configured, use the `switchconfig` command, in the following format, to back up the switch configuration to a file on the admin node:

    `switchconfig pull_switch_config -s` *switch_ID* `-f` *file* `[--debug]`

    The arguments are as follows:

Argument	Specification
switch_ID	The switch system ID. For example, this could be `mgmtsw0`, `mgmtsw1`, `mgmtsw2`, and so on. The output from the `cnodes` command includes this information.
file	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.

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

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

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

    Issue one of these commands per switch.

    In the future, if you need to replace a switch, you can save configuration time if you push out this configuration file from the admin node to the new switch.

11. Use the `discover` command, in the following format, to save the MAC addresses to a cluster definition file:

    # **`discover --show-configfile > path`**

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

    You can use this cluster definition file if you ever have to configure the switches again.

12. After all management switches have been configured, proceed as follows:

    - If you have MCells, proceed to the following:

      "(Conditional) Configuring the Cooling Racks and Cooling Distribution Units (CDUs) on the MCell Network's Switch Ports" on page 103

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

      "Configuring the Cluster With the `discover` Command" on page 106

(Conditional) Configuring the Cooling Racks and Cooling Distribution Units (CDUs) on the MCell Network's Switch Ports

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

A cluster contains 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, "SGI ICE X MCell Network IP Addresses" on page 189

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

Procedure 3-17 To configure MCell switches

1. Gather information about the MCell switches in your cluster.

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

2. Log in as the root user to the admin node.

3. Type following command to retrieve information about the virtual local area networks (VLANs) that are configured at this time:

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

   The preceding output shows that the MCell VLAN is 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 -b none -d` *vlan_num* `-p` *ports* `-s` *switch*

   Type an individual `switchconfig set` command for each switch on the cluster 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. Make sure that the labels on the cables correspond to the labels on the switch ports.

Example 1. The following command configures 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 -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 -b none -d 3 -p 2/31,2/32,2/33 -s mgmtsw0
```

Note: 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
```

---

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

   Procedure 3-17, step 4 on page 104

   If you encounter errors, issue a `switchconfig set` command again.

6. Save the configuration to the nonvolatile memory (flash) on the switches.

Note: This step is very important. In the event of a power outage or other interruption, the switch stack boots with the saved configuration.

Type the `switchconfig` command in the following format:

```
switchconfig save_running_config -s mgmtsw0[,mgmtsw1,mgmtsw2,...]
```

Include the parameters `mgmtsw1`, `mgmtsw2`, and so on, only if there are switches in addition to the spine switch (`mgmtsw0`).

7. Type the following command to back up the switch configuration to a file on the admin node:

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

The arguments are as follows:

Argument	Specification
switch_ID	The switch system ID. For example, this could be `mgmtsw0`, `mgmtsw1`, `mgmtsw2`, and so on. The output from the `cnodes` command includes this information.
file	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 admin node:

```
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`.

In the future, if you need to replace the switch, you can save configuration time if you push out this configuration file from the admin node to the new switch.

8. After all switches have been configured, proceed to the following:

"Configuring the Cluster With the `discover` Command" on page 106

Configuring the Cluster With the `discover` Command

The `discover` command finds and configures all non-admin nodes and all external switches. If you have a cluster definition file, this procedure can complete more quickly. The procedure in this topic includes configuration steps that explain how to complete the procedure both with and without a cluster definition file.

The following procedure explains how to configure the RLCs (if present), the SGI ICE compute nodes (if present), the flat compute nodes, and the external switches.

Procedure 3-18 To configure the nodes and switches

1. Visually inspect your cluster and note the labels on the nodes.

   RLCs are numbered starting with 1. For example, `r1lead`, `r2lead`, and so on.

SGI ICE compute nodes are numbered starting with 0. The numbering depends on the RLC number and on the IRU number within the RLC. For example, the first blade on RLC 1, IRU 1 is numbered as `r1i0n0`, and if there are eight IRUs in the rack, the last blade on the last IRU of RLC 1 is numbered `r1i0n7`.

Flat compute nodes are numbered starting with 0. For example, `n0`, `n1`, `n2`, and so on.

2. Check the power cords on all nodes, as follows:

   - If you have a cluster definition file, make sure all 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.

   - If you do not have a cluster definition file, make sure that all nodes are unplugged from their power sources.

3. Through an `ssh` connection, log into the admin node as the root user.

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

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

   The nodes determine the integrated Ethernet devices by accepting DHCP leases that belong only to the cluster. Cluster systems use DHCP 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 belonging to the cluster's DCHP server, which can lead to an installation failure. Change this option code only if absolutely necessary.

   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) Plug in all the racks and all the flat compute nodes.

   Perform this step if you have a cluster definition file.

   A cluster definition file contains information about the cluster, including the MAC addresses for the nodes. If you use a cluster definition file, the configuration

process can complete more quickly. Contact your SGI representative to find out if a cluster definition file is available. For more information about cluster definition files, see "Configuring the Switches" on page 86.

6. Use one or more `discover` commands to configure the cluster nodes.

   Example 1. If you have a cluster definition file, use the following `discover` command:

   `discover --configfile` *path_to_CDF* `--all`

   For *path_to_CDF*, specify the full path to your cluster definition file. This command configures all the nodes that appear in the cluster definition file.

   Example 2. If you do not have a cluster definition file, or if you want to configure only selected nodes, use a `discover` command with parameters that specify each node. If you have an SGI ICE X cluster, make sure to specify the `--leader` parameter to configure the rack leader controllers (RLCs). For example, use the discover command in the following format:

`discover [--leader[set]` *ID*`[,mic=`*mic_num*`]] --node[set]` *specs*`[,mic=`*mic_num*`]`
`[,dhcp_bootfile=ipxe][--configfile` *cluster_def_file*`]`

The arguments are as follows:

Argument	Specification
ID	Used only for SGI ICE X clusters. Specifications used to configure the SGI ICE compute nodes, including the ID number(s) for the rack(s) that you want to configure. For example, if you want to configure one rack for an SGI ICE X cluster, specify `--leader` and the system ID number that corresponds to that rack. If you want to configure a range of racks, specify `--leaderset`, the starting system ID number, a comma (`,`), and the ending system ID number. Examples: `--leader 2` Configures RLC 2. The RLC is configured with a `temponame` of `r2lead`.

`--leaderset 1,3` Configures RLCs 1, 2, and 3. The RLCs are configured with the following `temponames`: `r1lead`, `r2lead`, `r3lead`.

mic_num Specify the number of Intel® Many Integrated Core (MIC) devices that reside on each node. By default, the `discover` command assumes zero (0). If you have MIC devices on any SGI ICE compute nodes, specify the number you have, which can be `1` or `2`. For flat compute nodes, *mic_num* can be `1`, `2`, `3`, or `4`.

Specify the `,mic=`*mic_num* parameter only if your cluster includes MIC devices.

specs Specifications used to configure the flat compute nodes. After the cluster is completely configured, you can configure services, for example DNS or Lustre, on one or more of these nodes.

To configure only one node, specify the `--node` parameter. To configure a series of nodes, specify the `--nodeset` parameter.

Examples:

`--node 2`

Configures flat compute node 2.

The node is configured with a `temponame` of `service2` and a hostname of `n2`.

`--nodeset 1,3`

Configures flat compute nodes 1, 2, and 3.

The nodes are configured with the following `temponames`: `service1`, `service2`, and `service3`.

The nodes are configured with the following hostnames: `n1`, `n2`, and `n3`.

`--node 200,100,hostname1=snXXX`

Configures flat compute nodes 200 through 299.

The nodes are configured with the following temponames: `service200`, `service201`, and so on up to `service299`.

The nodes are configured with the following hostnames: `sn200`, `sn201`, and so on up to `sn299`.

Use the `dhcp_bootfile=ipxe` parameter in troubleshooting situations. If you already issued a `discover` command and one or more nodes failed to boot, specify the `dhcp_bootfile=ipxe` parameter, which directs the server boot agent to load iPXE rather than GRUB version 2. When this parameter is used, the iPXE software loads GRUB version 2.

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

Example 1. The following command uses a cluster definition file to configure rack 1 and flat compute node 0:

```
# discover --leader 1 --node 0
```

Example 2. If you have one rack of SGI ICE compute nodes and one flat compute node, type the following command:

```
# discover --leader 1 --node 0
```

Example 3. If you have five racks of SGI ICE compute nodes and three flat compute nodes, type the following command:

```
# discover --leaderset 1,5 --nodeset 1,3
```

Example 4. If you have one rack of SGI ICE compute nodes, one flat compute node, two MIC devices attached to each blade on the SGI ICE X compute node rack, and two MIC devices attached to the flat compute node, type the following command:

```
# discover --leader 1,mic=2 --node 0,mic=2,image=mic_serv_image
```

In example 4, for *mic_serv_image*, specify the label (name) of the flat compute node image that you created in "Creating Images for the Flat Compute Nodes That Include MIC Devices" on page 82.

7. (Conditional) When prompted to do so by the system, plug in each individual rack or flat compute node.

Perform this step if you did not use a cluster definition 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 compute 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 commands to update the configuration files:

   # **update-configs**

9. Type the following commands to save the configuration to the nonvolatile memory (NVM) on the switches:

   ```
   switchconfig save_running_config -s mgmtsw0[,mgmtsw1,mgmtsw2,...]
   ```

   Include the parameters `mgmtsw1`, `mgmtsw2`, and so on, only if there are switches in addition to management switch 0.

10. (Conditional) Confirm the status of the MIC devices.

    Perform this step if your cluster has MIC devices.

    Complete the following steps:

    - Type the following command to make sure that the MIC devices came online:

      # **cexec --all 'micctrl --status'**

      The following example output shows that the MIC devices came online correctly:

      ```
      ************************ rack_1 ************************
      ************************ rack_1 ************************
      --------- r1i0n6---------
      mic0: online (mode: linux image: /lib/firmware/mic/uos.img)
      mic1: online (mode: linux image: /lib/firmware/mic/uos.img)
      --------- r1i0n7---------
      mic0: online (mode: linux image: /lib/firmware/mic/uos.img)
      mic1: online (mode: linux image: /lib/firmware/mic/uos.img)
      ```

```
--------- r1i0n8---------
mic0: online (mode: linux image: /lib/firmware/mic/uos.img)
mic1: online (mode: linux image: /lib/firmware/mic/uos.img)
```

- Type the following command to verify the IP addresses and MTU size of the MIC devices:

  # **cexec --all 'micctrl --config | grep -E \"MIC IP|Bits|MtuSize\"'**

  The following example output shows that all the MIC devices have addresses on the 10.157.1.0/24 and the 10.158.1.0/24 networks and that the MIC devices use the correct MTU size of 9000:

```
************************ rack_1 ************************
************************ rack_1 ************************
--------- r1i0n4---------
        MIC IP:    10.157.1.6
        Net Bits:  24
        MtuSize:   9000
        MIC IP:    10.158.1.6
        Net Bits:  24
        MtuSize:   9000
--------- r1i0n5---------
        MIC IP:    10.157.1.7
        Net Bits:  24
        MtuSize:   9000
        MIC IP:    10.158.1.7
        Net Bits:  24
        MtuSize:   9000
--------- r1i0n6---------
        MIC IP:    10.157.1.8
        Net Bits:  24
        MtuSize:   9000
        MIC IP:    10.158.1.8
        Net Bits:  24
        MtuSize:   9000
--------- r1i0n7---------
        MIC IP:    10.157.1.9
        Net Bits:  24
        MtuSize:   9000
        MIC IP:    10.158.1.9
        Net Bits:  24
```

```
        MtuSize:   9000
--------- r1i0n8---------
        MIC IP:    10.157.1.10
        Net Bits:  24
        MtuSize:   9000
        MIC IP:    10.158.1.10
        Net Bits:  24
        MtuSize:   9000
```

11. Proceed as follows:

    - 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 113

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

      "(Conditional) Configuring the InfiniBand Subnetworks" on page 114

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

Typically, the DNS on the admin node provides name services for the cluster. When you configure a backup DNS, however, the cluster can use a flat compute node as a secondary DNS server if the admin node 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 flat compute node to act as a DNS.

Procedure 3-19 To enable a backup DNS

1. Through an `ssh` connection, log into the admin node as the root user.
2. Type the following command to retrieve a list of available flat compute nodes:

   # **cnodes --compute**

   The flat compute 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 flat compute node.

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 flat compute node that you want to designate as the backup DNS, and select **OK**.

   For example, you could configure flat compute node `n101` as the host for the backup DNS server.

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

6. (Conditional) Proceed to the following:

   "(Conditional) Configuring the InfiniBand Subnetworks" on page 114

(Conditional) Configuring the InfiniBand Subnetworks

Perform the procedures in this topic as follows:

- If you have an SGI ICE X cluster. On this platform, you can configure the InfiniBand subnetwork either on a rack leader controller (RLC) or on a flat compute node.
- If you have an SGI Rackable cluster, and you want to configure the InfiniBand subnetwork on one of the flat compute nodes. Some InfiniBand switches on SGI Rackable clusters come configured for an InfiniBand subnetwork. You can perform the procedure in this topic if your switch is not preconfigured for InfiniBand or if you prefer to configure this service on a flat compute node.

The InfiniBand network on the cluster uses Open Fabrics Enterprise Distribution (OFED) software. For information about OFED, see the following website:

http://www.openfabrics.org

For more information about the InfiniBand fabric implementation on SGI clusters, see the *SGI Management Center Administration Guide for Clusters*.

Each cluster has two InfiniBand fabric network cards, `ib0` and `ib1`. Each subnetwork has a subnet manager, which runs on an RLC or on a flat compute node.

The following procedures explain how to configure the master and the standby components and how to verify the configuration:

- "Configuring the InfiniBand Subnetworks" on page 115
- "Verifying That the InfiniBand Subnetwork is Working (SGI ICE X Clusters)" on page 119
- "Verifying That the InfiniBand Subnetwork is Working (SGI Rackable Clusters)" on page 120

Configuring the InfiniBand Subnetworks

The following procedure explains how to configure the InfiniBand subnetwork master and standby components on an SGI ICE X cluster or on an SGI Rackable cluster.

Procedure 3-20 To configure the InfiniBand subnetworks

1. Through an `ssh` connection, log into the admin node 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:

     ```
     # /opt/sgi/sbin/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 component identifiers for the master (primary) and the standby (backup, secondary) subnetwork, and select **Select**.

   Example 1. On an SGI ICE X cluster, 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.

   Figure 3-12 on page 117 shows a completed screen.

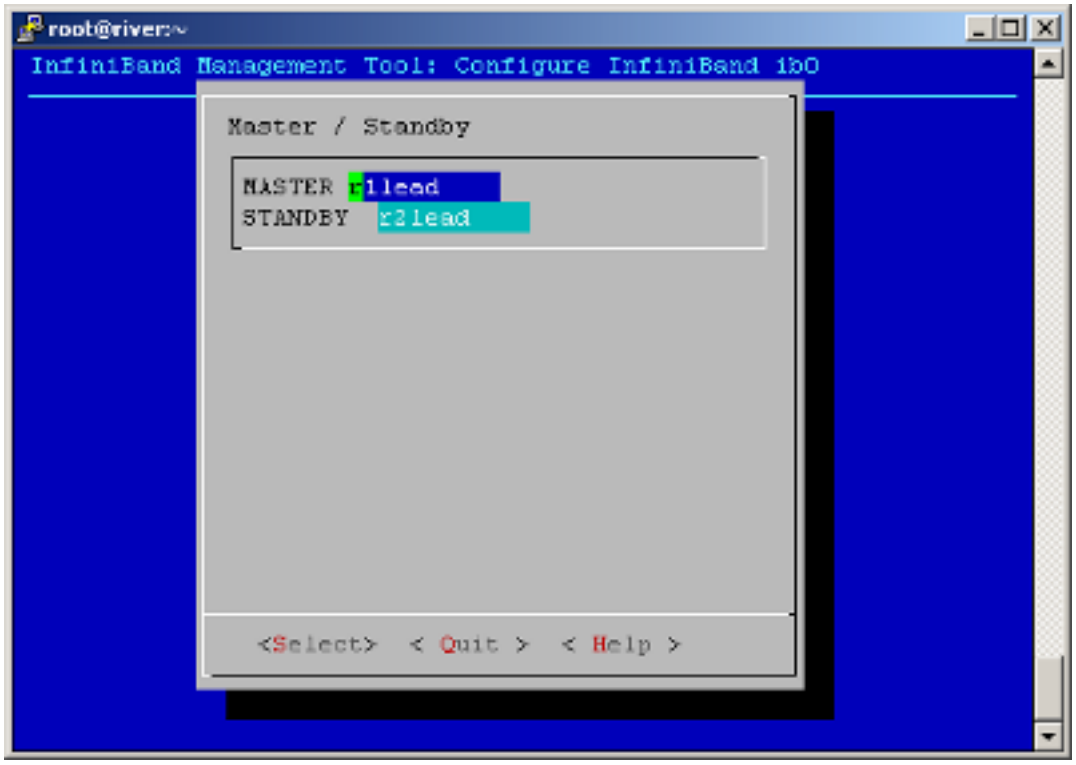

Figure 3-12 Completed InfiniBand (`ib0`) Master / Standby Screen

Example 2. On an SGI Rackable cluster, type `n1` in the **MASTER** field, and type `n101` in the **STANDBY** field.

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:

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

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

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

    Example 1. On an SGI ICE X cluster, 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.

    Example 2. On an SGI Rackable cluster, type `n101` in the **MASTER** field, and type `n1` in the **STANDBY** field.

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. Proceed to one of the following platform-specific verification procedures:
    - "Verifying That the InfiniBand Subnetwork is Working (SGI ICE X Clusters)" on page 119
    - "Verifying That the InfiniBand Subnetwork is Working (SGI Rackable Clusters)" on page 120

Verifying That the InfiniBand Subnetwork is Working (SGI ICE X Clusters)

The following procedure explains how to verify that the InfiniBand subnetwork is configured correctly on an SGI ICE X cluster.

Procedure 3-21 To verify the InfiniBand configuration on an SGI ICE X cluster

1. Type the following command to retrieve the list of rack leader controller (RLC) IDs:

   ```
   admin node:~ # cnodes --leader
   r1lead
   r2lead
   ```

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

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

For example:

```
admin node:~ # ssh r2lead
```

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

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

4. 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.

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

```
# cattr unset disableIbSwitchMonitoring
```

6. (Optional) Configure additional features.

   The SGI ICE 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 Additional Features" on page 123

Verifying That the InfiniBand Subnetwork is Working (SGI Rackable Clusters)

The following procedure explains how to verify that the InfiniBand subnetwork is configured correctly on an SGI Rackable cluster.

Procedure 3-22 To verify the InfiniBand configuration on an SGI Rackable cluster

1.

   (Optional) Configure additional features.

The SGI ICE 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 Additional Features" on page 123

Configuring Additional Features

This chapter includes the following topics:

Enabling Hardware Event Tracker (HET) Notifications

The following topics contain information about HET:

Note: This documentation includes information about all SGI systems for the sake of completeness.

About HET

All of your SGI system's IPMI controllers send SNMP traps to the management node. On SGI ICE X clusters and SGI Rackable clusters, the management node is the admin node. On an SGI UV system, the management node is the system management node (SMN). The SGI Foundation Software's HET tools process these system alerts and send an email notification after critical hardware events occur.

The HET tools are configured by default. You do not need to perform any additional system configuration, but SGI recommends that you customize the email address to which the HET tools send critical event notifications. The `het`(8) man page contains information about HET defaults and internal processes.

HET accumulates information about system events in the following log file:

`/var/log/het/het_trap_processor.log`

As an event-driven system monitoring tool, HET listens for system events. When HET receives information about an event, it converts the message from coded numbers into a readable form, as follows:

- When a noncritical event occurs, HET simply logs the event. As an option, you can configure an email address to receive noncritical event notifications.
- When a critical event occurs, HET logs the event and send an email message. SGI recommends that you edit file `/etc/sysconfig/het` and specify an email address specific to your site. By default, HET sends event information to `root@localhost`. For more information about how to customize HET notifications, see "Customizing HET Notifications" on page 124.

  The firmware for each baseboard management controller (BMC) and the firmware for each cooling node on an SGI ICE system includes threshold values for each component. If a system condition becomes too low or to high for its threshold, the BMC sends a critical event alert. The following are examples of critical system events that cause an alert:

  – Ambient air temperature outside of recommended range
  – Voltage sensor unable to attain a critical low voltage
  – Power supply failure
  – Loss of redundant power supply
  – Fan speed unable to attain a critical threshold or a loss of fan redundancy
  – Board processor modules that exceed a critical temperature threshold
  – Memory uncorrectable errors

Customizing HET Notifications

You can customize the email addresses to which event information about `NON-RECOVERABLE` events is sent. As an option, you can specify a site-specific email address for less-severe events, or all HET events, too.

The HET log file, `/var/log/het/het_trap_processor.log`, contains information about all HET events. You can consult this file periodically to monitor noncritical events.

The following procedure explains how to configure an email address or email alias to receive HET notifications.

Procedure 4-1 To customize HET notifications

1. Log in as root and open the following file with a text editor:

   `/etc/sysconfig/het`

   On an SGI cluster, log into the admin node.

   On an SGI UV system, log into the system management node (SMN). HET requires an SMN. If your system does not include an SMN, you cannot enable HET.

2. Search the file for the following string:

   `HET_MAIL_TROUBLE_TO`

3. Change the default recipient, `root`, to be the email address of a person or the email alias of a group who can attend to the system when `NON-RECOVERABLE` events occur.

4. (Optional) Configure an email recipient for notifications about `CRITICAL` events.

   Search the file for the following string:

   `HET_MAIL_NEWS_TO`

   Specify an email address or alias to receive `CRITICAL` event notifications.

5. Save and close file `/etc/sysconfig/het`.

6. (Optional) Configure an email recipient for all HET events.

   Complete the following steps:

   - Open file `/etc/het.action.d/het_mail` with a text editor.
   - Search for the following lines in `/etc/het.action.d/het_mail`:

     ```
     # NOTE: Adjust if needed
     # Default is an empty mailing list audience for
     # non (NON-RECOVERABLE or CRITICAL) events.
     ```

```
to=""
```

- Edit the `to=""` line to specify an email address or an email alias between the quotation marks.
- Save and close the file.

HET Examples

The following is an example of a HET log file that contains critical information:

```
dump     2013-10-23.07.13.21 [het_process_thread:2] # begin ---------------------------------
dump     2013-10-23.07.13.21 [het_process_thread:2] agentAddr        172.24.0.2
dump     2013-10-23.07.13.21 [het_process_thread:2] het_type         ipmi
dump     2013-10-23.07.13.21 [het_process_thread:2] guid             r1lead
dump     2013-10-23.07.13.21 [het_process_thread:2] sn               X1------
dump     2013-10-23.07.13.21 [het_process_thread:2] alertSeverity    NONE
dump     2013-10-23.07.13.21 [het_process_thread:2] event            uncorrectableECC
dump     2013-10-23.07.13.21 [het_process_thread:2] sensorName       None-memory
dump     2013-10-23.07.13.21 [het_process_thread:2] sensorNumber     0x00
dump     2013-10-23.07.13.21 [het_process_thread:2] sensorTypeName   memory
dump     2013-10-23.07.13.21 [het_process_thread:2] eventClassName   discrete
dump     2013-10-23.07.13.21 [het_process_thread:2] event1           0x51
dump     2013-10-23.07.13.21 [het_process_thread:2] event2           0xff
dump     2013-10-23.07.13.21 [het_process_thread:2] event3           0x51
dump     2013-10-23.07.13.21 [het_process_thread:2] flap_count       1
dump     2013-10-23.07.13.21 [het_process_thread:2] # end -----------------------------------
```

The corresponding email message that HET sends is as follows:

```
X-Original-To: root
Delivered-To: root@saturn9-1.americas.sgi.com
Date: Wed, 18 Dec 2013 14:36:52 -0600
From: HET.ALERT.donotreply@saturn9-1.americas.sgi.com
To: root@saturn9-1.americas.sgi.com
Subject: HET ALERT from cb9 - NON-RECOVERABLE
User-Agent: Heirloom mailx 12.2 01/07/07

The following HET(Hardware Environment Tracking) event has been recorded:
HET ALERT from cb9 - NON-RECOVERABLE
```

```
Event Details:

        EVENT                           uncorrectableECC
        HET                             r1i0n4
        LOCATION                        r1i0n4
        SENSOR                          None-memory
        SENSORNUMBER                    0x00
        SENSORTHRESHOLD                 81
        SENSORTYPE                      memory
        SENSORVALUE                     255
        SEVERITY                        NON-RECOVERABLE
        SN                              X1------
        TYPE                            ipmi
```

CPU Frequency Scaling

CPU frequency scaling allows the operating system to automatically and dynamically scale the processor frequency. CPU frequency scaling needs to be enabled in a compute node image if you want to take advantage of the Intel Turbo Boost technology that is built into each processor.

The Intel Turbo Boost Technology allows processor cores to run faster than the base operating frequency as long as they are operating below the limits set for power, current, and temperature. The CPU frequency scaling setting also affects power consumption and enables you to manage power consumption. For example, you can theoretically cut power consumption in half if you clock the computer's processors from 2 GHz down to 1 GHz.

The following procedures pertain to CPU frequency:

- "Enabling or Disabling CPU Frequency Scaling" on page 127
- "(Optional) Changing the Governor Setting and Configuring Turbo Mode" on page 129

Enabling or Disabling CPU Frequency Scaling

The procedure in this topic explains how to enable or disable CPU frequency scaling. CPU frequency scaling is disabled by default on SGI clusters.

The following procedure explains how to change your CPU frequency scaling setting.

Procedure 4-2 To control CPU frequency scaling

1. Log into the admin node as `root`.
2. Use the `cimage --list-images` command to retrieve a list of the compute node images you can edit:

   For example:

   ```
   # cimage --list-images
   image: ice-compute-sles11sp3.mpt
           kernel: 3.0.76-0.11-default
           kernel: 3.0.76-0.11-trace
   image: ice-compute-sles11sp3
           kernel: 3.0.76-0.11-default
   ```

   The previous example shows the names of two images: `ice-compute-sles11sp3.mpt` and `ice-compute-sles11sp3`.
3. Type the following command to install the system images that support CPU frequency scaling:

   # **cinstallman --yum-image --image *image_name* install sgi-base-configuration**

   For *image_name*, specify one of the compute images. For example, `ice-compute-sles11sp3.mpt` and `ice-compute-sles11sp3`.
4. Type the following command to change to the directory that contains the image you want to edit:

   # **chroot /var/lib/systemimager/images/*image_name***

   For *image_name*, specify one of the compute node image names that the `cimage` command returned. For example, using the output from the preceding step, specify either `ice-compute-sles11sp3.mpt` or `ice-compute-sles11sp3`.
5. Use a text editor to open file `/etc/modprobe.d/acpi-cpufreq.conf`.
6. Note the following line in this file:

   ```
   install acpi-cpufreq /bin/true
   ```

To enable CPU frequency scaling, insert a pound (#) character as the first character in this line, which makes the line appear as follows:

```
#install acpi-cpufreq /bin/true
```

To disable CPU frequency scaling, make sure that the `install acpi-cpufreq /bin/true` line does not contain a # character in column 1, which makes the line appear as follows:

```
install acpi-cpufreq /bin/true
```

7. Save and close file `/etc/modprobe.d/acpi-cpufreq.conf`.
8. Push the changes out to the compute nodes.

   Perform the procedure in the following topic:

   "(Conditional) Pushing Changes to SGI ICE Compute Nodes" on page 28
9. (Optional) Change the CPU frequency governor setting and configure turbo mode.

   The default governor setting and the default turbo mode setting are appropriate for most SGI ICE systems. If you want to change these settings, proceed to the following:

   "(Optional) Changing the Governor Setting and Configuring Turbo Mode" on page 129

(Optional) Changing the Governor Setting and Configuring Turbo Mode

Use the procedure in this topic to change the governor setting and, optionally, to configure turbo mode. When you enable turbo mode, you enable the CPU frequency to exceed its nominal level for short periods of time, depending on the processor, temperature, current, power, and other factors. For general information about turbo mode, see the following website:

https://www-ssl.intel.com/content/www/us/en/architecture-and-technology/turbo-boost/turbo-boost-technology.html

The following procedure explains how to set the CPU frequency governor appropriately and how to configure turbo mode.

Procedure 4-3 To change the governor setting and configure turbo mode

1. Make sure that CPU frequency is enabled.

For information, see "Enabling or Disabling CPU Frequency Scaling" on page 127.

2. Examine the following list and choose a power governor setting:

***governor* Setting**	**Effect**
`ondemand`	Dynamically switches between the available CPUs if at 95% of CPU load. Default.
`performance`	Runs the CPUs at the maximum frequency.
`conservative`	Dynamically switches between the available CPUs if at 75% of CPU load.
`powersave`	Runs the CPUs at the minimum frequency.
`userspace`	Runs the CPUs at user-specified frequencies.

3. Use the `cimage --list-images` command to retrieve a list of the compute node images you can edit:

   For example:

```
# cimage --list-images
image: ice-compute-sles11sp3.mpt
        kernel: 3.0.76-0.11-default
        kernel: 3.0.76-0.11-trace
image: ice-compute-sles11sp3
        kernel: 3.0.76-0.11-default
```

   The previous example shows the names of two images: `ice-compute-sles11sp3.mpt` and `ice-compute-sles11sp3`.

4. Type the following command to change to the directory that contains the image you want to edit:

   # **chroot /var/lib/systemimager/images/*image_name***

   For *image_name*, specify one of the compute node image names that the `cimage` command returned. For example, using the output from the preceding step, specify either `ice-compute-sles11sp3.mpt` or `ice-compute-sles11sp3`.

5. Use one of the following platform-specific methods to change the setting:

   - On RHEL platforms, complete the following steps:

     1. Open file `/etc/sysconfig/cpuspeed`.

2. Search for the `GOVERNOR=` string.
3. Edit the setting, adding the *governor* setting you chose in the previous step.
4. Save and close the file.
5. Type the following command:

   # **service cpuspeed restart**

- On SLES platforms, complete the following steps:

  1. Type the following command:

     # **cpupower frequency-set -g *governor***

     For *governor*, specify the setting you chose in the previous step.

  2. Type the following command:

     # **cpupower frequency-info**

  3. Verify that the *governor* setting you specified appears in the command in the output in the `current policy` field.
  4. Use a text editor to edit the `/etc/init.d/after.local` file and add the following line:

     `cpupower freqency-set -g` *governor*

     The preceding line ensures that after each boot, the system sets the *governor* setting you specified.

6. Push the changes out to the compute nodes.

   Perform the procedure in the following topic:

   "(Conditional) Pushing Changes to SGI ICE Compute Nodes" on page 28

7. Use the `cat`(1) command to retrieve the list of available frequencies. For example:

```
# cat /sys/devices/system/cpu/cpu0/cpufreq/scaling_available_frequencies
3301000 3300000 3200000 3100000 3000000 2900000 2800000 2700000 2600000
2500000 2400000 2300000 2200000 2100000 2000000 1900000 1800000 1700000
1600000 1500000 1400000 1300000 1200000
```

The preceding output shows the available frequencies, listed in order from the highest, 3301000 KHz, to the lowest, 1200000 KHz.

On SGI systems, the second frequency listed is always the processor's nominal frequency. This is a 3.3 GHz processor, so 3300000 KHz is the nominal frequency.

You can also obtain the nominal frequency by typing the following command and examining the information in the model name field:

```
# cat /proc/cpuinfo
```

8. Use the `cpupower` command to set the frequency to the nominal frequency of 3.3 GHz plus 1 MHz.

   That is, specify a frequency of 3301 MHz. For example:

   ```
   # cpupower frequency-set -u 3301MHz
   ```

   Later, if you want to disable turbo mode, type the following command to set the maximum frequency back to the nominal frequency:

   ```
   # cpupower frequency-set -u 3300MHz
   ```

Configuring Array Services for MPI Programs

You can configure compute nodes into an array. After you configure a set of compute nodes into an array, the Array Services software can perform authentication and coordination functions when MPI programs are running. For more information about MPI, see the *Message Passing Toolkit (MPT) User Guide*.

You cannot include the admin node or any rack leader controller (RLC) nodes in an array.

For general Array Services configuration information, see the following:

- `arrayconfig`(8), which describes how to use the `arrayconfig`(8) command to configure Array Services.
- `arrayconfig_tempo`(8), which describes Array Services configuration characteristics that are specific to clusters.

The procedures in the following topics assume that you want to create new a master image for the compute nodes and a new master image for the flat compute nodes that you configured with user services. After you create these images, you can push out

the new images to the compute nodes and to the compute services nodes. The alternative is to configure Array Services directly on the nodes themselves, but this method leaves you with an Array Services configuration that is overwritten the next time someone pushes new software images to the cluster's nodes.

Note: The procedures in the following topics assume that you want to install Array Services on an SGI ICE X cluster. The steps you need to complete for an SGI Rackable cluster are very similar with the major difference being that you do not need to perform steps that pertain to racks. On an SGI Rackable cluster, the steps in the following procedures that pertain to racks and rack leader controllers (RLCs) do not need to be performed.

The following procedures explain how to configure Array Services.

- "Planning the Configuration" on page 133
- "Preparing the Images" on page 135
- "Power Cycling the Nodes and Pushing Out the New Images" on page 142

Planning the Configuration

The following procedure explains preliminary steps that you can take to plan your array and the security you want to enable.

Procedure 4-4 To plan the array

1. Verify that the SGI Performance Suite software is installed on the cluster.
2. Log into the admin node as the root user.
3. Use the `cnodes` command to display a list of available nodes, and decide which nodes you want to include in the array.

   For example:

   ```
   # cnodes --compute
   ```

   The command shows all the compute nodes, including those that might be configured as compute services nodes at this time.
4. Use the `cinstallman` command to display a list of the available system images, and decide which images you want to edit.

For example, the following output is for an example SGI ICE X cluster that is running in production mode:

```
# cinstallman --show-images
Image Name                  BT VCS Compat_Distro
lead-sles11sp3              0  1   sles11sp3        # Default RLC node image
lead-sles11sp3.prod1        0  1   sles11sp3        # Site's production RLC image
sles11sp3                   0  1   sles11sp3        # Default flat compute node image
sles11sp3.prod1             0  1   sles11sp3        # Site's production compute services node image
ice-sles11sp3               0  1   sles11sp3        # Default SGI ICE compute node image
ice-sles11sp3.prod1         0  1   sles11sp3        # Site's production SGI ICE compute node image
```

The output includes image `sles11sp3.prod1`, which is the image installed on a flat compute node that is configured as a compute services node. Image `sles11sp3.prod1` is based on image `sles11sp3`, but can includes software to support user logins and a backup DNS server.

All system images are stored in `/var/lib/systemimager/images`.

The preceding output shows the orginal, factory-shipped system images for the RLCs, the flat compute nodes, and the SGI ICE compute nodes. These original files are as follows:

- `lead-sles11sp3`
- `sles11sp3`
- `ice-sles11sp3`

The output also shows customized images for this SGI ICE X cluster. These file names end in `.prod1`, for production use, and are as follows:

- `lead-sles11sp3.prod1`
- `sles11sp3.prod1`. This is the image that resides on the flat compute services node.
- `ice-sles11sp3.prod1`

The examples in this Array Services configuration procedure add the Array Services information to the customized, production images with the `.prod1` suffix.

5. Decide what kind of security you want to enable.

Array Services includes its own authentication and security, but if your site requires additional security, you can configure MUNGE security, which the installation includes. Your security choices are as follows:

- `munge` on all the nodes you want to include in the array. Configures additional security provided by MUNGE. The installation process installs MUNGE by default. If you decide to use MUNGE, the SGI MPT configuration process explains how to enable MUNGE at the appropriate time.
- `none` on the compute services nodes and `none` on the compute nodes **OR** `noremote` on the compute services nodes and `none` on the the compute nodes. These specifications have the following effects:
  - When you specify `none` on all the nodes you want to include in the array, all authentication is disabled.
  - When you specify `noremote` on the compute services nodes and specify `none` on the compute nodes, users must run their jobs directly from the compute services nodes. In this case, users cannot submit SGI MPI jobs remotely.
- `simple` (default). Generates hostname/key pairs by using either the OpenSSL, `rand` command, 64–bit values (if available) or by using `$RANDOM` Bash facilities.

6. Proceed to the following:

"Preparing the Images" on page 135

Preparing the Images

Before you can create images that include Array Services, you need to copy, or *clone*, the production system images your system is using at this time.

The following procedure explains how to prepare the images.

Procedure 4-5 To prepare the system images

1. Log into the admin node as the root user, and use two `cinstallman` commands, in the following format, to clone (1) one of the images that resides on a compute services node and (2) one of the images that reside on a compute node:

   `cinstallman --clone-image` *existing_image new_image*

   For *existing_image*, specify the name of one of the existing images.

For *new_image*, specify the new name for that to want to give to the image.

For example, the following commands copy the first-generation production images to new, second-generation production images:

```
# cinstallman --create-image --clone --source sles11sp3.prod1 --image sles11sp3.prod2
# cinstallman --create-image --clone --source ice-sles11sp3.prod1 --image ice-sles11sp3.prod2
```

2. Type the following command to change to the system images directory:

   ```
   # cd /var/lib/systemimager/images
   ```

3. (Optional) Use the cp(1) command to copy the MUNGE key from the new compute services node image to the new compute node image.

   Complete this step if you want to configure the additional security that MUNGE provides.

   The MUNGE key resides in `/etc/munge/munge.key` and must be identical on all the nodes that you want to include in the array. The copy command is as follows:

   `cp /var/lib/systemimager/images/`*new_computeservices_image*`/etc/munge/munge.key \`
   `/var/lib/systemimager/images/`*new_compute_image*`/etc/munge/munge.key`

   For *new_computeservices_image*, specify the name of the new compute services node image you created.

   For *new_compute_image*, specify the name of the new compute node image you created.

   For example:

```
# cp /var/lib/systemimager/images/sles11sp3.prod2/etc/munge/munge.key \
/var/lib/systemimager/images/ice-sles11sp3.prod2/etc/munge/munge.key
```

Note: The commands and formats in this step use the backslash (\) continuation character.

4. Use the `cinstallman` command to install the new image on the compute services node.

   `cinstallman --assign-image --node` *hostname(s)* `--image` *new_computeservices_image*

For *hostname*, specify the hostname or hostnames of the compute services node that you want users to log into when they log into the array.

For example, the following command installs the new image on node `n1`:

```
# cinstallman --assign-image --node n1 --image sles11sp3.prod2
```

5. Use the `ssh`(1) command to log into the compute services node from which you expect users to run SGI MPI programs.

   For example, log into `n1`.

6. Use the `arrayconfig`(8) command to configure the compute service node(s) and compute nodes into an array.

   The `arrayconfig`(8) command creates the following files on the compute service node to which you are logged in:

   - `/etc/array/arrayd.conf`
   - `/etc/array/arrayd.auth`

   Type the `arrayconfig`(8) command in the following format:

   `/usr/sbin/arrayconfig -a` *arrayname* `-f -m -A` *method node node* `...`

   For *arrayname*, specify a name for the array. The default is `default`.

   For *method*, specify `munge`, `none`, or `simple`. A later step explains how to specify `noremote` for a compute services node.

   For each *node*, specify a list of node IDs.

   Example 1. To specify that array `myarray` use MUNGE security and include all compute service nodes and all compute nodes, type the following command:

```
# /usr/bin/arrayconfig -a myarray -f -m -A munge $(cnodes --compute --ice-compute)
```

   Example 2. To specify that array `yourarray` use no security, include one compute service node, and include all compute nodes, type the following command:

```
# /usr/bin/arrayconfig -a yourarray -f -m -A none n0 $(cnodes --ice-compute)
```

7. Proceed to the following:

   "Configuring the Authentication Files in the New System Images on the Admin Node" on page 138

Configuring the Authentication Files in the New System Images on the Admin Node

Complete one of the following procedures, based upon whether you want to permit remote access to the compute services node:

- If you specified `-A munge` or `-A simple` for authentication

  OR

  If you specified `-A none` for authentication, and you want to permit users to log into a compute services node remotely to submit MPI programs, proceed to the following:

  "Permitting Remote Access to the Compute Services Node" on page 138

- If you specified `-A none` for authentication, and you want to prevent users from logging into a compute services node remotely to submit MPI programs, proceed to the following:

  "Preventing Remote Access to the Service Node" on page 139

Permitting Remote Access to the Compute Services Node

The following procedure assumes that you want to permit job queries and commands on the compute services node. It explains how to copy the array daemon files to the admin node.

Procedure 4-6 To permit remote access to the compute services node

1. Log into one of the compute services nodes as the root user.
2. Copy the `arrayd.auth` file and the `arrayd.conf` files from the compute services node to the new compute services node image on the admin node.

   Type the following command:

   ```
   # scp /etc/array/arrayd.* \
   admin:/var/lib/systemimager/images/computeservices_image/etc/arrayd.*
   ```

   For *computeservices_image*, specify the compute services node image on the admin node.

   Type this command all on one line. Note that the command in this step uses a backslash (\) character to continue the command to the following line.

For example:

```
# scp /etc/array/arrayd.* \
admin:/var/lib/systemimager/images/sles11sp3.prod2/etc/arrayd.*
```

3. Copy the `arrayd.auth` file and the `arrayd.conf` files from the compute services node to the new compute node image on the admin node.

   Type the following command:

   ```
   # scp /etc/array/arrayd.* \
   admin:/var/lib/systemimager/images/compute_image/etc/arrayd.*
   ```

   For *compute_image*, specify the SGI ICE compute node image on the admin node.

   Type this command all on one line. Note that the command in this step uses a backslash (\) character to continue the command to the following line.

   For example:

   ```
   # scp /etc/array/arrayd.* \
   admin:/var/lib/systemimager/images/ice-sles11sp3.prod2/etc/arrayd.*
   ```

4. Proceed to the following:

   "Distributing Images to all the Nodes in the Array" on page 141

Preventing Remote Access to the Service Node

You can prevent a compute services node from receiving any requests from other computers on the network. In this case, the compute services node can send requests to all remote nodes, but it does not listen on TCP port 5434 for any incoming requests. Complete the procedure in this topic if this is the behavior your site requires.

The following procedure explains how to configure the Array Services files to prevent remote access and how to copy the array daemon files to the admin node.

Procedure 4-7 To prevent remote access to the compute services node

1. Log into one of the compute services nodes as the root user.
2. Open the following file with a text editor:

   `/etc/array/arrayd.auth`

3. Type the following, all on one line:

   ```
   AUTHENTICATION NOREMOTE
   ```

4. Save and close the file.

   The file should contain only the one line.

5. Type the following command to copy `/etc/array/arrayd.auth` and `/etc/array/arrayd.conf` from the compute services node to the new compute services node image on the admin node:

   **# scp /etc/array/arrayd.* **
   admin:/var/lib/systemimager/images/*computeservices_image*/etc/arrayd.*

   For example:

   **# scp /etc/array/arrayd.* **
   admin:/var/lib/systemimager/images/sles11sp3.prod2/etc/arrayd.*

6. Log into the admin node as the root user.

7. Create file `/var/lib/systemimager/images/`*compute_image*`/etc/array/arrayd.auth`.

   For example:

 # vi /var/lib/systemimager/images/sles11sp3.prod2/etc/array/arrayd.auth

8. Add the following all on one line:

   ```
   AUTHENTICATION NONE
   ```

9. Save and close the file.

   The file should contain only the one line.

10. Type the following command to copy the `/etc/array/arrayd.conf` file to the SGI ICE compute nodes:

    **# scp /etc/array/arrayd.conf **
    admin:/var/lib/systemimager/images/*compute_image*/etc/arrayd.conf

    For example:

    **# scp /etc/array/arrayd.conf **
    admin:/var/lib/systemimager/images/ice-sles11sp3.prod2/etc/arrayd.conf

11. Proceed to the following:

    "Distributing Images to all the Nodes in the Array" on page 141

Distributing Images to all the Nodes in the Array

The following procedure explains how to assign the new compute services node image and the new compute node image to the compute services nodes and to the compute nodes.

Procedure 4-8 To assign images

1. Log into the admin node as the root user.
2. Use one or more `cinstallman` commands in the following format to assign the new compute services node image to the compute service nodes:

   `cinstallman --assign-image --node` *hostname* `--image` *new_computeservice_image*

   For *hostname*, specify the hostname for one or more of the compute services nodes. In the cluster definition file, this is the name that appears in the `hostname1` field.

   If you want to specify more than one hostname, and your nodes are similarly named, use the * wildcard character to represent a string of identical characters in this field. For example, if your hostnames are `n1`, `n2`, `n3`, and `n57`, specify `n*` in this field if you want to specify all compute services nodes.

   For *new_service_image*, specify the name of the new compute services node image you created.

   Example 1. The following command assigns the new compute services node image to all compute services nodes:

```
# cinstallman --assign-image --node n* --image sles11sp3.prod2
```

   Example 2. The following command assigns the new compute services node image to `n101`:

```
# cinstallman --assign-image --node n101 --image sles11sp3.prod2
```

3. Use one or more `cinstallman` commands in the following format to assign the new image to the compute services nodes the next time you boot the compute services nodes:

   `cinstallman --next boot --image` *hostname*

For *hostname*, specify the hostname for one or more of the compute services nodes. In the cluster definition file, this is the name that appears in the `hostname1` field. You can use the `*` wildcard character in this field if specifying more than one hostname.

For example:

```
# cinstallman --next boot --image n101
```

4. Proceed to the following:

"Power Cycling the Nodes and Pushing Out the New Images" on page 142

Power Cycling the Nodes and Pushing Out the New Images

The following procedure explains how to install the new images on the compute services nodes and the compute nodes that you want to include in the array.

Procedure 4-9 To propagate the images

1. Type the following command to reboot the compute services nodes and the compute nodes:

```
# cpower --system --noleader --reboot
```

2. Use one or more `cpower` commands in the following format to power off the compute nodes that you want to reimage:

`cpower --off` *hostname*

For *hostname*, specify the hostname(s) of the compute nodes.

You can use wildcard characters if you have many compute nodes.

For example, the following command powers off all the compute nodes on an SGI ICE X cluster:

```
# cpower --off r*i*n*
```

Issue as many `cpower` commands as needed.

3. (Conditional) Use the `cimage` command to push the new compute node image out to the compute nodes.

Complete this step only on SGI ICE X clusters. You do not need to complete this step on SGI Rackable clusters.

Use the following format:

`cimage --push-rack` *new_compute_image rack*

For *new_compute_image*, specify the name of the new compute node image you created.

For *rack*, specify the IDs of the racks in which the compute nodes reside.

For example, the following command pushes the compute node images to all compute nodes in all racks:

```
# cimage --push-rack ice-sles11sp3.prod2 r*
```

4. Use one or more `cpower` command in the following format to start the compute nodes:

   `cpower --on` *hostname*

   For *hostname*, specify the hostnames of the compute nodes.

   You can use wildcard characters if you have many compute nodes.

   For example, the following command powers on all compute nodes:

   ```
   # cpower --on r*i*n*
   ```

   Issue as many `cpower` commands as needed.

Enabling the Mellanox OpenFabrics Enterprise Distribution for Linux (MLNX_OFED) Software

The following procedure explains how to enable Mellanox's OFED software package and remove the OFED software included by default in the operating system software on your cluster.

Procedure 4-10 To enable Mellanox OFED

1. Log in as root to the admin node.
2. Download the Mellanox software package to the `/tmp` directory on the admin node from its location on SGI Supportfolio.

The packages are in `tar` format and are as follows:

```
mlnx-ofed-v2.3-1.0.1-rhel65-rpms.tgz
mlnx-ofed-v2.3-1.0.1-sles11sp3-rpms.tgz
```

3. Use the `mkdir`(1) command in the following format to create a repository on the admin node:

   `mkdir -p /tftpboot/mlnx/mlnx-ofed-2.3-1.0.1-`*distro*

   For *distro*, specify `rhel6.5` or `sles11sp3`.

4. Use the `cd`(1) command to change to the repository you created.

5. Use the `tar`(1) command in the following format to extract the `tar`(1) package into the repository directory:

`tar vxzf mlnx-ofed-v2.3-1.01-`*distro*`-rpms.tgz -C /tftpboot/mlnx/mlnx-ofed-v2.3-1.01-`*distro*

   For *distro*, specify `rhel6.5` or `sles11sp3`.

6. Use the `crepo` command in the following format to create the repository:

   `crepo --create /tftpboot/mlnx --custom MLNX_OFED-2.3-1.0.1-`*distro*

   For *distro*, specify `rhel6.5` or `sles11sp3`.

7. Clone the following images as appropriate for your cluster:

   - The rack leader controller (RLC) image.

     SLES example:

```
# cinstallman --create-image --clone --source lead-sles11sp3 --image lead-sles11sp3.mlnx-ofed23
```

     RHEL example:

```
# cinstallman --create-image --clone --source lead-rhel6.6 --image lead-rhel6.6.mlnx-ofed23
```

   - The SGI ICE compute image.

     SLES example:

```
# cinstallman --create-image --clone --source ice-sles11sp3 --image ice-sles11sp3.mlnx-ofed23
```

     RHEL example:

```
# cinstallman --create-image --clone --source ice-rhel6.6 --image ice-rhel6.6.mlnx-ofed23
```

- The flat compute image.

  SLES example:

```
# cinstallman --create-image --clone --source sles11sp3 --image sles11sp3.mlnx-ofed23
```

  RHEL example:

```
# cinstallman --create-image --clone --source rhel6.6 --image rhel6.6.mlnx-ofed23
```

8. Unselect the distro repository.

   RHEL example:

   # **crepo --unselect Red-Hat-Enterprise-Linux-6.5**

   SLES example:

   # **crepo --unselect SUSE-Linux-Enterprise-Server-11-SP3**

9. Select the newly created MLNX_OFED repository.

   RHEL example:

   # **crepo --select MLNX_OFED-2.3-1.0.1-rhel6.5**

   SLES example:

   # **crepo --select MLNX_OFED-2.3-1.0.1-sles11sp3**

10. Remove the `distro-ofed` package from each ot the newly created images.

    That is, remove the following software packages from each of the RLC, SGI ICE compute, and flat compute node images:

    - From the RLC image, remove `sgi-lead-distro-ofed`.
    - From the SGI ICE compute node image, remove `sgi-compute-distro-ofed`.
    - From the flat compute node image, remove `sgi-compute-distro-ofed`.

    For example, the following commands remove the images from SLES packages:

```
# cinstallman --yum-image --image service-sles11sp3.mofed23 remove sgi-service-distro-ofed
# cinstallman --yum-image --image compute-sles11sp3.mofed23 remove sgi-service-distro-ofed
# cinstallman --yum-image --image lead-sles11sp3.mofed23 remove sgi-service-distro-ofed
```

11. Install the MLNX_OFED packages into the newly created images.

    For example, the following commands install the packages into SLES images:

```
# cinstallman --yum-image --image service-sles11sp3.mofed23 install libibverbs.x86_64 \
libibverbs-utils.x86_64 libibumad.x86_64 libibmad.x86_64 librdmacm.x86_64 \
librdmacm-utils.x86_64 libmlx4.x86_64 libmlx5.x86_64 opensm-libs.x86_64 opensm.x86_64 \
infiniband-diags.x86_64 libibcm.x86_64 ofed-scripts.x86_64 ibutils.x86_64 ibutils2.x86_64 \
dapl.x86_64 dapl-utils.x86_64 mstflint.x86_64 mft.x86_64 ibdump.x86_64 perftest.x86_64
# cinstallman --yum-image --image service-sles11sp3.mofed23 install mlnx-ofa_kernel.x86_64 \
mlnx-ofa_kernel-kmp-default.x86_64 kernel-mft-mlnx-kmp-default.x86_64 srp-kmp-default.x86_64
```

    For example, the following commands install the packages into RHEL images:

```
# cinstallman --yum-image --image service-rhel6.5.mofed23 install libibverbs.x86_64 \
libibverbs-utils.x86_64 libibumad.x86_64 libibmad.x86_64 librdmacm.x86_64 \
librdmacm-utils.x86_64 libmlx4.x86_64 libmlx5.x86_64 opensm-libs.x86_64 opensm.x86_64 \
infiniband-diags.x86_64 libibcm.x86_64 ofed-scripts.x86_64 ibutils.x86_64 ibutils2.x86_64 \
dapl.x86_64 dapl-utils.x86_64 mstflint.x86_64 mft.x86_64 ibdump.x86_64 perftest.x86_64
# cinstallman --yum-image --image service-rhel6.5.mofed23 install mlnx-ofa_kernel.x86_64 \
kmod-mlnx-ofa_kernel.x86_64 kmod-kernel-mft-mlnx.x86_64 kmod-srp.x86_64
```

Troubleshooting Configuration Changes

If a configuration change does not affect the cluster in the intended manner, try one of the following approaches:

- Edit the node image on the admin node. For example, you can reconfigure the image for the compute nodes that you use for user services on the admin node and reimage the compute services nodes with that new image.
- Edit the node customization scripts. For example, the compute node update scripts reside on the admin node in the /opt/sgi/share/per-host-customization/global directory.

Troubleshooting

This chapter covers the following topics:

About Troubleshooting

This chapter provides answers to some common problems users encounter when installing or upgrading a cluster. It includes diagnosis and troubleshooting information.

Using the `switchconfig` Command

The `switchconfig` command displays switch settings and enables you to configure switches.

To retrieve help output online, type the following:

```
# switchconfig set --help
```

Unless you want to retrieve help output, the following parameters are required:

- `--default-vlan | -d` *default_vlan*
- `--switches | -s` *hostname_or_IP*`[,`*hostname_or_IP*`][, ...]`
- `--macs | m` *mac_addr*`[,`*mac_addr*`][, ...]` or `--ports | -p` *port_num*`[,`*port_num*`][, ...]`

The command's format is as follows:

```
switchconfig set subcommand
[--bonding | -b lacp|manual|none]
[--debug]
[--default-vlan | -d vlan_number]
[--help | -h]
[--log | -l file]
[--macs | m mac_addr[,mac_addr][, ...]]
[--ports | -p port_num[,port_num][, ...]]
[--redundant | -r NO|YES]
[--switches | -s hostname_or_IP[,hostname_or_IP][, ...]]
[--vlan | -v vlan_number]
```

For *subcommand*, specify one of the following:

Table 5-1 `switchconfig` Subcommands

Group	Subcommand	Purpose
Informational	`list`	Displays current settings.
Configuration	`set`	Assigns virtual local area networks (VLANs) and other settings to one or more MAC addresses.
	`unset`	Returns ports to default settings.
IP management	`list_ip`	Returns the IP address that is configured for the switch.
	`set_ip`	Adds an IP address for the VLAN on the switch. Used to route traffic from MIC devices.
	`unset_ip`	Removes the IP address for the VLAN on the switch.
OSPF management	`list_ospf`	Returns the open shortest path first (OSPF) router protocol used to route the MIC traffic between the switches.
	`set_ospf`	Sets the OSPF router protocol used to route traffic between the switches.
	`unset_ospf`	Disables OSPF and all the network statements that are associated with OSPF.
MTU management	`list_mtu`	Displays the assigned maximum transportation unit (MTU) value for all ports in the switch stack.
	`set_mtu`	Assigns all ports in the switch stacks to the MTU value.
Gateway management	`list_default_gateway`	Displays the default gateway assigned to the switch network.
	`set_default_gateway`	Sets the default gateway for the switch network.

Group	Subcommand	Purpose
	`unset_default_gateway`	Clears the default gateway that is assigned to the switch network.
Propagate configuration	`pull_switch_config`	Copies the start-up switch's configuration file from the TFTP server to the switch(es) you specify on the command and loads the switch.
	`push_switch_config`	Copies the switch configuration file from the switch you specify and saves it to the TFTP server.
Specify the SNMP community	`set_snmp_community`	Defines the standard network management protocol (SNMP) community settings. Defines the components or devices in a specific SNMP community. A component can reside within more than one community.
	`list_snmp_community`	Displays the SNMP community settings.
Miscellaneous	`change_password`	Changes the administrator password on the switch. This subcommand accepts arguments of `-c` *old_password* and `-n` *new_password*. For *old_password*, specify the old password that you want to change. By default, this password is the same as the admin node password, which is `admin`. For *new_password*, specify the new password that you want to use on the switches.

Group	Subcommand	Purpose
	`find`	Maps one or more MAC addresses to the mananagement switch to which the MAC address is physically connected. Uses forwarding database (FDB) and link layer discovery protocol (LLDP).
	`list_route`	Displays the routing table.
	`reset_factory_defaults`	Reverts to the default, factory configuration and reboots.
	`sanity_check`	Runs a sanity check on the switch configuration. For example, it checks the switch ports for trunking. It returns configuration information regarding misconfigured elements and points out anomalies.
	`save_running_config`	Saves the current switch configuration to the switch's nonvolatile memory.

The `--redundant | -r` parameter accepts `NO` or `YES` as arguments. The default is `NO`. If you specify `YES`, the command adds the same port associated with a 2–stack of switches to a bonding group.

The variables in the `switchconfig` command arguments are as follows:

Argument	**Specification**
vlan_number	The VLAN number. Must be a positive integer.
file	The full path to the file to which `switchconfig` can write log output.
mac_address	The MAC address of the switch.
port_num	The port number.

hostname_or_IP The hostname or IP address of the switch.

Example 1. The following `switchconfig` command returns help text for the entire `switchconfig` command:

```
# switchconfig --help
```

Example 2. The following `switchconfig` command returns help text for only the `set_ip` parameter:

```
# switchconfig set_ip --help
```

The preceding command shows how to display help output for `set_ip`, which is only one of the `switchconfig` subcommands. You can display output for any of the other `switchconfig` subcommands in the same way.

Example 3. The following `switchconfig` command shows how to set the IP address to `100.100.100.100` on for VLAN `101` on switch `mgmtsw0`:

```
# switchconfig set_ip --switch mgmtsw0 --vlan 101 --ip 100.100.100.100
```

Example 4. The following `switchconfig` command adds an IP address for the VLAN on the switches that route traffic for the MIC devices on the compute nodes:

```
# switchconfig set_ip --switch mgmtsw0 --vlan 101 --ip 10.159.2.54 --netmask 255.255.255.0
```

The following command is equivalent:

```
# switchconfig set_ip --s mgmtsw0 --v 101 --i 10.159.2.54 --n 255.255.255.0
```

Example 5. The following two `switchconfig` commands return the switch IP addresses used for routing.

Command one:

```
# switchconfig show_ip -s mgmtsw0 -v 1
VLAN 1 is Administrative Up - Link Up
  Address is B4-0E-DC-39-C4-83
  Index: 1001, MTU: 1500
  Address Mode is DHCP
  IP Address: 172.23.0.254 Mask: 255.255.0.0
  Proxy ARP is disabled
```

Command two:

```
# switchconfig show_ip -s mgmtsw0 -v 101
VLAN 101 is Administrative Up - Link Up
  Address is B4-0E-DC-39-C4-83
  Index: 1101, MTU: 1500
  Address Mode is User specified
  IP Address: 10.159.1.254 Mask: 255.255.255.0
  IP Address: 10.157.1.254 Mask: 255.255.255.0 Secondary
  IP Address: 10.158.1.254 Mask: 255.255.255.0 Secondary
  IP Address: 10.160.1.254 Mask: 255.255.255.0 Secondary
  Proxy ARP is disabled
```

Example 6. All of the switches for an cluster must have the same password. The following command changes the switch password for a system with two switches:

```
# switchconfig change_password -c admin -n mynewpassword --switches mgmtsw0,mgmtsw1
```

SGI ICE Compute Nodes Are Taking Too Long To Boot (SGI ICE X Clusters Only)

If the SGI ICE compute nodes on an SGI ICE X cluster are taking a long time to boot, perform the following:

- See "Verify the Bonding Mode on the Rack Leader Controller (RLC) (SGI ICE X Clusters Only)" on page 154 to verify that the compute nodes have the proper bonding setup.
- Verify the rack leader controller (RLC) has a MegaRAID controller. 144 nodes do not boot well with 106x controllers, for example. You can verify this with `lspci` command.

  To verify the MegaRAID 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) (SGI ICE X Clusters Only)

The redundant management network (RMN) is configured by default. To verify the bonding mode, log into an RLC and type the following command:

```
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 cluster configuration tool's **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 100 Mb/s 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 admin node, 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 admin 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`.

`cimage --push-rack` Pushes Too Many (or Too Few) Expansions (SGI ICE X Clusters Only)

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) (SGI ICE X Clusters Only)

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 admin node. 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 the cluster configuration tool, 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 "Restarting the `blademond` Daemon (SGI ICE X Clusters Only)" on page 160.

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 "Troubleshooting a Rack Leader Controller (RLC) With Misconfigured Switch Information (SGI ICE X Clusters Only)" on page 165).

Confirm the wiring rules. See "Switch Wiring Rules" on page 167.

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 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 admin node `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.

Restarting the `blademond` Daemon (SGI ICE X Clusters Only)

From the rack leader controller (RLC), perform the following steps:

1. Stop the daemon:

   ```
   r1lead:~ # service blademond stop
   ```

2. Remove `etc/dhcpd.conf.d/ice.conf` or `/etc/dhcp/dhcpd.conf.d/ice.conf`:

   `rm ice.conf dhcpd.conf`

3. Remove `slot_map`:

   `r1lead:~ #` **`rm /var/opt/sgi/lib/blademond/slot_map`**

4. Start the daemon:

   `r1lead:~ #` **`service blademond start`**

Log Files

All of the log files reside in the `/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 admin node, 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 admin node, `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 admin node, 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) (SGI ICE X Clusters Only)" on page 158.
- 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:
  - Dump the slot map from each CMC to your screen:

    ```
    r1lead:~ # /opt/sgi/lib/dump-cmc-slot-tables
    ```

  - Query an individual slot map:

    ```
    r1lead:~ # echo STATUS | netcat r1i0c 4502
    ```

    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 admin node.

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.
```

Resolving CMC Slot Map Ordering Issues

If there are ssh(1) key failures or if the compute node hosts seem to be BMCs, it is possible that there are problems with the CMC slot map might be corrupted.

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 "Restarting the blademond Daemon (SGI ICE X Clusters Only)" on page 160). 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_cache
```

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 "Restarting the `blademond` Daemon (SGI ICE X Clusters Only)" on page 160.

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 164).

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 admin node and rack leader controller (RLC) have the correct time
- Make sure the admin node and RLCs are synchronized with `ntp`. An admin node 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
  ```

- Make 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
```

Troubleshooting a Rack Leader Controller (RLC) With Misconfigured Switch Information (SGI ICE X Clusters Only)

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 167.
2. Make sure all management switches are reachable from the admin node.
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 clusters 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 admin nodes, rack leader controllers (RLCs), compute services 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 connect `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 primary 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 to prevent looping.
  - The `discover` command will tell you when it is safe to plug in the second leg of the trunk. This avoids circuit loops.

Admin Node `eth2` Link in the Bond is Down

A problem occasionally occurs, especially in SGI XE270 admin nodes, 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 admin node comes back up with `eth2` and `eth3` properly ordered.

Installing SGI Tempo Versions Older than 2.9.0

After you upgrade or intstall SGI Tempo 2.9.0 on any slot, the boot manager is changed to GRUB version 2. At this point, you can no longer install SGI Tempo versions earlier than 2.9.0 on any slot. The procedure in this topic explains how to install an earlier SGI Tempo version. For information about booting an SGI Tempo 2.9.0 system, see the following:

"Booting the System" on page 43

The following procedure explains how to install a version of SGI Tempo that is earlier than 2.9.0. After you run this procedure, the boot system is again GRUB version 1. Subsequently, if you install SGI Tempo 2.9.0, the boot system changes again to GRUB version 2.

Procedure 5-1 To install an SGI Tempo version that is earlier than SGI Tempo 2.9.0

1. Log into the slot upon which you installed SGI Tempo 2.9.0 as the root user.
2. Type the following command to revert to the GRUB version 1 boot manager on the admin node:

   # **/opt/sgi/lib/revert-admin-to-legacy-grub**

3. Use the SGI DVD to install the older version of SGI Tempo that you want to use.

Booting Nodes With iPXE After an Upgrade

If a node fails to boot after an upgrade, you might need to specify that iPXE load first and that iPXE load GRUB version 2. From the admin node, type the following command to specify the nodes:

```
cadmin --set-dhcp-bootfile --node node_ID ipxe
```

For *node_ID*, specify the identifier of the node that did not boot. For example, for a compute node, specify its hostname.

To verify whether a node is enabled to load iPXE first, type the following command:

```
cadmin --show-dhcp-bootfile
```

YAST Navigation

The following list shows SLES YAST 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 YAST interface.

Subnetwork Information

This chapter contains the following topics:

About the Cluster Subnetworks

Cluster 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 a cluster.

This appendix section includes VLAN information for cluster systems. The VLAN information for the MIC devices pertains only to cluster systems that include MIC devices on the compute nodes. Within this appendix section, the subnetworks are named as you see them in the cluster configuration tool's **Subnet Network Address** menu.

SGI Rackable Head VLAN Ethernet Network Configurations

An SGI Rackable cluster 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. The VLAN configuration file is `/opt/sgi/lib/discover-ice-backend`.

shows an SGI Rackable cluster with three VLANs.

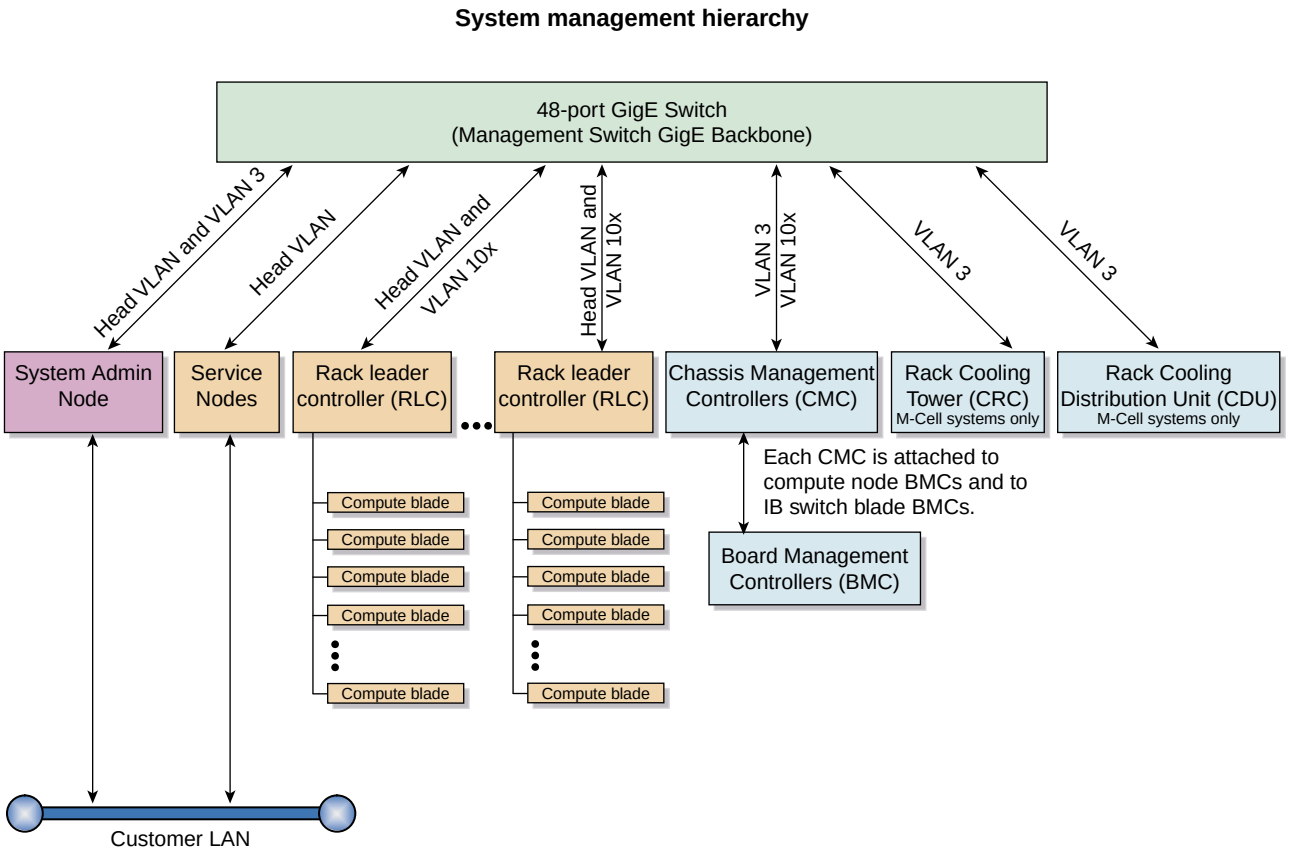

Figure B-1 SGI Rackable VLAN Logical Overview

If your SGI Rackable cluster is equipped with MIC devices,

shows how the components in an SGI Rackable cluster connect to the master switch stack.

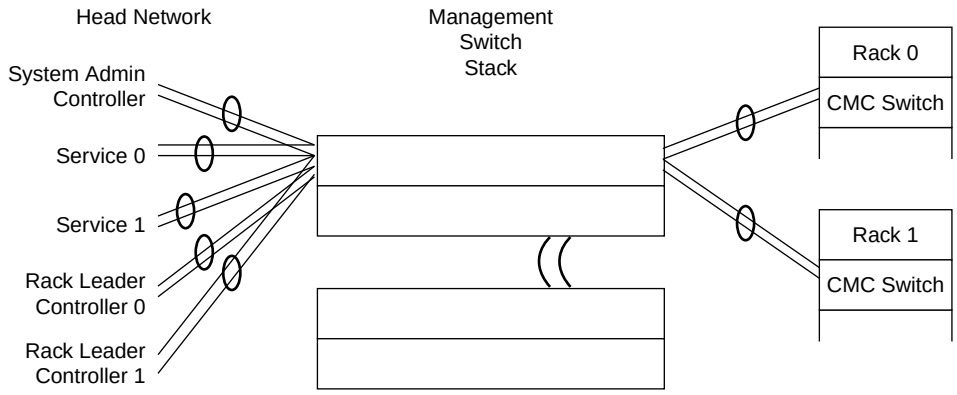

Figure B-2 SGI Rackable Physical Topology Management Switch Connections

The following topics contain more information about SGI Rackable VLANs:

•

SGI Rackable Head VLAN Configuration

The head VLAN is VLAN 1. The head VLAN includes the admin node and all compute nodes. The head VLAN is always configured as untagged. Any untagged packets coming into an admin node or compute node are associated with the head VLAN. In the cluster configuration tool's menus, the head VLAN appears as head.

SGI Rackable Additional Head Network VLAN Configurations

There is one rack VLAN for each rack in the system. The VLANs are numbered incrementally. The VLAN for rack 1 is `101`. The VLAN for rack 2 is `102`. The VLAN for rack 99 is `199`. At the maximum, the VLAN number for rack 1000 is `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.

SGI ICE X Head VLAN Ethernet Network Configurations

An SGI ICE X cluster 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. The VLAN configuration file is `/opt/sgi/lib/discover-ice-backend`.

Figure B-3 on page 177 shows an SGI ICE system with three VLANs.

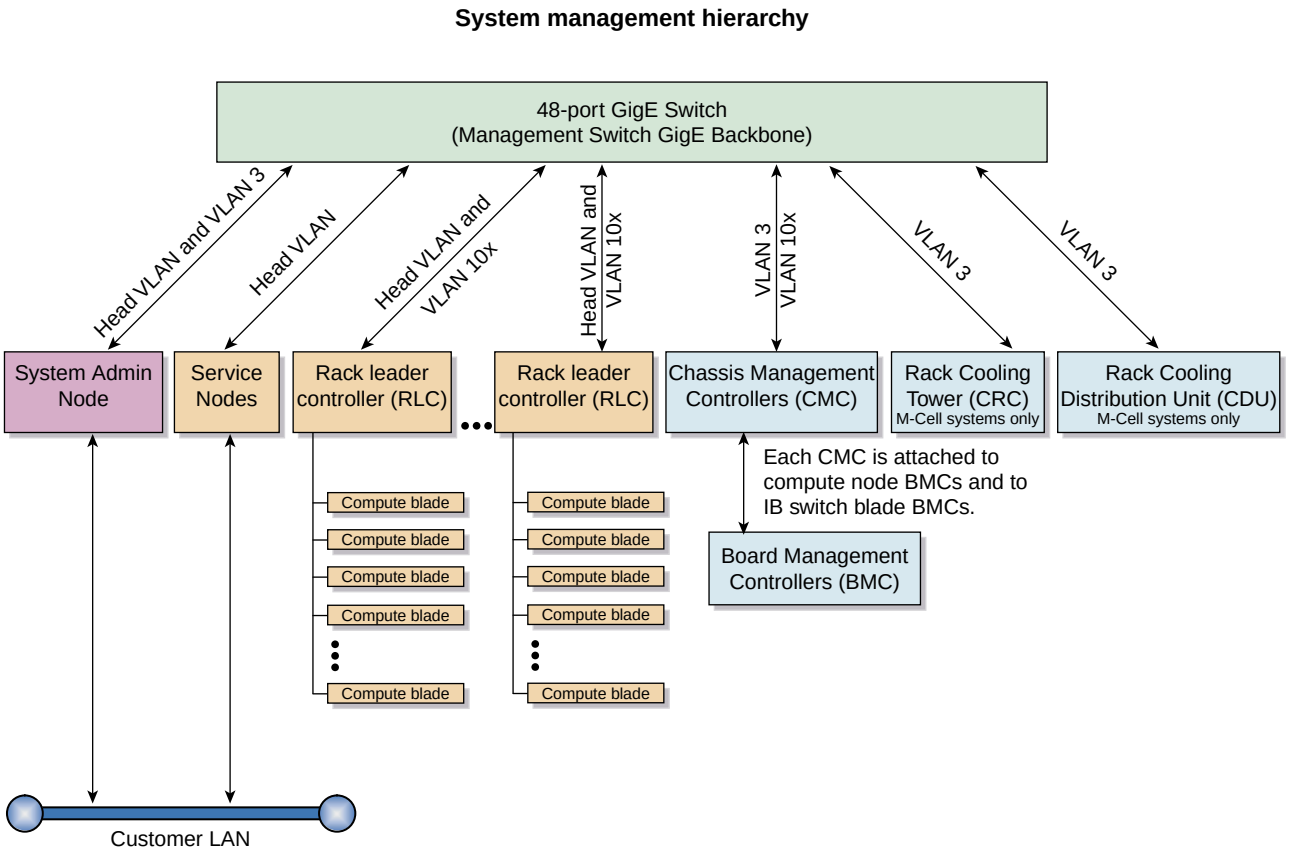

Figure B-3 SGI ICE X VLAN Logical Overview

If your SGI ICE X system is equipped with MIC devices, each compute blade includes one or two IP addresses for each device. Only one cable connects each compute blade to the network, but each MIC device requires its own, unique IP address on your network.

Figure B-4 on page 178 shows how the components in an SGI ICE system connect to the master switch stack.

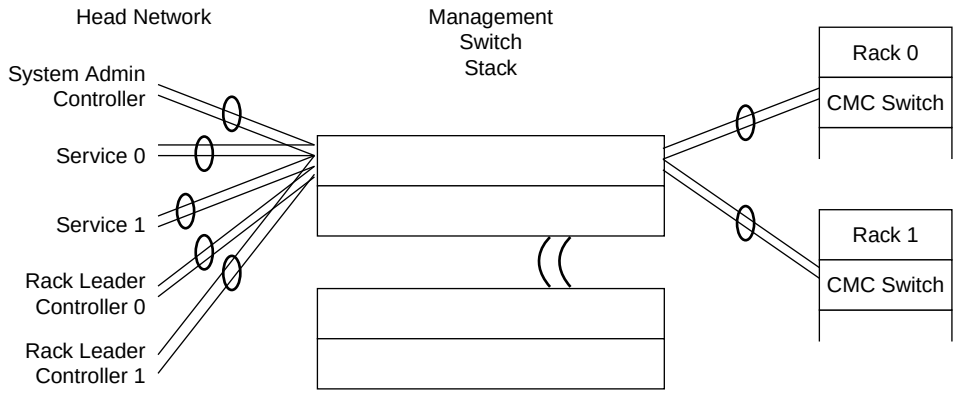

Figure B-4 SGI ICE X Physical Topology Management Switch Connections

The following topics contain more information about the head and rack VLANs:

- "SGI ICE X Head VLAN Configuration" on page 178
- "SGI ICE X Rack VLAN Configurations" on page 178
- "SGI ICE X MCell Cooling VLAN Configurations" on page 179

SGI ICE X Head VLAN Configuration

The head VLAN is VLAN 1. The head VLAN includes the admin node, all rack leader controllers (RLCs), and all flat compute nodes. The head VLAN is always configured as untagged. Any untagged packets coming into an admin node, RLC, or service node port are associated with the head VLAN. In the cluster configuration tool's menus, the head VLAN appears as `Head`.

SGI ICE X Rack VLAN Configurations

There is one rack VLAN for each rack in the system. The VLANs are numbered incrementally. The VLAN for rack 1 is `101`. The VLAN for rack 2 is `102`. The VLAN for rack 99 is `199`. At the maximum, the VLAN number for rack 1000 is `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.

SGI ICE X MCell Cooling VLAN Configurations

If the SGI ICE X cluster is equipped with MCells, the MCell cooling system has its own VLAN. This VLAN resides on your system only if you have MCells.

Address Ranges and VLANs for Management and Application Software

Table B-1 on page 180 shows the system-wide IP address ranges that the cluster management software uses. The following notes pertain to this table:

- The `head_bmc` network is a separate IP subnetwork.
- MIC devices, if present, are included on the `head` node network.

Table B-1 System-wide IP Address Ranges for the Head Network

VLAN	Subnetwork Name	IP Range	Nodes
1	`head`	172.23.0.0/16	Admin Node and MICs Compute nodes and MICs RLCs and MICs
1	`head_bmc`	172.24.0.0/16	Admin Node BMC Compute node BMCs RLC BMCs

Table B-2 on page 180 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

Rack VLAN Number	Subnetwork Name	IP Range	Components
`101`	`gbe`	10.159.1.0/24	Rack 1's RLC and MICs Rack 1's CMCs and MICs Rack 1's compute nodes and MICs
`101`	`bmc`	10.160.1.0/24	Rack 1's RLC's BMC Rack 1's CMCs' BMC Rack 1's compute nodes' BMCs

Rack VLAN Number	Subnetwork Name	IP Range	Components
102	gbe	10.159.2.0/24	Rack 2's RLC and MICs Rack 2's CMCs and MICs Rack 2's compute nodes and MICs
102	bmc	10.160.2.0/24	Rack 2's RLC's BMC Rack 2's CMCs' BMC Rack 2's compute nodes' BMCs
X	gbe	10.159.*X*.0/24	Rack *X*'s RLC and MICs Rack *X*'s CMCs and MICs Rack *X*'s compute nodes and MICs
X	bmc	10.160.*X*.0/24	Rack *X*'s RLC's BMC Rack *X*'s CMCs' BMCs Rack *X*'s compute nodes' BMCs

shows the additional subnetworks that are used on large SGI Rackable configurations of

Table B-3 on page 181 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 and MICs Some RLCs and MICs Compute nodes and MICs
IB1	ib-0	10.149.0.0/16	Service nodes and MICs Some RLCs and MICs Compute blades and MICs

Component Naming Conventions

SMC commands enable you to perform some procedures on only one component or on a range of similar components. Addressing methods differ depending on the command, 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: admin node, RLC, and so on.
- `temponame`. The system-wide unique identifier for the component as it appears in the `hostname1` field in the cluster definition file.
- `hostname1`. The host name for the component as it appears in the `hostname1` field in the cluster definition file.

Log in as the root user and type the following command to generate a copy of the cluster definition file:

`discover --show-configfile >` *out_file*

Table B-4 on page 182 explains how to specify components when you run administrative and user commands. *x* is always an integer number.

Table B-4 Naming Conventions

Component	`temponame`	`hostname1`	Examples

Admin node	N/A	Site-defined hostname	`icex1`
			`mysiteicex`
			`sleet`

Management switch	`mgmtsw`*x*	N/A	`mgmtsw0`

Component	temponame	hostname1	Examples
RLC	r*x*lead	N/A	r1lead, the RLC on the first rack r2lead, the RLC on the second rack
RLC BMC	r*x*lead-bmc	N/A	r2lead-bmc, the BMC on the second rack
SGI ICE compute node	r*x*i*x*n*x* r*x*i*x*n*x*-eth r*x*i*x*n*x*-mic0 r*x*i*x*n*x*-mic1	N/A	r1i3n10, the compute node on the first rack, in the fourth IRU, in position 10
SGI ICE compute node BMC	r*x*i*x*n*x*-bmc	N/A	r1i3n10-bmc, the BMC on the first rack, in the fourth IRU, in position 10
SGI Rackable (flat) compute node	service*x* service*x*-mic0 service*x*-mic1 service*x*-mic2 service*x*-mic3	n*x* n*x* n*x* n*x*-mic n*x*-mic	service0, the first compute node service3, the fourth compute node
SGI Rackable (flat) compute node BMC	service*x*-bmc	N/A	service1-bmc, the BMC on the second service node
InfiniBand switch	ibswitch*x*-bmc	N/A	ibswitch1-bmc, the BMC on the second InfiniBand switch
CMC	r*x*i*x*c	N/A	r1i1c, the CMC for the first rack, in the second IRU

System Control Configuration

The following topic describes the system control configuration for SGI clusters:

SGI ICE X System Control Configuration

The system control network for an SGI ICE 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 admin node, 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-5 on page 185 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 admin node, RLCs, and CMCs. Figure B-6 on page 186 shows the switches in a nonredundant management network configuration.

Figure B-5 on page 185 shows a redundant management network cascaded switch configuration.

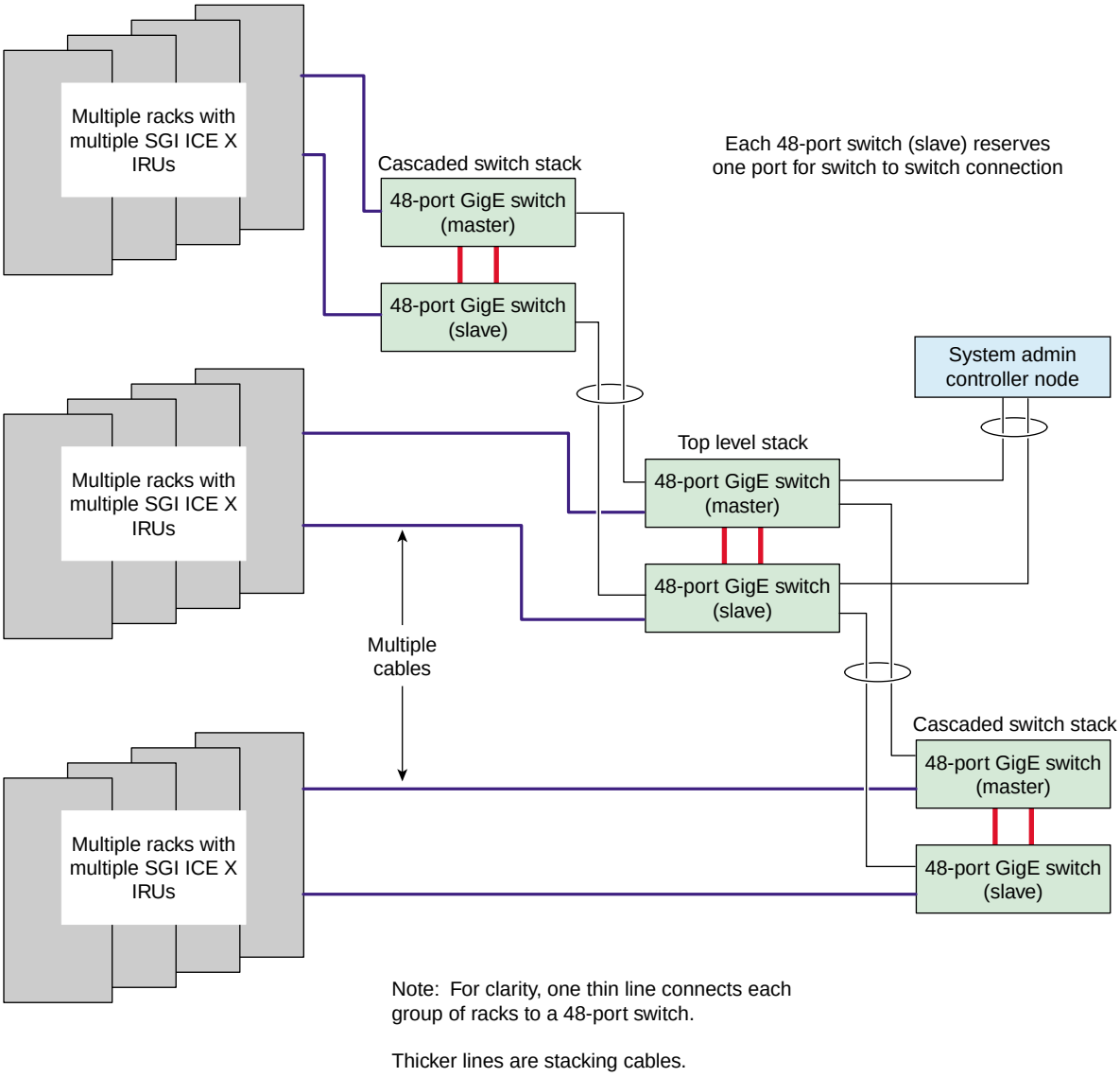

Figure B-5 Redundant Cascaded Switch Configuration

Figure B-6 on page 186 shows a nonredundant management network cascaded switch configuration.

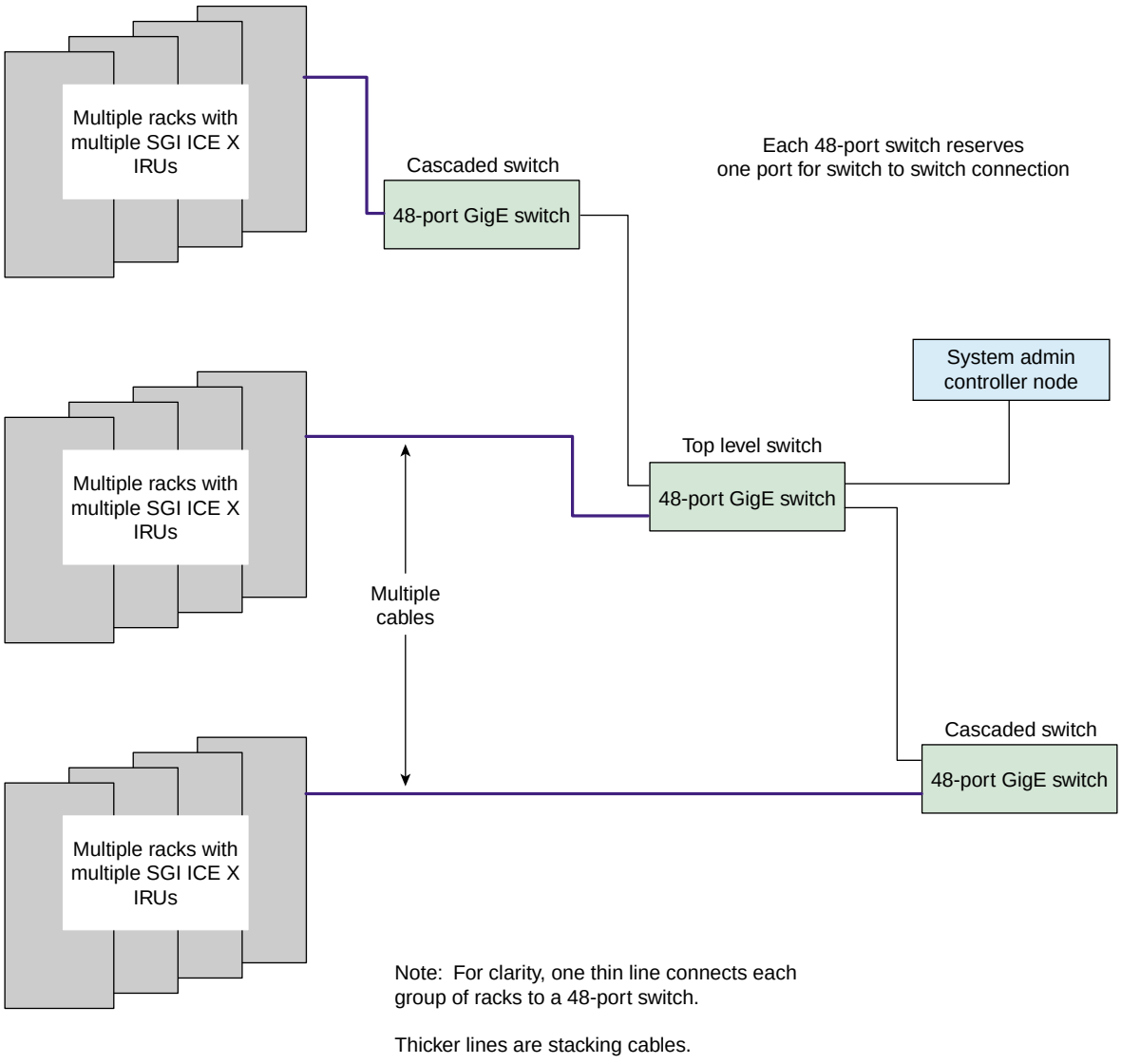

Figure B-6 Nonredundant Cascaded Switch Network Configuration

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*.

SGI ICE X MCell Network IP Addresses

If you need to troubleshoot the MCell cooling equipment on an SGI ICE X cluster, 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 189 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

Partition Layout Information

This appendix section incudes the following topics:

About the Partition Layout on SGI Clusters

As of the SGI Tempo 2.9.0 release, the partition layout scheme changed for SGI ICE X systems. The SMC 3.0 and later releases use the partition layout scheme that SGI Tempo 2.9.0 introduced.

If you upgrade an SGI ICE X cluster to an SGI Tempo 2.9.0 or later release or to an SMC 3.0 or later release, it is possible for your system to have the following mixture of slots:

- Slots with the legacy, SGI Tempo pre-2.9.0 partition scheme
- Slots with the SGI Tempo 2.9.0 or later or the SMC 3.0 or later partitioning scheme

The SGI Tempo 2.9.0 release introduced a partition layout that uses the GUID partition table (GPT) and a new boot system, which is GRUB version 2. In previous SGI Tempo versions, the partition layout was the MSDOS layout, and the boot system was GRUB version 1. These changes affect system operations in the following ways:

- If your cluster was originally installed with is SGI Tempo 2.9.0 or later or with SMC 3.0. or later, then the software on your cluster uses the new partition layout. You can use the `cadmin` command, as shown in the SGI documentation, to manage all the slots.
- If you upgrade a slot from an SGI Tempo release that is earlier than the SGI Tempo 2.9.0 release, note the following:
  - If you upgrade any slot to the current SGI Tempo or SMC release, the admin node converts the boot loader installed in the master boot record (MBR) to GRUB version 2. This boot loader chooses a slot when you boot the admin node.

  - If you do not upgrade all the slots to SGI Tempo 2.9.0 or later or to SMC 3.0 or later, the only administrative action you can perform on the slots with the older SGI Tempo software is to boot the slot. You can boot the slots that host older versions of SGI Tempo at the console or by selecting the slot on the GRUB version 2 boot menu. Because the admin node MBR is now upgraded to GRUB version 2, you can no longer use the `cadmin` command to manage any slots that host older versions of SGI Tempo.
  - You cannot install SGI Tempo releases older than SGI Tempo 2.9.0 on any of the system's slots. If you perform any future, from-scratch installations, these installations must be to SGI Tempo 2.9.0 or later or to SMC 3.0 or later. You can, however, use the `updatetempo` upgrade script to upgrade slots to releases other than SGI Tempo 2.9.0 or SMC 3.0.

  Disks with the new GPT layout can exist on the system alongside disks with the MSDOS partition layout. On a cluster system newly installed with SGI Tempo 2.9.0 or later or with SMC 3.0 or later, the software configures the GPT layout on all system disks and all disks that you clear, except for the SGI ICE compute nodes, if present. After you upgrade one or more slots to SGI Tempo 2.9.0 or later or to SMC 3.0 or later, if you add new system disks or you clear existing system disks, the software configures the GPT layout on system disks that you add and on all disks that you clear, except for the compute nodes.

- For the admin node, rack leader controller (RLC), and flat compute nodes, when you install SGI Tempo 2.9.0 or later or SMC 3.0 or later with blank system disks, or with system disks that have been cleared, the software uses the new GPT layout. If the disks have the old MSDOS partition table format, the software continues to use the old MSDOS partition table format until you manually clear the disks and reinstall (if desired).

The following topics explain the partition layout for slots installed with the current release and for slots installed with legacy software:

- "About the Current Release's Partition Layout" on page 196
- "About the Legacy Partition Layout" on page 200

About the Current Release's Partition Layout

The following topics show the partition layout that the current release creates:

- "Partition Layout for a One-slot Cluster" on page 197

Partition Layout for a One-slot Cluster

Table D-1 on page 197 shows the partition layout for a one-slot cluster. This layout 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 D-1 Partition Layout for a Single-boot Cluster

Partition	File System Type	File System Label	Notes
1	Ext3	`sgidata`	Contains slot information. On the admin node, contains GRUB version 2 data for choosing root slots at boot time.
2	swap	`sgiswap`	Swap partition.
3-10	N/A	N/A	N/A
11	Ext3	`sgiboot1`	Slot 1 `/boot` partition.
12-20	N/A	N/A	N/A
21	VFAT	`sgiefi`	`/boot/efi` partition. EFI BIOS clusters only. On x86_64 BIOS clusters, this partition is unused.
22-30	N/A	N/A	N/A
31	Ext3 on admin node and flat compute nodes. XFS on RLCs.	`sgiroot1`	Slot 1 `/` partition.

Partition Layout for a Two-slot Cluster (Default)

Table D-2 on page 198 shows the partition layout for a two-slot cluster. This layout yields two boot partitions.

Table D-2 Partition Layout for a Dual-boot Cluster (Default Layout)

Partition	File System Type	File System Label	Notes
1	Ext3	`sgidata`	Contains slot information. On the admin node, contains GRUB version 2 data for choosing root slots at boot time.
2	swap	`sgiswap`	Swap partition.
3-10	N/A	N/A	N/A
11	Ext3	`sgiboot1`	Slot 1 `/boot` partition.
12	Ext3	`sgiboot2`	Slot 2 `/boot` partition.
13-20	N/A	N/A	N/A
21	VFAT	`sgiefi`	Slot 1 `/boot/efi` partition. EFI BIOS clusters only. On x86_64 BIOS clusters, this partition is unused.
22	VFAT	`sgiefi2`	Slot 2 `/boot/efi` partition. EFI BIOS clusters only. On x86_64 BIOS clusters, this partition is unused.
23-30	N/A	N/A	N/A
31	Ext3 on admin node, flat compute nodes. XFS on RLCs.	`sgiroot1`	Slot 1 `/` partition.
32	Ext3 on admin node, flat compute nodes. XFS on RLCs.	`sgiroot2`	Slot 2 `/` partition.

Partition Layout for a Five-slot Cluster

Table D-3 on page 199 shows the partition layout for a five-slot cluster. This layout yields five boot partitions.

Table D-3 Partition Layout for a Quintuple–boot Cluster

Partition	File System Type	File System Label	Notes
1	Ext3	`sgidata`	Contains slot information. On the admin node, contains GRUB version 2 for choosing root slots at boot time.
2	swap	`sgiswap`	Swap partition.
3-10	N/A	N/A	N/A
11	Ext3	`sgiboot1`	Slot 1 `/boot` partition.
12	Ext3	`sgiboot2`	Slot 2 `/boot` partition.
13	Ext3	`sgiboot3`	Slot 3 `/boot` partition.
14	Ext3	`sgiboot4`	Slot 4 `/boot` partition.
15	Ext3	`sgiboot5`	Slot 5 `/boot` partition.
16-20	N/A	N/A	N/A
21	VFAT	`sgiefi`	Slot 1 `/boot/efi` partition. EFI BIOS clusters only. On x86_64 BIOS clusters, this partition is unused.
22	VFAT	`sgiefi2`	Slot 2 `/boot/efi` partition. EFI BIOS clusters only. On x86_64 BIOS clusters, this partition is unused.
23	VFAT	`sgiefi3`	Slot 3 `/boot/efi` partition. EFI BIOS clusters only. On x86_64 BIOS clusters, this partition is unused.
24	VFAT	`sgiefi4`	Slot 4 `/boot/efi` partition. EFI BIOS clusters only. On x86_64 BIOS clusters, this partition is unused.

Partition	File System Type	File System Label	Notes
25	VFAT	`sgiefi5`	Slot 5 `/boot/efi` partition. EFI BIOS clusters only. On x86_64 BIOS clusters, this partition is unused.
26-30	N/A	N/A	N/A
31	Ext3 on admin node, flat compute nodes. XFS on RLCs.	`sgiroot1`	Slot 1 `/` partition.
32	Ext3 on admin node, flat compute nodes. XFS on RLCs.	`sgiroot2`	Slot 2 `/` partition.
33	Ext3 on admin node, flat compute nodes. XFS on RLCs.	`sgiroot3`	Slot 3 `/` partition.
34	Ext3 on admin node, flat compute nodes. XFS on RLCs.	`sgiroot4`	Slot 4 `/` partition.
35	Ext3 on admin node, flat compute nodes. XFS on RLCs.	`sgiroot5`	Slot 5 `/` partition.

About the Legacy Partition Layout

The following topics show the legacy partition layouts for various slot configurations:

Legacy Partition Layout for a One-slot SGI ICE X Cluster

Table D-4 on page 201 shows the legacy partition layout for a one-slot SGI ICE X cluster. This layout yields one boot partition.

Table D-4 Partition Layout for a Single-boot SGI ICE X Cluster

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 on admin node, service nodes. XFS on RLCs.	`sgiroot`	`/` partition.

Legacy Partition Layout for a Two-slot SGI ICE X Cluster

Table D-5 on page 201 shows the legacy partition layout for a two-slot SGI ICE X cluster. This layout yields two boot partitions.

Table D-5 Partition Layout for a Dual-boot SGI ICE X Cluster (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.

Partition	File System Type	File System Label	Notes
6	ext3 on admin node, service nodes. XFS on RLCs.	`sgiroot`	Slot 1 `/` partition.
7	ext3	`sgiboot`	Slot 2 `/boot` partition.
8	ext3 on admin node, service nodes. XFS on RLCs.	`sgiroot`	Slot 2 `/` partition.

Legacy Partition Layout for a Five-slot SGI ICE X Cluster

Table D-6 on page 202 shows the legacy partition layout for a five-slot SGI ICE X cluster. This layout yields five boot partitions.

Table D-6 Legacy Partition Layout for a Quintuple-boot SGI ICE X Cluster

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.

Partition	File System Type	File System Label	Notes
12	ext3 or XFS	`sgiroot`	Slot 4 `/` partition.
13	ext3	`sgiboot`	Slot 5 `/boot` partition.
14	ext3 or XFS	`sgiroot`	Slot 5 `/` partition.

Specifying Configuration Attributes

This appendix includes the following topics:

About Configuration Attributes

SGI cluster configuration information can be specified in several ways. For example:

- When you configure the cluster for the first time, you can provide configuration information in the cluster definition file or you can provide information by responding to the prompts in the online cluster configuration tool.
- When you add nodes to a cluster, you can specify node attributes as parameters to the `discover` command that you use to configure the nodes.
- When you use the `cadmin` command you set and apply an attribute.
- When you use the `cattr` command, you set an attribute.

SMC supports several global cluster attributes, and some attributes can be specified in more than one way. The following topics describe each attribute, show each attribute's default value, show additional other accepted values or ranges of values, and show the commands or files in which you can specify the value:

UDPcast Options

The following configuration attributes control UDPcast operations:

edns_udp_size

Specifies the edns-udp-size option in /etc/named.conf. This value is the default packet size, in bytes. This is the packet size that remote servers can receive.

Default = 512.

Values = any positive integer number.

Accepted by:

- cattr command

udpcast_max_bitrate

Specifies the maximum numbers of bits that are conveyed or processed per unit of time, expressed as a number followed by a unit of measure.

Default = 900m.

Values = any positive integer number followed by one of the following: m (megabytes), ...

Accepted by:

- Cluster definition file
- `cadmin` command
- `cattr` command
- `discover` command

udpcast_max_wait

Specifies the greatest amount of time that can elapse between when the first client node connects and any other client nodes connect. Clients that connect after this time has elapsed receive their software in a subsequent broadcast.

Default = `10`.

Values = any positive integer number.

Accepted by:

- Cluster definition file
- `cadmin` command
- `cattr` command
- `discover` command

udpcast_mcast_rdv_addr

Specifies the UDPcast rondezvous (RDV) address. Used for senders and receivers to find each other.

Note that the admin node default address and the global (leader node) default address are different. If you change the global setting, which is used by leaders, also make the following changes:

- Adjust the `--set-udpcast-rexmit-hello-interval` .
- Use the `cimage` command to push an image and initiate changes on the RLCs.

Default for the admin node = `239.0.0.1`.

Default for the global (leaders) = `224.0.0.1`.

Values = any valid IP address.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cadmin` command
- `cattr` command
- `discover` command

udpcast_min_receivers

Specifies the minimum number of receiver nodes for UDPcast.

Default = `1`.

Values = any positive integer number.

Accepted by:

- Cluster definition file
- `cadmin` command
- `cattr` command
- `discover` command

udpcast_min_wait

Specifies the minimum amount of time that the system waits, while allowing clients to connect, before the software broadcast begins. This is the time between when the first client node connects and any other client nodes connect. The UDPcast distributes the software to all clients that connect during this interval.

Default = `10`.

Values = any positive integer number.

Accepted by:

- Cluster definition file
- `cadmin` command
- `cattr` command
- `discover` command

udpcast_rexmit_hello_interval

Specifies the frequency with which the UDP sender transmits `hello` packets.

Note: NOTE: The admin node has a different default than the RLCs.

Default for the admin node = `5000` (5 seconds).

Default for the global (leaders) = `0`.

Values = any positive integer number.

Accepted by:

- Cluster definition file
- `cadmin` command
- `cattr` command
- `discover` command

udpcast_ttl

Sets the UDPcast time to live (TTL).

Default = `1`.

Values = any positive integer number.

Accepted by:

- Cluster definition file
- `cattr` command

- `discover` command

VLAN and General Network Options

The following configuration attributes control the VLAN configuration and other aspects of the cluster network:

- "`head_vlan`" on page 210
- "`mcell_network`" on page 211
- "`mcell_vlan`" on page 211
- "`mgmt_vlan_end`" on page 211
- "`mgmt_vlan_start`" on page 212
- "`rack_vlan_end`" on page 212
- "`rack_vlan_start`" on page 213
- "`redundant_mgmt_network`" on page 213
- "`switch_mgmt_network`" on page 213

`head_vlan`

Specifies the number of the head network VLAN. SGI recommends that you do not change this value.

Default = `1`.

Range = `1` <= *arg* <= `4096`.

Accepted by:

- Cluster definition file
- `cattr` command

mcell_network

Specifies whether the cluster includes MCells. This value must be set to `yes` when the cluster includes MCell cooling equipment. This value can be set to `yes` or `no` for clusters that do not include MCells.

Values = `yes` (default) or `no`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cadmin` command
- `cattr` command

mcell_vlan

Specifies the cooling network VLAN. SGI recommends that you do not change this value.

Default = `3`.

Range = `1` <= *arg* <= `4096`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cattr` command

mgmt_vlan_end

Specifies the last flat compute node rack VLAN. Use caution when changing this value. Take care not to overlap other VLAN settings.

Default = `2500`.

Range = `1` <= *arg* <= `4096`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cattr` command

`mgmt_vlan_start`

Specifies the first flat compute node rack VLAN. Use caution when changing this value. Take care not to overlap other VLAN settings.

Default = `2001`.

Range = `1` <= *arg* <= `4096`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cattr` command

`rack_vlan_end`

Specifies the last SGI ICE X rack VLAN. Use caution when changing this value. Take care not to overlap other VLAN settings.

Default = `1100`.

Range = `1` <= *arg* <= `4096`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cattr` command

rack_vlan_start

Specifies the first SGI ICE X rack VLAN. Use caution when changing this value. Take care not to overlap other VLAN settings.

Default = `101`.

Range = `1` <= *arg* <= `4096`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cattr` command

redundant_mgmt_network

Specifies the default setting for the redundant management network. If no value is supplied to the `discover` command at configuration time, the installer populates the node(s) attribute(s) with this value.

Values = `yes` (default) or `no`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cadmin` command
- `cattr` command
- `discover` command

switch_mgmt_network

Specifies the default setting for the switch management network. If no value is supplied to the `discover` command at configuration time, the installer populates the node(s) attribute(s) with this value. Must be set to `no` when discovering 8200/8400 racks.

Values = `yes` (default) or `no`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cadmin` command
- `cattr` command
- `discover` command

Console Server Options

The admin node and the rack leader controller (RLC) nodes manage other nodes. On an SGI Rackable cluster, the admin node manages flat compute nodes. On an SGI ICE X system the admin node manages RLCs and flat compute nodes, and the RLCs manage SGI ICE compute nodes.

On the management nodes (the admin node and the RLCs), there are files in the `/var/log/consoles` directory for each node that the admin node or the RLC manages. The files contain log information about the baseboard management controller (BMC) on each node under the admin node's or RLC's control.

The console server options let you control the quantity and frequency of log information that is collected. SMC logs BMC output to the `/var/log/consoles` directory. In the `/var/log/consoles` directory, there is a file for each node in the cluster. If you tune the console server options, you can limit the amount of traffic between the console and the cluster. Set these options if you need to minimize network contention.

The following options control conserver operations:

- "`conserver_logging`" on page 214
- "`conserver_ondemand`" on page 215

`conserver_logging`

Specifies conserver logging. If set to `yes`, the conserver logs messages to the console via IPMItool. This feature uses some network bandwidth.

Values = `yes` (default) or `no`.

Accepted by:

- Cluster definition file
- `cadmin` command
- `cattr` command
- `discover` command

`conserver_ondemand`

Specifies conserver logging frequency. When set to `no`, logging is enabled all the time. When set to `yes`, logging is enabled only when someone is connected.

Values = `yes` or `no` (default).

Accepted by:

- Cluster definition file
- `cadmin` command
- `cattr` command
- `discover` command

Miscellaneous Options

The following are miscellaneous options that control the cluster:

- "`blademond_scan_interval`" on page 216
- "`cluster_domain`" on page 216
- "`dhcp_bootfile`" on page 217
- "`discover_skip_switchconfig`" on page 217
- "`max_rack_irus`" on page 217
- "`mic`" on page 218

- "my_sql_replication" on page 218
- "tempo_dhcp_option" on page 219

blademond_scan_interval

Specifies how often the blade monitor detects changes in the blades. For example, if you are changing blades or doing other blade maintenance, you could set this option to a very high value so the daemon does not run during the maintenance period.

Specifies the sleep time for the `blademond` daemon. The daemon waits the specified number of seconds in between checking if the CMC slot maps have changed.

Default = `120`.

Values = can be `0` or any positive integer.

Accepted by:

- Cluster definition file
- `cadmin` command
- `cattr` command

cluster_domain

Specifies the cluster domain name. SGI recommends that users change this value.

Default = `smc-default-americas.sgi.com`.

Values = must be a standard domain name.

Accepted by:

- Cluster definition file
- Cluster configuration tool
- `cadmin` command
- `cattr`

dhcp_bootfile

Specifies that if `ipxe` is selected, the server boot agent loads iPXE instead of GRUB2, and then iPXE loads GRUB2.

Values = `grub2` (default) or `ipxe`.

Accepted by:

- Cluster definition file
- `cadmin` command
- `cattr` command
- `discover` command

discover_skip_switchconfig

Signals the installer to omit the switch configuration steps. When set to `yes`, the installer does not configure the switches. Set this to `yes` when you want to perform a quick configuration change, but you do not need to update the switch configuration. This value is not saved in the cluster definition file, but it can be specified there.

Values = `yes` or `no` (default).

Accepted by:

- Cluster definition file
- `cadmin` command
- `cattr` command
- `discover` command

max_rack_irus

Specifies the maximum number of IRUs in the cluster. When you use the `discover` command, the installer autopopulates the database with this value during the rack configuration. This value is not saved in the cluster definition file, but it can be specified there.

Default = `8`.

Range = `1` <= *arg* <= `16`.

Accepted by:

- Cluster definition file
- Cluster configuration tool
- `cadmin` command
- `cattr` command
- `discover` command

mic

Specifies the default number of Intel Phi (MIC) devices on each compute node. The default is 4 or fewer for flat compute nodes. The default is 2 or fewer for SGI ICE X compute nodes. This value is not saved in the cluster definition file, but it can be specified there.

Values = `0` (default), `1`, `2`, `3`, `4`.

Accepted by:

- Cluster definition file
- `cattr` command
- `discover` command

my_sql_replication

Specifies if MySQL replication is disabled or enabled on any node. If set to `yes`, MySQL replication is enabled. Unless otherwise specified, this value is set on each node when the `discover` command runs. If you want to enable MySQL replication for any node, you need to make sure that MySQL replication is set to `yes` on the admin node.

For more information about MySQL replication, see the following:

SGI Management Center Administration Guide for Clusters

Values = `yes` (default) or `no`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cadmin` command
- `cattr` command
- `discover` command

tempo_dhcp_option

Resets the DHCP option code to a cluster-specific option code.

The nodes in the cluster accept DHCP leases that belong only to the cluster. By default, this is `149`. If your site's DHCP server is also configured to use option code `149`, you need to change this value. To determine your site's DHCP option code, type the following command:

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

The DHCP option code used on your site network and the DHCP option code used within the cluster need to be different. If these codes are not different, the installation can fail when the cluster nodes mistake your site's DHCP server for one of the nodes in the cluster.

Default = `149`.

Range = `149`, .

Accepted by:

- `cadmin` command
- `cattr` command

Index