



# HPE Performance Cluster Manager Administration Guide

## Abstract

This publication describes how to use the HPE Performance Cluster Manager 1.0, patch 1, software to administer and maintain HPE cluster systems.

Part Number: 007-6499-002  
Published: June 2018  
Edition: 1

## **Notices**

The information contained herein is subject to change without notice. The only warranties for Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.

Confidential computer software. Valid license from Hewlett Packard Enterprise required for possession, use, or copying. Consistent with FAR 12.211 and 12.212, Commercial Computer Software, Computer Software Documentation, and Technical Data for Commercial Items are licensed to the U.S. Government under vendor's standard commercial license.

Links to third-party websites take you outside the Hewlett Packard Enterprise website. Hewlett Packard Enterprise has no control over and is not responsible for information outside the Hewlett Packard Enterprise website.

## **Acknowledgments**

Intel®<sup>®</sup>, Itanium®<sup>®</sup>, Xeon®<sup>®</sup>, and Xeon Phi™<sup>™</sup> are trademarks of Intel Corporation in the U.S. and other countries.

Linux®<sup>®</sup> is the registered trademark of Linus Torvalds in the U.S. and other countries.

Oracle®<sup>®</sup> and Lustre®<sup>®</sup> are registered trademarks of Oracle and/or its affiliates.

Red Hat®<sup>®</sup> and Red Hat Enterprise Linux®<sup>®</sup> are registered trademarks of Red Hat, Inc., in the United States and other countries.

# Contents

<b>Administering a cluster with HPE Performance Cluster Manager.....</b>	<b>11</b>
<b>Administering the cluster with GUI actions and commands in the style of the HPE Insight Cluster Manager Utility.....</b>	<b>13</b>
<b>HPE Performance Cluster Manager GUI operations.....</b>	<b>14</b>
Launching the GUI.....	14
<b>Managing nodes.....</b>	<b>17</b>
Scanning nodes.....	17
Adding nodes.....	19
Modifying nodes.....	21
Importing nodes.....	21
Deleting nodes.....	22
Exporting node information to a flat text file.....	22
Retrieving node information from the GUI.....	23
Retrieving node information from the command line.....	23
Replacing a node and scanning in the new node.....	24
Adding network groups and deleting network groups.....	24
Adding network groups.....	24
Deleting network groups.....	25
Uploading files to the admin node.....	25
<b>Provisioning leader nodes and flat compute nodes with an operating system.....</b>	<b>26</b>
Creating and managing image groups.....	27
Creating image groups.....	27
Adding nodes to an image group.....	28
Deleting autoinstall image groups.....	28
Renaming image groups.....	28
Autoinstalling an operating system image.....	29
Preparing to autoinstall an operating system image.....	29
Autoinstalling with the GUI.....	32
Autoinstalling with the CLI.....	34
Completing the autoinstall.....	36
Troubleshooting and special cases.....	37
Capturing an image from a leader node or flat compute node using the GUI.....	41
Provisioning an image.....	41
Provisioning a leader node or a flat compute node using the GUI.....	42
Reviewing the success of the provisioning operation.....	43
<b>Monitoring a cluster.....</b>	<b>44</b>
Monitoring and cluster security.....	44
(Conditional) Installing the monitoring client.....	45

Monitoring the cluster.....	45
Node and group status.....	46
Changing the pingStaleDelay timeout value.....	47
Right pane display.....	47
Stopping monitoring.....	60
Customizing monitoring, alerting, and reactions.....	60
Action and alert files.....	60
Modifying sensors, alerts, and alert reactions.....	65
Restarting the monitoring daemons.....	65
Using <code>collectl</code> to gather monitoring data .....	66
Monitoring NVIDIA GPUs.....	71
Monitoring cluster manager alerts in HPE Systems Insight Manager.....	73
Extended metric support.....	74
Metric arrays.....	85
Changing the monitoring interval.....	89
Transitional display.....	90

## **Managing a cluster.....** 92

Administrator menu.....	92
SSH connection.....	93
Changing the default gateway IP address of a node from the GUI.....	93
Management card connection.....	94
Virtual serial port connection.....	94
Shutdown.....	94
Power off.....	95
Boot.....	95
Reboot.....	95
Change UID LED status.....	96
Multiple windows broadcast.....	96
PDSH (using cmdiff) - single window <code>pdsh</code> .....	97
PDPC (Distributed Copy) - parallel distributed copy.....	101
Custom group management.....	101
Adding custom groups.....	102
Deleting custom groups.....	102
Firmware management.....	102
Viewing and analyzing BIOS settings.....	103
Checking BIOS versions.....	104
Installing and upgrading firmware.....	104
Requesting firmware versions.....	104
Saving user settings.....	105
CLI actions.....	105
Starting a CLI interactive session.....	105
Basic commands.....	105
Specifying nodes.....	108
Administration and deployment commands.....	109
Administration utilities <code>pdcp</code> and <code>pdsh</code> .....	115
Linux shell commands.....	115

## **Advanced topics.....** 116

Enabling nonroot user access to commands and GUI actions.....	116
Enabling nonroot users to run commands based on <code>pdsh</code> .....	116
Granting nonroot users the ability to run cluster manager commands.....	117
Granting nonroot users the ability to use the GUI to manage the cluster.....	118
Configuring the GUI to work with <code>sudo</code> .....	120

Commands and GUI actions.....	120
Customizing the cluster manager commands.....	123
Customizing netboot kernel arguments on flat compute nodes.....	124
Remote hardware control API (flat clusters only).....	126
Support for ScaleMP.....	127
Ansible integration.....	128
<b>Troubleshooting.....</b>	<b>129</b>
Retrieving cluster manager service status information.....	129
Log files.....	129
cmuserver log files .....	129
Monitoring log files.....	129
Network boot problems.....	130
Troubleshooting switch problems.....	130
Troubleshooting network boot.....	132
Administration command problems.....	133
GUI problems.....	133
GUI cannot be launched from browser.....	133
GUI cannot contact the remote cluster manager service.....	134
GUI is running, but the monitoring sensors are not updated.....	135
Failed to validate certificate error displays.....	135
<b>Manpages.....</b>	<b>137</b>
<b>Administering the cluster in the style of SGI Management Center</b>	<b>138</b>
<b>Configuring optional features on flat compute nodes.....</b>	<b>139</b>
Configuring ICE compute nodes to use a flat compute node as a network address (NAT) gateway.....	139
Configuring a flat compute node as an NFS server.....	140
Configuring scratch disk space on system disks.....	140
Configuring scratch disk space on leader and flat compute nodes.....	141
Configuring scratch disk space on an admin node.....	142
Configuring software RAID on cluster nodes.....	142
Configuring software RAID on an admin node.....	143
Configuring software RAID on leader nodes and flat compute nodes.....	143
Configuring trusted boot nodes .....	145
Prerequisites and constraints.....	145
Configuring trusted boot nodes at configuration time.....	146
Configuring trusted boot nodes after the initial installation and configuration....	147
Building trusted boot node images.....	147
Verifying that a node booted in trusted mode.....	148
Managing trusted boot nodes.....	149
<b>System operation.....</b>	<b>150</b>
Changing global cluster configuration settings.....	150
Changing the network time protocol (NTP) server.....	151
Changing the site domain name service (DNS) server information.....	151
Enabling or disabling a backup domain name service (DNS) server .....	152
Configuring a redundant management network (RMN).....	152

Configuring the <code>blademond</code> rescan interval .....	154
discover command .....	154
Using the generic hardware type .....	155
Configuring a compute node to use a nondefault image.....	155
Skipping a node while configuring.....	156
Omitting unneeded switch configurations when reconfiguring.....	156
Managing slots.....	157
Retrieving slot information.....	157
Booting from a different slot.....	157
Cloning a slot.....	158
Customizing slot labels.....	159
Modifying boot options.....	160
Powering on and powering off cluster systems and cluster system components.....	160
Using the <code>cpower</code> command .....	161
Power commands for the entire cluster.....	164
Power commands for ICE compute nodes and flat compute nodes.....	165
Managing rack leaders.....	166
Managing ICE compute IRUs.....	167
Managing ICE compute blade switches.....	167
Power and energy management.....	168
<code>pdsh</code> and <code>pdcp</code> commands .....	168
<code>pdsh</code> command examples.....	169
Creating custom <code>pdsh</code> group files .....	169
Using the <code>cadmin</code> command, the administrative interface .....	170
Bringing a node online or setting a node offline.....	170
Creating and displaying node-specific notes.....	171
Changing compute node configuration elements.....	171
Changing the admin node hostname and IP address on the house network .....	172
Displaying network information.....	173
Changing switch management network settings.....	174
Changing console management settings.....	174
Managing UDP multicast (UDPcast) provisioning.....	175
Console management.....	179
Synchronizing system time.....	180
Admin node NTP.....	180
Leader node NTP.....	181
BMC setup with NTP.....	181
Compute node NTP.....	181
ICE compute node NTP.....	181
NTP workarounds.....	181
Booting leaders or flat compute nodes from a local disk.....	182
Disjoint boot mode.....	182
Admin-assisted boot mode.....	182
Changing the size of <code>/tmp</code> on ICE compute nodes .....	183
Switching ICE compute nodes to a <code>tmpfs</code> root .....	183
Switching flat compute nodes to a <code>tmpfs</code> root .....	184
Configuring flat compute nodes to boot a <code>tmpfs</code> root file system on RHEL 7 nodes.....	184
Configuring flat compute nodes to boot a <code>tmpfs</code> root file system on SLES 12 nodes.....	186
Configuring local storage space for swap and scratch disk space .....	187
Retrieving the status of a local storage space setting .....	189
Enabling, disabling, or respecifying a local storage space setting .....	190
Using the <code>cattr</code> command to modify system attributes .....	191
About disk quotas.....	192
Retrieving quota information.....	192

Setting quotas.....	194
Viewing the ICE compute node read/write quotas.....	195
Creating custom partitions.....	196
Custom partitioning notes, constraints, and cautions.....	196
Configuring custom partitioning for flat compute and leader nodes .....	197
Configuring custom partitioning for admin nodes.....	199
Managing custom partitions.....	199
Backing up and restoring the system database.....	199
Backing up the cluster database.....	200
Restoring the cluster database.....	201
Enabling EDNS.....	201

## **Managing software images.....** 202

About cluster images.....	202
Node types and default image names.....	202
Image management commands.....	204
About installation repositories.....	205
Repository metadata.....	205
Remote repositories.....	206
General repository management parameters.....	206
RPM lists.....	207
Adding and updating software images.....	207
Adding software to the cluster manager repository database.....	208
Standard repository.....	208
Custom repository.....	209
Multiple media sources.....	209
Nested repositories.....	209
Selecting the repositories to be used in the installation.....	209
Explicit selection using the <code>crepo</code> command .....	210
Selection using a group assignment.....	210
Selection using the <code>cinstallman</code> command .....	211
Creating images to host new software.....	211
Cloning an existing <code>cinstallman</code> image .....	212
Using the cluster manager version control system (VCS).....	212
Capturing an image from a running compute node.....	212
Installing new software into new images.....	213
Installing packages from repositories into an image.....	213
Installing miscellaneous RPMs into an image.....	214
Installing packages from repositories onto running compute nodes or leader nodes.....	215
Installing miscellaneous RPMs onto running compute nodes or leader nodes...	215
Associating nondefault images with targeted nodes.....	215
Associating a nondefault image with flat compute nodes and leader nodes.....	216
Associating a nondefault image with ICE compute nodes.....	216
Pushing images from the admin node to the targeted nodes.....	217
Pushing images to flat compute nodes and leader nodes.....	217
Pushing images to ICE compute nodes.....	218
Miscellaneous image management tasks.....	218
Using a custom repository for site packages.....	219
Creating images in an environment with multiple operating systems.....	221
Comparing the image on a running node with images on the admin node.....	223
Performing ICE compute node per-host customization.....	224
Changing the services on the ICE compute nodes.....	224
Using the <code>cimage</code> command to manage ICE compute node images .....	225
Using <code>cinstallman</code> to install packages into software images .....	227

<b>Using the version control system.....</b>	<b>231</b>
VCS terminology.....	231
Creating images.....	232
Managing clones.....	232
Committing the working copy.....	232
Reverting the working copy to a specified revision.....	232
Reviewing revision history.....	233
Reviewing changes between revisions and the working copy.....	233
Amending a commit message.....	233
Removing revisions.....	234
VCS examples.....	234
Adding a revision and querying changes.....	234
Reverting to a previous revision.....	235
Cloning an image.....	236
Deleting all revisions permanently.....	237
<b>Fabric management.....</b>	<b>239</b>
Omni-Path fabric management.....	240
Starting and stopping the Omni-Path fabric managers.....	240
Managing Omni-Path fabric software.....	240
InfiniBand fabric management.....	242
InfiniBand fabric overview (hierarchical clusters).....	242
Using the InfiniBand management tool GUI.....	243
Fabric management commands.....	245
Automatic InfiniBand fabric management.....	248
Network topology.....	248
Utilities and diagnostics for Omni-Path fabrics and InfiniBand fabrics.....	249
ibstat and ibstatus commands (Omni-Path and InfiniBand) .....	249
perfquery command (InfiniBand) .....	250
ibnetdiscover command (InfiniBand) .....	250
opareport command (Omni-Path) .....	251
ibdiagnet command (InfiniBand) .....	251
Logging and debugging options.....	252
<b>System monitoring.....</b>	<b>253</b>
Hardware event tracker (HET) notifications.....	253
About HET.....	253
Customizing the default HET notification script.....	254
Using environment variables to create a site-specific HET notification .....	254
HET example.....	256
Ganglia.....	256
Accessing the Ganglia system monitor.....	257
Monitoring system metrics.....	257
Hardware event logs.....	258
Heartbeat daemon.....	259
Nagios.....	260
Accessing Nagios.....	260
Examining the cluster components configured for Nagios monitoring .....	261
Performance Co-Pilot.....	263
Monitoring SDR metrics.....	263
Starting the pmgcluster cluster performance monitor (flat clusters) .....	264

<b>System maintenance and troubleshooting.....</b>	<b>265</b>
Hardware maintenance procedures.....	265
Taking one ICE compute node or flat compute node offline for maintenance temporarily.....	265
Taking one leader node in a highly available (HA) leader node configuration offline for maintenance temporarily.....	266
Replacing a failed blade.....	267
Replacing a management switch.....	267
Node replacement process for cold spares.....	270
Ensure that the spare is an appropriate replacement.....	270
Identify the failed unit and replace it.....	270
Troubleshooting IRU power up and automatic power down problems (hierarchical clusters).....	274
About the power on process (hierarchical clusters).....	275
CMC monitoring (hierarchical clusters).....	275
Power cycling the IRUs (hierarchical clusters).....	275
Power supplies and the watchdog timer (hierarchical clusters).....	276
Interpreting the power supply LEDs (hierarchical clusters).....	277
Troubleshooting the devices on the CAN bus interface (hierarchical clusters only).....	277
Flashing the firmware on a power shelf or fan controller (hierarchical clusters).....	278
Troubleshooting a missing power shelf (hierarchical clusters).....	279
Power consumption log files (hierarchical clusters).....	282
Retrieving information about the power supplies (hierarchical clusters).....	282
Retrieving information about the PMBus registers (hierarchical clusters).....	284
Miscellaneous troubleshooting tools.....	284
cm_info_gather command .....	284
cminfo command .....	285
kdump utility .....	285
Retrieving system firmware information.....	289
Booting a flat compute node or a leader node on an installed cluster.....	290
Phase 1 - Initiating the boot.....	290
Phase 2 - Loading the kernel for the node.....	291
Phase 3 - Loading the miniroot.....	292
Phase 4 - Starting the operating system on the node.....	293
Overriding installation scripts.....	293
<b>Security features.....</b>	<b>295</b>
Secret creation.....	295
Packaging and file residence.....	295
Recreating secrets.....	296
Secure provisioning.....	296
Safeguards against unauthorized requests.....	296
Image encryption and authentication.....	297
Restricted node-to-node login access.....	297
Cluster manager ssh zones .....	298
Default images and customizing your ssh configuration .....	299
Security recommendations.....	299
Cluster manager database security.....	299
<b>Using Singularity containers.....</b>	<b>300</b>
Installing the container software.....	300

Building a container.....	301
Running containers.....	304
<b>Hierarchical cluster system configuration framework information.....</b>	<b>306</b>
About the hierarchical cluster system configuration framework.....	306
About the cluster configuration repository.....	308
Automatic updates.....	308
Custom configuration scripts.....	308
Preserving custom configuration changes.....	309
<b>YaST navigation.....</b>	<b>310</b>
<b>Support and other resources.....</b>	<b>311</b>
Accessing Hewlett Packard Enterprise Support.....	311
Accessing updates.....	311
Customer self repair.....	312
Remote support.....	312
Warranty information.....	312
Regulatory information.....	313
Documentation feedback.....	313
<b>Websites.....</b>	<b>314</b>

# Administering a cluster with HPE Performance Cluster Manager

This guide is for system administrators of the HPE Performance Cluster Manager. Use the information in this guide to administer and maintain HPE clusters. The cluster manager release notes include a list of the specific hardware and operating systems that the cluster manager supports. This documentation typically uses the terms **flat cluster** and **hierarchical cluster**. These clusters are as follows:

- The HPE Apollo 480 is an example of a flat cluster. A flat cluster includes the following types of nodes:
  - Admin node
  - Flat compute nodes
- The HPE SGI 8600 is an example of a hierarchical cluster. A hierarchical cluster includes a leader node, which adds a communication hierarchy to the cluster system. A hierarchical cluster includes the following types of nodes:
  - Admin node
  - Leader nodes
  - ICE compute nodes, which communicate to the admin node through a leader node
  - Flat compute nodes

For more information about cluster terminology, see the glossary in the following:

## [HPE Performance Cluster Manager Getting Started Guide](#)

To determine which release version of the cluster manager is installed, enter the following command on the admin node:

```
# rpm -q sgi-admin-node-release
```

---

**NOTE:** Unless otherwise noted, assume that information that pertains to RHEL platforms also pertains to CentOS platforms.

---

The following list shows the cluster manager documentation:

- The HPE Performance Software - Cluster Manager Release Notes. To access the release notes, follow the links on the following website:  
<https://www.hpe.com/software/hpcm>
- [HPE Performance Cluster Manager Getting Started Guide](#)
- [HPE Performance Cluster Manager Installation Guide](#)
- [HPE Performance Cluster Manager Power Management Guide](#)

Hardware documentation might also interest you. The hardware documentation typically includes the following:

- A system architecture overview
- A description of the major components
- Standard procedures for powering on and powering off the system
- Basic troubleshooting information
- Important safety and regulatory specifications

# Administering the cluster with GUI actions and commands in the style of the HPE Insight Cluster Manager Utility

The chapters that follow explain how to administer the cluster with GUI actions and commands in the style of the HPE Insight Cluster Manager Utility.

# HPE Performance Cluster Manager GUI operations

The cluster manager GUI is a Java application. You can download a copy of the GUI client to a computer that is connected to the admin node over your site network. For example, you can download the client to a laptop or desktop computer.

To close an unwanted dialog window, press the `ESC` key.

For information about how to download and install a GUI client on your laptop or desktop computer, see the following:

- [Launching the GUI](#) on page 14
- [HPE Performance Cluster Manager Installation Guide](#)

## Launching the GUI

### Procedure

1. Open a web browser on your client computer.
2. In the browser address bar, enter the address of the cluster admin node.

Use the following format:

`https://admin_node_addr`

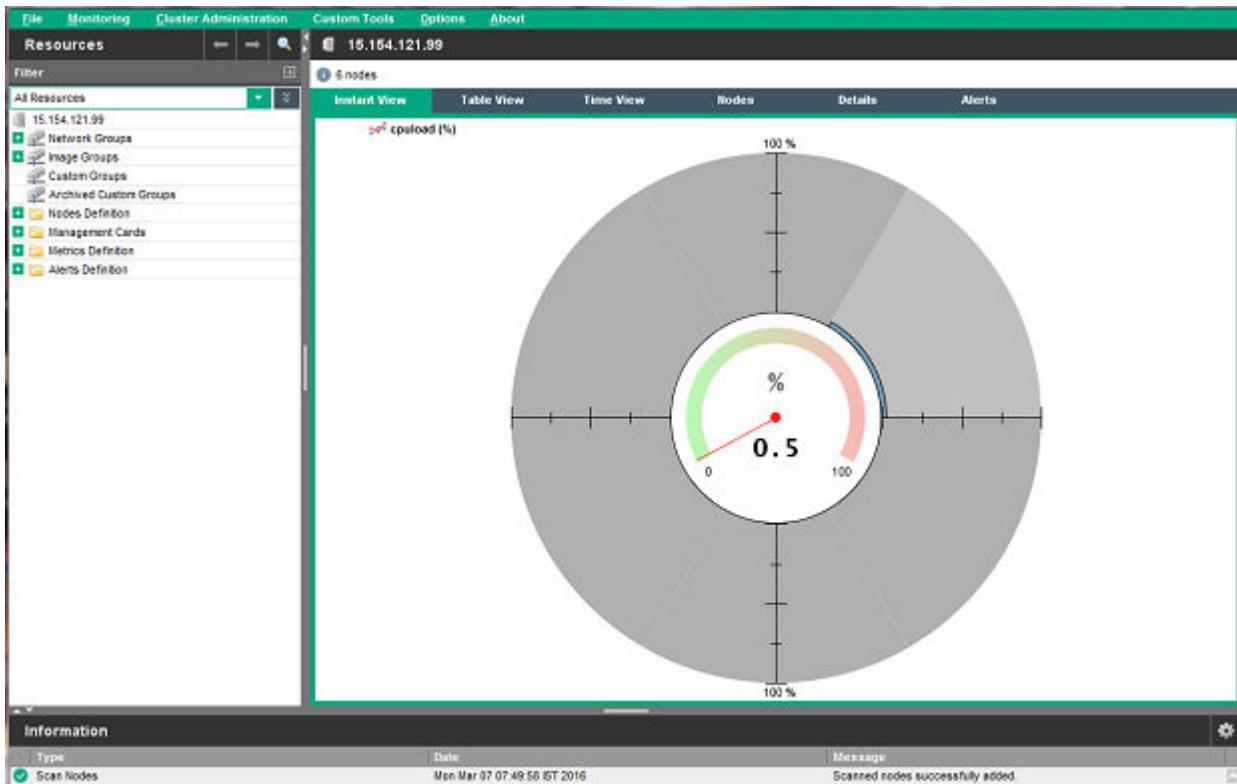
For `admin_node_addr`, specify the IP address or fully qualified domain name of the cluster admin node.

3. On the cluster manager splash page, click **Launch Cluster Management Utility GUI**.
4. Follow the onscreen prompts to launch the GUI.

Depending on your configuration, the cluster manager might prompt you to enter more information. For example:

- The cluster manager might prompt you to enter the IP address of the admin node.
- Your client computer might have more than one network interface. In this case, the cluster manager might prompt you to enter the correct network interface to use for communication with the admin node.

5. Observe the main window and notice the functional areas.



**Figure 1: Main window**

The main window contains the following functional areas:

- The top bar. This bar allows you to click or select GUI actions.
- The left pane. This pane lists resources such as **Network Groups**, **Image Groups**, and **Nodes Definitions**.

Click the + button to expand a resource.

The left pane contains a filter that allows you to select resources for display.

- The right pane, which displays the global cluster view.
- The bottom pane, which displays log information.

#### 6. (Optional) Click **Options > Properties** to adjust the GUI to suit your viewing preferences.

There are several settings you can adjust to change font sizes, colors, and other aspects of the visual representation.

If you adjust any settings, exit the GUI and restart the GUI. The cluster manager records your preferences and reloads them the next time you launch the GUI.

#### 7. (Optional) Assume the administrator role.

Administrator privileges enable you to perform cluster configuration tasks. You can perform cluster configuration tasks on only one instance of the GUI at a time. Complete the following steps:

- Click **Options > Enter Admin mode**.
- On the popup window that appears, provide the cluster administrator login credentials and click **OK**.

To exit administrator mode, click **Options > Leave Admin mode**.

The GUI is a Java application. You can use Java Web Start to download the GUI from the web server that runs on the admin node. Then, manually copy the GUI Java file onto the client.

The cluster manager automatically starts a minimal web server on port 80 of the admin node. Port 80 serves only the GUI. If an HTTP service is already running on admin node port 80, the web service does not run. To specify a different port number, edit the `CMU_THTTPD_PORT` environment variable in the following file:

`/opt/clmgr/etc/cmuserver.conf`

# Managing nodes

The following procedure explains how to display GUI features for node management.

## Procedure

### 1. Click Cluster Administration > Node Management

In the left pane, observe that **Nodes** is highlighted.

### 2. In the left pane, click **Nodes**.

Observe that the information in the right pane updates to show possible node actions. For example, adding, deleting, modifying nodes.

### 3. Click the + next to **Nodes** to expand the node list.

### 4. Select a node in the node list.

Observe that the information in the right pane changes to display information about the selected node.

### 5. Right-click the selected node.

Observe the actions available in this contextual menu. For example, you can boot the node or display BIOS settings.

# Scanning nodes

When you click **Scan Node** button in the **Node Management** window, the cluster manager scans the selected nodes into the cluster database, retrieves the MAC address, hardware addresses, and configures IP addresses.

You can also manually add node information. For information about how to add nodes manually, see the following:

### [Adding nodes](#) on page 19

## Prerequisites

Use the **Scan Node** method for adding nodes to the cluster when you want to add one or more nodes in a batch. This method requires the following:

- Each iLO IP address is statically (and incrementally) preassigned.
- The iLO credentials are already configured on each iLO and in the cluster manager. If yours is a factory-configured cluster, be aware that the HPE factory typically configures these credentials during the assembly process.

## Procedure

### 1. Click Cluster Administration > Node Management > Scan Node

### 2. Complete the **Scan Node** dialog box and click **OK**.

For help on each parameter, click the question mark (?) button.

**Scan Node**

Management card type :	iLOCM	?
Number of iLCCM to scan :	1	?
First iLOCM IP :	10.117.23.101	?
Node name pattern :	m700_c%2c-n%	?
example : node%3i , n_%3L-%3c-%n , node%4i_%L-%c-%n		
Initial value of %i token :	1	?
<input type="checkbox"/> Use custom %i increment for undiscovered cartridge		
Custom %i token increment :		
First node IP :	10.117.23.1	?
Subnet mask :	255.255.0.0	?
NIC index :	1	?
Bios Boot Mode :	AUTO	?
CMU server IP address :	default	?
Default Gateway IP :	default	?
ISCSI Root Boot String :	none	?
<b>Clear</b>		<b>OK</b>
		<b>Cancel</b>

**Figure 2: Scan node**

Notes:

- If the cluster has Moonshot cartridges based on ARM64 architecture, install the appropriate cluster manager add-on packages before scanning those nodes.
- Make sure to select the correct management card (also known as the baseboard management controller (BMC)) type for your compute nodes (iLO/iLOCM).

3. In the confirmation window, click **OK**.

4. On the **Management card password** window, complete the following steps:

- Enter the login name for the management card (also known as the baseboard management controller (BMC)).
- Enter the password for the management card.
- Click **OK**.

The **Management card password** window appears only once, for the first scan operation. For subsequent scans, the **Management card password** window does not appear.

5. On the **Scan Nodes Result** window, select **Add to cluster** to add the scanned nodes.

## Adding nodes

The **Add Node Dialog** window adds one node to the cluster.

You can also scan several nodes and add them to the database in a batch. For information about how to add nodes, see the following:

[\*\*Scanning nodes\*\* on page 17](#)

### Prerequisites

Know the MAC address of the node you want to add to the cluster.

### Procedure

1. Click **Cluster Administration > Node Management**.
2. In the right pane, click **Add Node**.

The following dialog box displays.

**Add Node Dialog**

Hostname:	cn005
IP Address:	10.0.0.5
Subnet Mask:	255.255.0.0
Ethernet Mac Address:	38-EA-A7-A6-DA-D4
Management Card:	iLO
Management Card IP address:	10.1.0.105
Architecture:	x86_64
Cartridge ID:	
Node ID:	
Platform:	generic
Serial Port:	default
Serial Port Speed:	default
Vendor Arguments:	default
Cloning Block Device:	default
Bios Boot Mode:	AUTO
CMU server IP address:	default
Default Gateway IP Address:	default
ISCSI Root Boot String:	none

**OK**   **Cancel**

**Figure 3: Add node dialog**

3. Complete the fields in the dialog box and click **OK**.

Notice that when adding Moonshot cartridges based on ARM64, select the appropriate architecture (aarch64) using the **Architecture** field in the **Add Node Dialog** window.

A dialog box displays the successful addition of a node completion.

4. Click **OK**.

A dialog box asks if you want to add another node.

5. Add the new node or nodes to a network group.

Proceed to the following:

[\*\*Adding network groups\*\*](#) on page 24

# Modifying nodes

## Procedure

1. Click **Cluster Administration > Node Management**.
2. In the left pane, locate the node you want to modify.  
If necessary, click **+** to expand the node group that includes the node.
3. Right-click the node and select **Manage > Modify Node**.  
The **Modify Node Dialog** window displays.
4. Update information for the node in the **Modify Node Dialog** box, and click **OK**.  
You cannot change the name of the node in the **Modify Node Dialog** dialog box.

# Importing nodes

You can import node information from a flat text file. You can manually create and edit this file. Incorrect formatting can break the operation.

All imported nodes belong to the default image group.

## Procedure

1. Create a text file that describes the node information for the node that you want to import into the cluster database.

Within the file, enter information for an individual node all on one line. The fields in the line are as follows:

Node name  
IP address  
Subnet mask  
MAC address  
Image name  
BMC IP address  
BMC type  
Architecture  
Cartridge ID  
Node ID  
Platform name  
Serial port  
Serial port speed  
Vendor arguments  
Cloning bock device  
BIOS mode  
Management server IP address  
Default gateway  
ISCSI root  
BMC MAC address

For example:

```
n0 10.117.31.2 255.255.255.0 c8-cb-b8-cb-d4-c6 rhel7.4 10.117.30.2 IPMI  
x86_64 -1 -1 generic ttys1 115200 "default" default auto default default  
"none" c8-cb-b8-cb-d4-c9  
n1 10.117.31.3 255.255.255.0 c8-cb-b8-cb-d5-96 rhel7.4 10.117.30.3 IPMI  
x86_64 -1 -1 generic ttys1 115200 "default" default auto default default  
"none" c8-cb-b8-cb-d5-99  
n2 10.117.31.4 255.255.255.0 c8-cb-b8-c6-c8-b8 rhel7.4 10.117.30.4 IPMI  
x86_64 -1 -1 generic ttys1 115200 "default" default auto default default  
"none" c8-cb-b8-c6-c8-bb
```

2. Click **Cluster Administration > Node Management**.
3. In the right pane, click **Import Nodes**.
4. Browse to the text file you created.
5. Click **Open** to add the nodes from the file to the cluster.
6. In the left pane, right-click the cluster name and select **Update Configs**.

## Deleting nodes

### Procedure

1. Click **Cluster Administration > Node Management**.
2. In the left pane, locate the node you want to delete.  
If necessary, click **+** to expand the node group that includes the node.
3. Right-click the node and select **Manage > Delete Node(s)**.  
After you delete a node, it cannot be recovered.

## Exporting node information to a flat text file

You can export node information to a text file. When you store the text file on a system outside the cluster, you have a backup copy of the cluster node information.

### Procedure

1. Click **Cluster Administration > Node Management**.
2. In the right pane, select one or more nodes to export.
3. Click **Export Nodes**.
4. In the **Export** dialog box, complete the following steps:
  - In the **File Name** field, enter a name for the exported file.
  - Click **Export**.
5. Save the file.

# Retrieving node information from the GUI

## Procedure

1. Click **Cluster Administration > Node Management**
2. In the left pane, locate the node you want to delete.  
If necessary, click **+** to expand the node group that includes the node.
3. Select a node.
4. In the right pane, click **Details**.

# Retrieving node information from the command line

## Procedure

1. Log into the admin node as the root user.
2. Enter one of the following commands:

- `/opt/clmgr/bin/cmu_add_attribute`
- `/opt/clmgr/bin/cmu_del_attribute`
- `/opt/clmgr/bin/cmu_show_attributes`

The preceding commands allow you to add and manage node information. For the command syntax, see the manpages or run these commands with `-h`.

The `cmu_show_attributes` command syntax mimics the output of the `pdsh` command. This design means that you can pipe command output it through the `cmu_diff` command to get a condensed display of all the node static information.

For example:

```
# cmu_show_attributes | cmu_diff
-----
Responses: 7 { n[0-5],o184i098 }
Reference: n0 - 20 lines
3 groups of similar nodes found:
  1 { n0 }
  5 { n[1-5] }
  1 { o184i098 }

Ignored:
<none>
-----

d  m| BIOS = HP-P75-12/01/2014          | (all different) not present in 5: n[1-5]
14% > BIOS = HP-P75-12/01/2014          | x   1: n0
14% > MCELL_NETWORK = no                | x   1: o184i098
d  | CONSERVER_LOGGING = yes            | not present in 1: o184i098
d  | CONSERVER_ONDEMAND = no            | not present in 1: o184i098
d  | CPU model = Intel(R) Xeon(R) CPU E5-2695 v2 @ 2.40GHz | not present in 6: n[1-5],o184i098
d  | Cluster dns domain name = ib0.cm.gre.smktg.hpecorp.net | not present in 6: n[1-5],o184i098
d  | DHCP_BOOTFILE = grub2              | not present in 1: o184i098
d  | DISK_BOOTLOADER = no               | not present in 1: o184i098
d  | Disk size = 2794 GB               | not present in 6: n[1-5],o184i098
d  | Disk type = ATA                  | not present in 6: n[1-5],o184i098
d  | Logical CPU number = 24          | not present in 6: n[1-5],o184i098
d  | MGMT_BONDING = active-backup    | not present in 1: o184i098
d  | Memory size = 64133 MB           | not present in 6: n[1-5],o184i098
d  | Native CPU speed = 2400 MHz      | not present in 6: n[1-5],o184i098
```

```

| PREDICTABLE_NET_NAMES = yes
m| REDUNDANT_MGMT_NETWORK = yes
85% > REDUNDANT_MGMT_NETWORK = yes
14% > REDUNDANT_MGMT_NETWORK = no
| SWITCH_MGMT_NETWORK = yes
d m| System model = ProLiant SL230s Gen8
14% > System model = ProLiant SL230s Gen8
14% > UDPCAST_MCAST_RDV_ADDR = 224.0.0.1
m| TEMPO_CPOWER = yes
85% > TEMPO_CPOWER = yes
14% > UDPCAST_PORTBASE = 9000
m| TPM_BOOT = no
85% > TPM_BOOT = no
14% > UDPCAST_REXMIT_HELLO_INTERVAL = 0
a+m| dns_domain = cm.gre.smktg.hpecorp.net
85% > dns_domain = cm.gre.smktg.hpecorp.net
14% > UDPCAST_TTL = 1
|
| (2 populations)
| x 6: n[0-5]
| x 1: o184i098
|
| (all different) not present in 5: n[1-5]
| x 1: n0
| x 1: o184i098
| (2 populations)
| x 6: n[0-5]
| x 1: o184i098
| (2 populations)
| x 6: n[0-5]
| x 1: o184i098
| (2 populations)
| x 6: n[0-5]
| x 1: o184i098
| (2 populations) more lines after in 1: o184i098
| x 6: n[0-5]
| x 1: o184i098

```

---

## Replacing a node and scanning in the new node

### Procedure

1. Use your cluster hardware documentation to replace the failed node.
2. Click **Cluster Administration > Node Management**
3. In the left pane, locate the node you want to scan.  
If necessary, click **+** to expand the node group that includes the node.
4. Right-click the node, and select **Refresh > Rescan MAC address**.

## Adding network groups and deleting network groups

A **network group** consists of one of the following:

- Flat compute nodes attached to a common switch.
- ICE compute nodes attached to a common leader node. These groups also include the associated chassis management controllers (CMCs) and InfiniBand switches.

A cluster can have multiple network groups. The following topics explain how to manage network groups:

- [Adding network groups](#) on page 24
- [Deleting network groups](#) on page 25

## Adding network groups

For the monitoring functions to work, each compute node must be included in a network group. The cluster manager automatically includes ICE compute nodes in network groups based on the rack number. Ensure that flat compute nodes are included in a network group.

The cluster monitoring function requires that each node be included in a network group.

### Procedure

1. Click **Cluster Administration > Network Group Management**
2. In the right pane, click **Create**.

3. In the **New Network Group** window, specify a name for the network group you want to create and click **OK**.

Limit the network group name to 32 characters.

4. Complete the following actions to move nodes into network groups:

- Select any number of nodes from the **Nodes not in any Network Group** section on the left.
- Use the arrows to move the nodes to the **Nodes in Network Group** section on the right.

You can include up to 288 nodes in a single network group.

A network group must correspond to an Ethernet switch. A switch in the cluster must physically represent a network group and the associated nodes must be connected to that switch.

## Deleting network groups

You can delete network groups. After you delete a network group, you can put the nodes from the group into another network group.

### Procedure

1. Click **Cluster Administration > Network Group Management**.
2. In the right pane, use the **Select a Network Group** pull-down menu to select a network group that you want to delete.
3. Click **Delete**.

After you delete a network group:

- You cannot recover that network group.
- You can put the nodes from the network group into another network group.
- You can create a new network group with the same name as the network group you deleted.

## Uploading files to the admin node

### Procedure

1. Right click the cluster name, and select **Upload file(s) to /tmp**.
2. In the **Upload file(s)** popup, specify file or files you want to upload.
3. Click **Upload file(s)**.

The cluster manager uploads the selected files to the `/tmp` folder on the admin node, and a progress window displays.

When the upload is complete, the progress window disappears and the following message displays in the bottom information panel:

Uploading: - Operation completed

# Provisioning leader nodes and flat compute nodes with an operating system

**Provisioning** is the process by which you install an operating system image and (optionally) cluster manager files on a node. The initial cluster installation process provisions each node with an operating system and the cluster manager files. The provisioning topics in this chapter explain the following:

- Defining an image group.

An **image group** consists of the group of nodes that all host the same image. That is, they all host the same operating system image and the associated cluster manager software. After you create an image group, you can specify nodes to include in the image group. The nodes you include in an image group do not have to be included in the same network group.

For more information, see the following:

[\*\*Creating and managing image groups\*\*](#) on page 27

- Autoinstalling a node.

The autoinstall process copies operating system distribution from a single ISO to one or more leader nodes or flat compute nodes. It assumes that the ISO file resides in the admin node local storage. You can autoinstall an image that is already under cluster manager control. To determine the images that are under cluster manager control, enter the following command on the admin node:

```
# crepo --show
```

If you want to install the cluster manager RPMs, you need to do that manually.

You can use the autoinstall process in the following situations:

- You added new flat compute nodes to the cluster. Now you can install an operating system on them.
- You want to change the operating system that resides on one or more flat compute nodes.
- You want to troubleshoot an operating system problem on a node. The autoinstall process installs only the operating system on the node.

The autoinstall process uses NFS for the file transfer. You cannot use `http` or `ftp`.

For more information, see the following:

[\*\*Autoinstalling an operating system image\*\*](#) on page 29

- Capturing an image.

The capture process creates a new image from a running leader node or flat compute node. The cluster manager stores the newly captured image in a new repository on the admin node.

For more information, see the following:

[\*\*Capturing an image from a leader node or flat compute node using the GUI\*\*](#) on page 41

- Deploying an image.

Use the deploy procedure to push a copy of the captured image to all the flat compute nodes or all leader nodes in the image group.

For more information, see the following:

---

**NOTE:** Do not use the procedures in this chapter to provision ICE compute nodes. For information about how to provision ICE compute nodes, see the following:

[Managing software images](#) on page 202

---

## Creating and managing image groups

An **image group** is a group of similar nodes that run the same image. After the cluster manager is installed and the cluster is configured, the following image groups exist:

- default
- *distro*

For example, `rhel7.4`

- `ice-distro`.

This image group exists only on hierarchical clusters. For example, `ice-rhel7.4`.

- `lead-distro`

This image group exists only on hierarchical clusters. For example, `lead-rhel7.4`.

Before you provision a new operating system image on any node or nodes, create an image group for the nodes.

When a node first becomes a member of an image group, the cluster manager marks the node as an inactive candidate of the group. This designation means that the node could be deployed with the image, but the node is not currently hosting the image. This situation is likely because the node is currently running another image. After you provision a node with a new image, it becomes an active member of the image group.

If you autoinstall a node, and the process fails, the cluster manager moves the node to the default image group or to the **unassigned** image group. The cluster manager puts nodes in the default image group or the **unassigned** image group if it cannot determine what image a node is running. This group is listed as `default` in the command line interface and is displayed as **unassigned** in the GUI.

## Creating image groups

### Procedure

1. Click **Options > Enter Admin mode** and specify the cluster credentials.
2. Click **Cluster Administration > Image Group Management**.
3. In the right pane, click **Create**.
4. In the **New Image Group** popup, complete the following fields:

- In the **Group name** field, specify a group name.
- In the **Group type** field, select **diskful** or **autoinstall**.
- In the **Capture disk name** field, accept the default of `sda`.

5. Click **OK**.

6. Proceed to the following to add nodes to the image group:

[Adding nodes to an image group](#) on page 28

## Adding nodes to an image group

### Procedure

1. Click **Cluster Administration > Image Group Management**.
2. In the right pane, use the **Select an Image Group** pull-down menu to select the image group you want to update.
3. Use the **=>** and **<=** buttons to move nodes to the **Nodes in Image Group** list.  
In the list, the `[non-active]` notation follows the name of a node that has not been cloned. These nodes are cloning candidates. After the node is cloned, the notation changes to `[active]`.
4. Proceed to the following to install a new operating system on the nodes in the image group:  
[Autoinstalling an operating system image](#) on page 29

## Deleting autoinstall image groups

### Procedure

1. Click **Cluster Administration > Image Group Management**.
2. In the **Select an Image Group** pull-down, select an image group to delete.
3. Click **Delete**.

When you delete an image group, the cluster manager preserves the content of the image group in the following directory:

`/opt/clmgr/image/group_name`

## Renaming image groups

### Procedure

1. Click **Cluster Administration > Image Group Management**.
2. In the **Select an image group** pull-down, select an image group.
3. Click **Rename**.
4. In the **Rename Image Group** popup, enter a new name for the image group, and click **OK**.

# Autoinstalling an operating system image

You can use either the GUI or the command line interface to autoinstall an operating system image.

Proceed to the following topic to prepare for an autoinstall:

[Preparing to autoinstall an operating system image](#) on page 29

## Preparing to autoinstall an operating system image

### Procedure

1. Verify that the operating system distribution ISO resides in a directory on the admin node.

If the ISO is already configured in the cluster manager, enter the following command on the admin node:

```
# crepo --show
```

Note the output from the `crepo` command. If the chosen distribution does not appear in the `crepo` output, use the `crepo` command in the following format to copy the ISO to the admin node:

```
crepo --add path_to_ISO_image
```

For `path_to_ISO_image`, specify the full path to where the ISO image resides. This can be a network location.

2. Create an autoinstall image group.

For information about how to create an autoinstall image group, see the following:

[Creating image groups](#) on page 27

3. Examine the list of autoinstall template files and determine the file you want to use.

On the admin node, the following directory contains the default autoinstall template files:

```
/opt/clmgr/templates/autoinstall
```

The following notes pertain to the default template files:

- UEFI-enabled nodes require a separate FAT partition (`/boot/efi`) to boot into the operating system. To autoinstall these nodes, plan to use one of the UEFI-specific autoinstall template files.
- You can create your own autoinstall template files. Use one of the default files as the base for your own, custom file. Make sure of the following when you create a custom file:
  - The file is compatible with the software release you want to autoinstall.
  - The NFS server and repository information is configured correctly.
- The autoinstall template files for RHEL are RHEL kickstart files.  
The autoinstall template files for SLES are AutoYaST XML files.
- The templates support specific operating systems and are as follows:

**Table 1: Autoinstall template files**

Autoinstall template name	Appropriateness
autoinst_rh6_rh7.templ	RHEL 7.X and RHEL 6.X on BIOS nodes
autoinst_rh6_rh7_uefi.templ	RHEL 7.X and RHEL 6.X on UEFI-enabled nodes
autoinst_rh6_rh7_moonshot_m710.templ	RHEL 7.X and RHEL 6.X on ProLiant m710 Moonshot server cartridges
autoinst_sles12.templ	SLES 12 SPX on BIOS nodes
autoinst_sles11.templ	SLES 11 SPX on BIOS nodes
autoinst_sles12_uefi.templ	SLES 12 SPX on UEFI-enabled nodes
autoinst_sles11_uefi.templ	SLES 11 SPX on UEFI-enabled nodes

The autoinstall engine performs keyword substitutions. All keywords begin with CMU\_. The autoinstall section of the following file describes many of the keywords:

/opt/clmgr/etc/cmuserver.conf

You can review the autoinstall templates in the following directory to see how the cluster manager uses the keywords to create image-specific and node-specific unattended operating system distribution configuration files:

/opt/clmgr/templates/autoinstall

#### 4. (Optional) Modify the autoinstall variables.

The /opt/clmgr/etc/cmuserver.conf configuration file includes an autoinstall section with the following two sets of variables:

- Variables that affect the autoinstall process behavior:
  - CMU\_AUTOINST\_INSTALL\_TIMEOUT  
You can increase this value if the autoinstall times out due to a long disk formatting time.
  - CMU\_AUTOINST\_PIPELINE\_SIZE
- Variables for keyword substitution into autoinstall templates:
  - CMU\_CN\_OS\_LANG

For example, when you set CMU\_CN\_OS\_LANG=en\_US in one of the template files, lang CMU\_CN\_OS\_LANG becomes lang en\_US.

- CMU\_CN\_OS\_TIMEZONE
- CMU\_CN\_OS\_CRYPT\_PWD
- CMU\_CN\_DEFAULT\_GW

For example, you can use this variable to specify a per-node default gateway value during autoinstall or cloning. The following settings are available:

- default, which configures the admin node default gateway IP address as the gateway for the autoinstalled node.
- cmumgt, which configures cmumgt to set the admin node admin IP address as the gateway for the autoinstalled node.
- The actual IP address of the gateway

You can also use the GUI to change the default gateway IP address of a node. For more information, see the following:

[\*\*Changing the default gateway IP address of a node from the GUI\*\*](#) on page 93

## 5. (Optional) Disable predictable NIC device names.

Complete this step if you want to install a RHEL image and if your image requires legacy names.

RHEL 7.x-specific predictable NIC device names (eno1 or ens1p1) can be disabled during the autoinstall, and legacy names (eth0 or eth1) can be used.

Complete the following steps:

- a. Append the kernel command line parameter net.ifnames=0 to the CMU\_KS\_KERNEL\_PARMS in the /opt/clmgr/etc/cmuserver.conf file.

For example:

```
CMU_KS_KERNEL_PARMS="lang=CMU_CN_OS_LANG devfs=nomount ramdisk_size=10240  
console=CMU_CN_SERIAL_PORT ksdevice=CMU_CN_MAC_COLON initrd=autoinst-  
initrd-CMU_IMAGE_NAME net.ifnames=0"
```

---

**NOTE:** Modifications made to /opt/clmgr/etc/cmuserver.conf apply to all the RHEL autoinstall image groups created from that point forward.

To disable predictable NIC names for a specific RHEL 7 autoinstall group, edit the pcmlinux\_template file in the following directory:

/opt/clmgr/image/group\_name

Add net.ifnames=0 to the existing kernel parameters list.

---

- b. Modify the autoinstall template to ensure that the net.ifnames=0 parameter is persistent during the subsequent disk boots.

To the bootloader line, append --append='net.ifnames=0'.

For example:

```
#System bootloader configuration
bootloader --location=mbr --
append='net.ifnames=0'
```

**6.** (Optional) Increase the number of nodes that can be autoinstalled simultaneously.

In many cases, you autoinstall an operating system on only one flat compute node, and this node becomes the golden node. However, by default, you can autoinstall 16 nodes or fewer simultaneously. To increase this number, edit the `CMU_AUTOINST_PIPELINE_SIZE` variable in the following file:

```
/opt/clmgr/etc/cmuserver.conf
```

**7.** (Conditional) Restrict RHEL operating system partitions to one disk or LUN.

Complete this step if you want to autoinstall RHEL 7 or RHEL 6 on a node that contains multiple disks or LUNs.

By default, the RHEL 7.X and RHEL 6.X installer spreads the operating system installation across multiple disks or LUNs. As a result, certain operating system partitions, such as `/boot` and `/`, can spread across multiple disks or LUNs.

To confine the operating system installation to a single disk or LUN, edit the following file:

```
/opt/clmgr/image/group_name/autoinst tmpl-orig
```

In the file, pass the `--ondisk=disk_id` option to the partitioning commands.

For example:

```
#Disk partitioning information
#USE THE APPROPRIATE DISK NAME
part /boot --fstype ext4 --size 1000 --asprimary --ondisk=sda
part swap --size 4096 --asprimary --ondisk=sda
part / --fstype ext4 --size 1 --grow --asprimary --ondisk=sda
```

---

**NOTE:** The alternative autoinstall command `ignoredisk --only-use=sda` can be used instead of specifying the `--on-disk=sda` option for every partition command.

---

**8.** Review the special situations in the following topic and decide whether to address any of the situations at this time:

**Troubleshooting and special cases** on page 37

**9.** Proceed to one of the following:

- **Autoinstalling with the GUI** on page 32
- **Autoinstalling with the CLI** on page 34

## Autoinstalling with the GUI

### Procedure

1. Click **Options > Enter Admin mode** and enter the cluster credentials.
2. Click **Cluster Administration > Image Group Management**.
3. On the **New Image Group** popup window, complete the fields as follows and click **OK**:

- In the **Group name** field, enter a name. For example: `rh7u4_autoinstall`.  
This name becomes a directory in `/opt/clmgr/image`.
- In the **Group type** field, select **autoinstall**.
- In the **Repository path** field, enter the path to the location of the operating system distribution files or click the folder icon to browse.
- In the **Autoinstall template** field, enter the path to the autoinstall template you want to use or click the folder icon to browse.

To retrieve the path to the operating system on the admin node, enter the following command:

```
# crepo --show
```

This step creates the following new directory on the admin node:

```
/opt/clmgr/image/group_name
```

For example, it creates directory `/opt/clmgr/image/rh7u4_autoinstall` with the following files:

- `autoinst tmpl orig` - An exact copy of the autoinstall file
- `repository` - A logical link to the autoinstall repository
- `README` - More information about node-specific customization of autoinstall and PXE boot files

4. (Optional) Create one or more node-specific custom prefixes or custom autoinstall files in the autoinstall group directory.

For example, to create a PXELINUX file specific to node `n1`, create the following file:

```
/opt/clmgr/image/group_name/pclinux-n1.custom
```

Customize the file according to the guidelines in the `/opt/clmgr/image/group_name/README` file.

5. Click **Cluster Administration > Image Group Management**

6. In the right pane, use the **Select an Image Group** pull-down menu to select the autoinstall image group you created.

7. Use the `=>` and `<=` buttons to add a compute node to the autoinstall image group.

This node is the one that will host the new operating system RPMs. This node becomes the golden node.

By default, you can add 16 or fewer nodes to the autoinstall image group. In this case, multiple nodes receive the new operating system image.

To add more than 16 nodes to the autoinstall image group, see the following:

#### [Preparing to autoinstall an operating system image](#) on page 29

8. In the left pane, left-click an image group and then right-click to select **Autoinstall (Kickstart | AutoYaST | Preseed | Unattended)**.

When the autoinstall finishes, new files are added to the following directory:

```
/opt/clmgr/image/group_name
```

The new files are as follows:

- `autoinst tmpl-cmu` - A copy of your autoinstall file with additional required directives.
- `autoinst-compute_node_hostname` - The autoinstall template with hard-coded node-specific information.
- `pcmlinux_template` - The PXELINUX boot parameter file template for this image group.
- `pcmlinux-compute_node_hostname` - The PXELINUX boot parameter file for a specific node.

The autoinstall process boots the selected compute nodes over the network and initiates a typical RHEL Kickstart or SLES AutoYaST installation. During the operation, the autoinstall log displays on the terminal.

**9.** Proceed to the following:

[Completing the autoinstall](#) on page 36

## Autoinstalling with the CLI

### Registering an autoinstall image group using the CLI

You can use one of the following command-based methods to add an autoinstall image group.

- Method 1:

1. Log into the admin node as the root user.
2. Enter the following command to access the `cmu` prompt:

```
# cmucli
```

3. Use the `add_ai_image_group` command in the following format to add the node:

```
add_ai_image_group image_group_name "repository_path"  
"path_to_autoinstall_file"
```

For example:

```
cmu> add_ai_image_group rh7u4_autoinst "/data/repositories/rh7u4_x86_64" "/data/repositories/rh7_x86_64.cfg"  
repository registration tool  
registration in progress...  
--> creating image directory  
--> copying config file  
--> creating link to repository in CMU image directory  
--> exporting CMU image directory via NFS  
--> registering the cmu image in cmu.conf  
  
==> registration finished  
***  
*** add nodes to this group  
*** before using cmu_autoinstall_node...  
*** press enter to exit
```

- Method 2:

1. Log into the admin node as the root user.
2. Enter the `crepo --show` command to determine the path to the operating system repository.

Use the directory path in the output from this command as input to the `-r` parameter of the `cmu_add_image_group` command.

For example:

```
# crepo --show
* Cluster-Manager-1.0.0-rhel74 : /opt/clmgr/repos/cm/Cluster-Manager-1.0.0-rhel74
* Red-Hat-Enterprise-Linux-7.4 : /opt/clmgr/repos/distro/rhel7.4
  SUSE-Linux-Enterprise-Server-12-SP3 : /opt/clmgr/repos/distro/sles12sp3
```

### 3. (Conditional) Add the operating system image to the admin node.

Complete this step if the operating system ISO you want to use does not appear in the `crepo` output.

Use the `crepo` command to add the ISO.

For example:

```
# crepo --add SLE-12-SP3-Server-DVD-x86_64-GM-DVD1.iso
Mounting ISO file loopback...
  Running: cp -a /tmp/ZSwuu1XTAO /opt/clmgr/repos/distro/sles12sp3
Exporting repository for use with yume....
  Exporting /opt/clmgr/repos/distro/sles12sp3 through httpd, http://node64/repo/opt/clmgr/repos/distro/sles12sp3
Updating default rpm lists...
  Updating: /opt/clmgr/image/rpmlists/generated/generated-rhel7.4.rpmlist
  Updating: /opt/clmgr/image/rpmlists/generated/generated-ice-rhel7.4.rpmlist
  Updating: /opt/clmgr/image/rpmlists/generated/generated-lead-rhel7.4.rpmlist
  Updating: /opt/clmgr/image/rpmlists/generated/generated-admin-rhel7.4.rpmlist
```

### 4. Use the `cmu_add_image_group` command at the admin node prompt.

```
# /opt/clmgr/bin/cmu_add_image_group -n rh7u4_autoinst -d autoinstall \
-r /data/repositories/rh7u4_x86_64 -t /data/repositories/rh7_x86_64.cfg
```

## Adding nodes to an autoinstall image group using the CLI

You can use one of the following command-based methods to add nodes to an autoinstall image group:

- Method 1:

1. Log into the admin node as the root user.
2. Enter the following command to access the `cmu` prompt:

```
# cmucli
```

3. Use the `add_to_image_group` command in the following format to add the node:

```
add_to_image_group node to image_group_name
```

For example:

```
cmu> add_to_image_group node1 to rh7u4_autoinst
selected nodes: node1
processing 1 node ...
cmu>
```

- Method 2:

1. Log into the admin node as the root user.
2. Use the `cmu_add_to_image_group_candidates` command at the admin node prompt.

For example:

```
# /opt/clmgr/bin/cmu_add_to_image_group_candidates -t rh7u4_autoinstall n1 n2
processing 2 nodes...
```

## Autoinstalling compute nodes using the CLI

You can use one of the following command-based methods to autoinstall compute nodes:

- Method 1:

1. Log into the admin node as the root user.
2. Enter the following command to access the `cmu` prompt:

```
# cmucli
```

3. Use the `autoinstall` command in the following format to add the node:

```
autoinstall "image" node
```

For example:

```
cmu> autoinstall "rh7u4_autoinst" node1
```

4. Proceed to the following:

[Completing the autoinstall](#) on page 36

- Method 2:

1. Log into the admin node as the root user.
2. Use a text editor to create a file called `nodes.txt`.

Within the file, include one node name per line.

3. Use the `cmu_autoinstall_node` command at the admin node prompt.

For example:

```
# /opt/clmgr/bin/cmu_autoinstall_node -l rh7u4_autoinst -f nodes.txt
```

4. Proceed to the following:

[Completing the autoinstall](#) on page 36

## Completing the autoinstall

When the autoinstall process is complete, the node or nodes reboot into the installed operating system.

## Procedure

1. Log into the admin node, and enter the following command to install root ssh keys on the node:

```
# /opt/clmgr/tools/copy_ssh_keys.exp node_name
```

For *node\_name*, specify the hostname of the node that received the autoinstalled image.

This step configures passwordless access for the root user on the autoinstalled node.

2. Respond to the system prompts for the root password.

3. Proceed to the following to install the cluster manager packages on the node:

[Managing software images](#) on page 202

## Troubleshooting and special cases

The following topics contain information you might need for specific operating systems and special environments:

- [Autoinstalling RHEL on nodes configured with HPE Dynamic Smart Array RAID \(B120i, B320i, B140\)](#) on page 37
- [Autoinstalling SLES on nodes configured with HPE Dynamic Smart Array RAID \(B120i, B320i, B140i\)](#) on page 39

## Autoinstalling RHEL on nodes configured with HPE Dynamic Smart Array RAID (B120i, B320i, B140)

To autoinstall RHEL 7 or RHEL 6 on nodes with HPE Dynamic Smart Array RAID configured, pass the *hpvsda* or *hpdsda* driver update diskette image to the kickstart environment. As a result, you can distribute the updated image to all new autoinstall image groups subsequently created.

To apply the customizations in this topic to all autoinstall image groups that you will ever create on this cluster, complete the procedure in this topic before you create an autoinstall image group. The customizations in this topic then apply to all the new autoinstall image groups that you create subsequently.

To apply the customizations in this topic to only a specific autoinstall image group, create the autoinstall image group first. Then, use the procedure in this topic to customize the group-specific template files in the following directory:

```
/opt/clmgr/image/group_name
```

## Procedure

1. Download the appropriate driver diskette image for the corresponding RHEL operating system version from the [Hewlett Packard Enterprise Support Center](#).

The *hpvsda* driver update diskette is required for B120i and B320i controllers. For example, for RHEL 7.4, download *hpvsda-version.rhel7u4.x86\_64.dd.gz*.

The *hpdsda* driver update diskette is required for the B140i controller. For example, for RHEL 7.4, download *hpdsda-version.rhel7u4.x86\_64.dd.gz*.

2. Expand (*gunzip*) the driver diskette image, and rename the extracted driver diskette image (.dd) with a .iso extension.

For example, *hpdsda-version.rhel7u4.x86\_64.iso*.

3. Copy the renamed file to the autoinstall repository directory that contains the RHEL DVD ISO contents.

This directory is automatically NFS exported by the autoinstall process.

4. Add a `driverdisk` line at the beginning of the RHEL autoinstall template file.

This `driverdisk` line must point to the expanded driver update diskette `.iso` file which was prepared in the previous step.

For example:

```
driverdisk --source=nfs:CMU_CN_MGT_IP:CMU_REPOSITORY_PATH/  
hpvs-a-1.2.12-110.rhel7u4.x86_64.iso
```

---

**NOTE:** `CMU_CN_MGT_IP` and `CMU_REPOSITORY_PATH` are automatically substituted with correct values during autoinstall. Optionally, these values can be hardcoded in the autoinstall template file.

If the driver update diskette is only required for a specific autoinstall image group, create the autoinstall image group first. Then, add the `driverdisk` line to that group-specific template file in the following directory:

```
/opt/clmgr/image/group_name/autoinst tmpl-orig.
```

5. (Conditional) Blacklist the `ahci` driver.

Complete this step on B120i- and B140i-controller-based nodes only. Do not complete this step on B320i-based nodes.

For nodes with B120i- or B140i-based Dynamic Smart Array RAIDs, the `ahci` driver conflicts with the `hpvs-a` and `hpdsa` drivers. To avoid this conflict, modify the `CMU_KS_KERNEL_PARMS` line in `/opt/clmgr/etc/cmuserver.conf` to explicitly blacklist the `ahci` driver during OS autoinstall on the target nodes. The blacklisting command may vary across different OS versions. For the OS version-specific kernel command-line parameters for blacklisting the `ahci` driver, see the following table:

---

OS	Kernel boot parameter for blacklisting <code>ahci</code>
RHEL 7.X	<code>modprobe.blacklist=ahci</code>
RHEL 6.X	<code>blacklist=ahci</code>

---

For example:

```
CMU_KS_KERNEL_PARMS="lang=CMU_CN_OS_LANG devfs=nomount ramdisk_size=10240  
console=CMU_CN_SERIAL_PORT ksdevice=CMU_CN_MAC_COLON initrd=autoinst-  
initrd-CMU_IMAGE_NAME blacklist=ahci"
```

If a node is enabled with B120i- or B140-based Dynamic Smart Array RAID mode, verify that the `hpvs-a` or the `hpdsa` driver diskette is inserted.

---

**NOTE:** If the blacklisting step is only required for a specific autoinstall image group, create that group using the GUI or CLI. Then, add nodes to that group. Launch the autoinstall operation on one node. After a few minutes, stop the process. This action creates the boot templates under the `/opt/clmgr/image/group_name/` directory. Edit the `pcmlinux_template` file to add the appropriate `ahci` blacklisting kernel parameter.

6. If a server contains extra Smart Array RAID disks (for example, P420i) in addition to the B120i-, B320i-, or B140i-driven Dynamic Smart Array RAID (for example, SL4540 has a B120i and one or more P420i controllers), then autoinstall may not work as expected due to those extra disks.

To avoid this situation, also blacklist the `hpsa` driver by editing `CMU_KS_KERNEL_PARMS` in `/opt/clmgr/etc/cmuserver.conf` as described in a previous step to ensure that the OS or bootloader is always installed on the Dynamic Smart Array disks.

For RHEL 7, the parameter is `modprobe.blacklist=hpsa`.

For RHEL 6, the parameter is `blacklist=hpsa`.

Also add lines to the `%post` section of the RHEL autoinstall template to ensure that the extra Smart Array disks are re-enabled after the OS installation.

- For RHEL 7 node templates, add the following:

- `sed -i 's/modprobe.blacklist=hpsa//g' /boot/grub2/grub.cfg`
- `sed -i 's/modprobe.blacklist=hpsa//g' /boot/efi/EFI/redhat/grub.cfg`
- `sed -i 's/blacklist hpsa//g' /etc/modprobe.d/anaconda-blacklist.conf`

- For RHEL 6 node templates, add the following:

- `sed -i 's/blacklist=hpsa//g' /boot/grub/grub.conf`
- `sed -i 's/blacklist=hpsa//g' /boot/efi/EFI/redhat/grub.conf`
- `sed -i 's/blacklist hpsa//g' /etc/modprobe.d/anaconda.conf`

## Autoinstalling SLES on nodes configured with HPE Dynamic Smart Array RAID (B120i, B320i, B140i)

To autoinstall SLES 12 or SLES 11 on nodes with HPE Dynamic Smart Array RAID configured, pass the `hpvsda` or `hpdsda` driver update diskette image to the AutoYaST environment. As a result, you can distribute the updated image to all new autoinstall image groups subsequently created.

To apply the customizations in this topic to all autoinstall image groups that you will ever create on this cluster, complete the procedure in this topic before you create an autoinstall image group. The customizations in this topic then apply to all the new autoinstall image groups that you create subsequently.

To apply the customizations in this topic to only a specific autoinstall image group, create the autoinstall image group first. Then, use the procedure in this topic to customize the group-specific template files in the following directory:

`/opt/clmgr/image/group_name`

### Procedure

1. Download the appropriate driver diskette image for the corresponding SLES OS version from the [\*\*Hewlett Packard Enterprise Support Center\*\*](#) website.

- The `hpvsda` driver update diskette is required for B120i and B320i controllers. For example, for SLES 11 SP4, download `hpvsda-version.sles11sp4.x86_64.dd.gz`.
- The `hpdsda` driver update diskette is required for the B140i controller. For example, for SLES 11 SP4, download `hpdsda-version.sles11sp4.x86_64.dd.gz`.

**2.** Expand (`gunzip`) the driver diskette image, and copy the `.dd` image file to the autoinstall repository directory.

The autoinstall directory is the directory that contains the SLES DVD ISO contents. This directory is automatically NFS exported by the autoinstall process.

**3.** Open the following file in a text editor:

```
/opt/clmgr/etc/cmuserver.conf
```

**4.** Modify `CMU_AY_KERNEL_PARMS` to append a Driver Update Disk `dud` parameter, which points to the `.dd` image file.

For example:

```
CMU_AY_KERNEL_PARMS="autoyast=nfs://CMU_CN_MGT_IP/opt/clmgr/image/CMU_IMAGE_NAME/autoinst-CMU_CN_HOSTNAME
install=nfs://CMU_CN_MGT_IP//CMU_REPOSITORY_PATH initrd=autoinst-initrd-CMU_IMAGE_NAMe netwait=20
dud=nfs://CMU_CN_MGT_IP//CMU_REPOSITORY_PATH/hpvsda-1.2.0-185.sles12sp3.x86_64.dd"
```

---

**NOTE:** `CMU_CN_MGT_IP` and `CMU_REPOSITORY_PATH` are automatically substituted with the appropriate values during the autoinstall process. These values can also be hardcoded in the template file.

If the Driver Update Disk `dud` is only required for a specific autoinstall image group, create that group using the GUI or CLI and add nodes to the group. Launch the autoinstall operation on one node, and after a few minutes, stop the process. This action creates the boot templates under the `/opt/clmgr/image/group_name` directory.

Edit the `pcmlinux_template` file to add the appropriate `ahci` and `hpsa` blacklisting kernel parameters.

**5.** (Conditional) Append `broken_modules=ahci` to `CMU_AY_KERNEL_PARMS`.

Complete this step for B120i- or B140i-based Dynamic Smart Array RAIDs.

For example:

```
CMU_AY_KERNEL_PARMS="autoyast=nfs://CMU_CN_MGT_IP/opt/clmgr/image/CMU_IMAGE_NAME/autoinst-CMU_CN_HOSTNAME
install=nfs://CMU_CN_MGT_IP//CMU_REPOSITORY_PATH initrd=autoinst-initrd-CMU_IMAGE_NAME netwait=20
dud=nfs://CMU_CN_MGT_IP//CMU_REPOSITORY_PATH/hpvsda-1.2.0-185.sles12sp3.x86_64.dd broken_modules=ahci,hpsa"
```

If a node is enabled with B120i- or B140i-based Dynamic Smart Array RAID mode, verify that the `hpvsda` or `hpdsda` driver diskette is inserted.

This step is not required for B320i-based nodes.

**6.** Blacklist the `hpsa` driver.

Complete this step on nodes that have additional Dynamic Smart Array RAID disks, such as P420i, in addition to the B1200i-, B320i-, or B140i-driven Dynamic Smart Array RAID. This action helps to avoid potential disk selection conflicts.

For example:

```
CMU_AY_KERNEL_PARMS="autoyast=nfs://CMU_CN_MGT_IP/opt/clmgr/image/CMU_IMAGE_NAME/autoinst-CMU_CN_HOSTNAME  
install=nfs://CMU_CN_MGT_IP//CMU_REPOSITORY_PATH initrd=autoinst-initrd-CMU_IMAGE_NAME netwait=20  
dud=nfs://CMU_CN_MGT_IP//CMU_REPOSITORY_PATH/hpsa-1.2.0-185.sles12sp3.x86_64.dd broken_modules=ahci,hpsa"
```

7. Save and close the following file:

```
/opt/clmgr/etc/cmuserver.conf
```

---

**NOTE:** If the blacklisting step is only required for a specific autoinstall image group, edit the `pcmlinu`\_template boot templates under the `/opt/clmgr/image/group_name` directory. Add the appropriate `ahci` and `hpsa` blacklisting kernel parameters.

---

## Capturing an image from a leader node or flat compute node using the GUI

The capture image operation captures the entire operating system from a flat compute node or leader node. This action stores the image in an image archive on the admin node. You can deploy this image to other nodes of the cluster. Each captured image is associated with an image group.

Prior to capturing an image, ensure that the node from which you want to capture an image contains all of the preferred services, such as NTP for synchronizing time across the cluster. Install any additional applications or libraries.

### Procedure

1. Click **Options > Enter Admin mode** and specify the cluster credentials.
2. Expand the node list in the left frame.
3. Select the node from which you want to capture an image.
4. Right-click the selected node and select **Capture Image**.
5. In the **Capture Image** popup, in the **Image Group Name** field, type in the name for the new image and click **OK**.

For example, if the image on the selected node is called `rhel7.4`, you could type `rhel7.4-new` into the **Image Group Name** field.

6. Proceed to the following:

[Provisioning an image](#) on page 41

## Provisioning an image

The provisioning (or deployment) operation copies a captured image from the admin node to a leader node or flat compute node. The provisioning operation destroys all information on the target node, so back up any information you want to retain. The new image is the same as the original, but the copy operation updates the following on the target node:

- The host name of the node.
- The IP address of the network used for deployment.
- The compute node default gateway. The copy operation updates the target node with the value of the Default gateway IP address field in the cluster database. This value can be one of the following:

- default
- cmumgt
- The IP address of the gateway.

For information about how to modify the gateway IP address of a node, see the following:

[\*\*Changing the default gateway IP address of a node from the GUI\*\*](#) on page 93

Proceed to the following to provision a node:

[\*\*Provisioning a leader node or a flat compute node using the GUI\*\*](#) on page 42

## Provisioning a leader node or a flat compute node using the GUI

### Prerequisites

Before provisioning, verify that the following prerequisites are met:

- A valid image group is created.
- An image associated with the image group exists.
- The nodes being deployed belong to the image group.
- The nodes being deployed belong to a network group.
- The image group has an image that is compatible with the node hardware.
- The nodes are ready to be powered on by the management card (also known as the baseboard management controller (BMC)) .

### Procedure

1. In the left pane, select the node, nodes, or node group that you want to host the captured image.
  2. Right-click the nodes, and select **Provision Image (Deploy)**.
  3. In the **Deploy Images** popup window, complete the following fields and click **OK**:
    - In the **Image group** field, select the new image group.
    - In the **Kernel** field, select the kernel that you want to deploy on the selected nodes.
    - In the **Rootfs type** field, select either **disk** or **tmpfs**.
  4. In the **Add to image group** popup window, click **Yes**.
  5. Examine the status window and check for any nodes that failed to deploy.
- When provisioning is complete, the final status displays. The compute nodes that provisioned successfully display in the chosen image group. The compute nodes that failed remain in the default image group.
6. Proceed to the following topic to review the success of the provisioning operation:

[\*\*Reviewing the success of the provisioning operation\*\*](#) on page 43

## Reviewing the success of the provisioning operation

Use the following conditions to determine if a deployment operation was successful.

- Successfully provisioned nodes are added to the image group containing the image. The remaining nodes are added to the default image group.
- If the node name has a suffix of [non-active] in the image group containing the image, then the node failed to provision. If the node name has a suffix of [active], then the node is provisioned correctly.
- The list of successfully provisioned nodes is available in the `/opt/clmgr/log/cmucerbere-pid.log` file.

# Monitoring a cluster

## Monitoring and cluster security

Monitoring consists of two conceptual parts:

- The back-end gathering and archiving of metrics and reactions to alerts
- The GUI metric and alert display

There are two methods of gathering metrics. The first method launches monitoring agents onto the compute nodes to gather in-band, operating system level metrics and aggregate them back to the admin node for storage and display. The second method supports invoking a program on the admin node. This program gathers metrics out-of-band, or outside of the running OS on the compute nodes, and feeds those metrics to the cluster manager for storage and display. An example of the out-of-band approach is the pre-configured support for gathering hardware metrics, such as power and temperature, from the iLO of each compute node.

For in-band metric gathering, launching the operating system agents requires that the root user on any node in the cluster be able to `ssh` to any other compute node, without a password, to (re)start the monitoring agent on those nodes. By default, the cluster manager copies the root `ssh` keys to all compute nodes to configure password-less `ssh` access one node to another.

You can configure a dedicated non-root user on the cluster for the sole purpose of (re)starting the monitoring agents. In `cmuserver.conf` this setting is `CMU_MONITORING_USER`. By default, this variable is set to `root`. If you change this setting to `cmu` and restart monitoring, the cluster manager configures a local `cmu` user account on all compute nodes and sets up `ssh` keys for this account so that the cluster manager can `ssh` between any two nodes without a password. The cluster manager then uses the account to start monitoring agents on the compute nodes.

---

**NOTE:** It is best to configure `CMU_MONITORING_USER` with a local non-root user account that does not already exist. The cluster manager creates this local account with its home directory in `/opt/clmgr/etc/` so that the default `ssh` key location is in the predetermined `/opt/clmgr/etc/.ssh/` directory on each node.

When the monitoring agents are configured to be started by a non-root `CMU_MONITORING_USER` user account, the configured metric action and alert commands must be executable by the non-root `CMU_MONITORING_USER` user account on the compute nodes. The default metrics and alerts are executable by a non-root user. If any metric action or alert command requires the root user, then action should be taken to resolve this issue. One solution is to configure passwordless `sudo` for the non-root `CMU_MONITORING_USER` on the compute nodes for the given command(s). Then, you can prefix those command(s) in the `ActionAndAlertsFile.txt` with `sudo`.

---

Alternatively, you can disable copying of the root `ssh` keys to all nodes. In `cmuserver.conf`, this setting is `CMU_DISTRIBUTE_ROOT_PRIVATE_SSH_KEY`. The default setting for this is `yes`. When this is set to `no`, the cluster manager configures the root account on the admin node with password-less access to all nodes. However, the root user on the compute nodes does not have password-less `ssh` access to other nodes because the private `ssh` key for the root user is not distributed by the cluster manager.

---

**NOTE:** When `CMU_DISTRIBUTE_ROOT_PRIVATE_SSH_KEY` is set to `no`, the cluster manager does not actively remove any `ssh` keys from an existing cluster. You can choose to do this manually or create and deploy new operating system images that do not contain the root private `ssh` key.

The purpose of these features is to improve security on the cluster by offering improved protection of the root ssh keys.

## (Conditional) Installing the monitoring client

The default cluster manager installation process installs the monitoring client. Use the procedure in this topic if you need to reinstall the monitoring client. For example, if you autoinstall some nodes, you might need to install the monitoring client on those nodes.

### Procedure

1. Click **Options > Enter Admin mode** to enable administrator mode.
2. In the left pane, locate the node upon which you want to install the monitoring client.  
Expand the node list as needed.
3. Right-click the node, and select **Refresh > Install Monitoring Client**.  
A window displays the status of the RPM installation.
4. When the installation is complete, press Enter to close the status window.

## Monitoring the cluster

In the main display, the left pane lists resources, such as **System Groups**, **Network Groups**, and **Image Groups**. Click the + button to expand a resource.

For example, to see nodes belonging to an image group, click the + button to expand the **Image Group** resource list. Then, expand a selected image group. The following display shows image groups.

The screenshot shows the Cluster Administration interface with the 'Monitoring' tab selected. The left pane displays a tree view of resources under 'All Resources'. Under 'Image Groups', there are three groups listed: 'ice-rhel7.4', 'lead-rhel7.4', and 'rhel7.4'. The 'rhel7.4' group is expanded, showing '6 non-active candidates' (n0-n5). The right pane shows a table titled 'Groups List' with columns: State, Name, Nodes, Creation time, Deletion time, Duration, and Show. The table contains three rows corresponding to the groups listed in the left pane.

State	Name	Nodes	Creation time	Deletion time	Duration	Show
Green	ice-rhel7.4	0	2018-05-22 - 15:31:23			Green arrow
Green	lead-rhel7.4	0	2018-05-22 - 15:17:21			Green arrow
Yellow	rhel7.4	6	2018-05-22 - 14:59:22			Yellow arrow

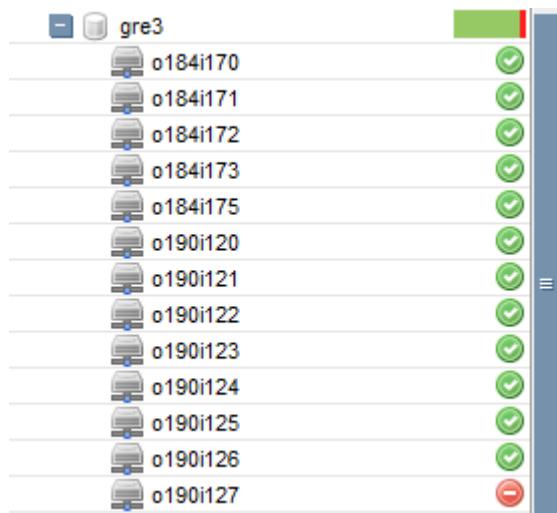
You can view compute nodes by network group, image group, custom group, or nodes definition.

When viewing by image group, nodes that are deployed successfully are listed in the active image group in which they are deployed. The GUI lists inactive nodes as non-active candidates.

If no network group is defined, or if a node is not included in a network group, the nodes appear in a default network group that contains the unclassified nodes.

## Node and group status

In left pane, when you expand a group, you expose the status indicators for the group as a whole and for the objects in the group. For example:



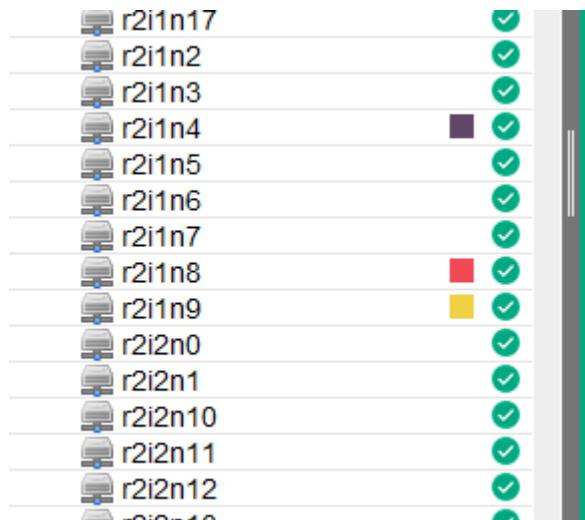
**Figure 4: Node status**

The status bar above the group represents the proportion of nodes in various states. In the preceding figure, the red/green status bar at the top displays the overall status for the group.

The following table describes the available statuses in the left-most part of the display.

Status	Description
✓	The object is in OK status. This means that node is registered to the cluster manager as booted.
⚠	Appears when the monitoring process detects a problem with the object and wants to alert you. Right-click the object in the tree to view the details of the alert. This icon can appear near a red or green circle icon.
✗	This means that the object is not registered to the cluster manager as booted. User action is required to identify the problem.
?	The object status is unknown. The object daemon is not monitored because it has failed or is late. This status changes when the monitoring software selects a new monitoring server for the object . No user action is required. For very large clusters of 2000 nodes or more, the back-end process that gathers the status information might take a long time to complete. When this occurs, a <code>pingStaleDelay</code> timeout is reached in the GUI, which causes all of the objects to display in an unknown status. By default, this timeout is 10 seconds. You can increase this timeout. For information about how to increase the timeout, see the following: <a href="#">Changing the pingStaleDelay timeout value</a> on page 47

When monitoring is enabled, additional colored boxes might appear in the display. These boxes show the colors used to represent the objects in the monitoring display. For example:



## Changing the pingStaleDelay timeout value

### Procedure

1. Open the following file:

```
/opt/clmgr/etc/cmuserver.conf
```

2. Add `-Dcmu.monitoring.pingStaleDelay=value` to the `CMU_JAVA_SERVER_ARGS` variable.

For `value`, specify a new delay value. For example, specify 30 to increase the timeout to 30 seconds, as follows:

```
-Dcmu.monitoring.pingStaleDelay=30
```

3. Enter the following commands to stop and start the `cmu` service:

```
# systemctl stop cmu  
# systemctl start cmu
```

## Right pane display

Information in the right pane appears according to the object selected in the left pane. When you select the cluster identifier at the top of the left pane, the right pane displays the global cluster view. When you select a different object or group of objects, the right pane displays information that pertains to that object or group.

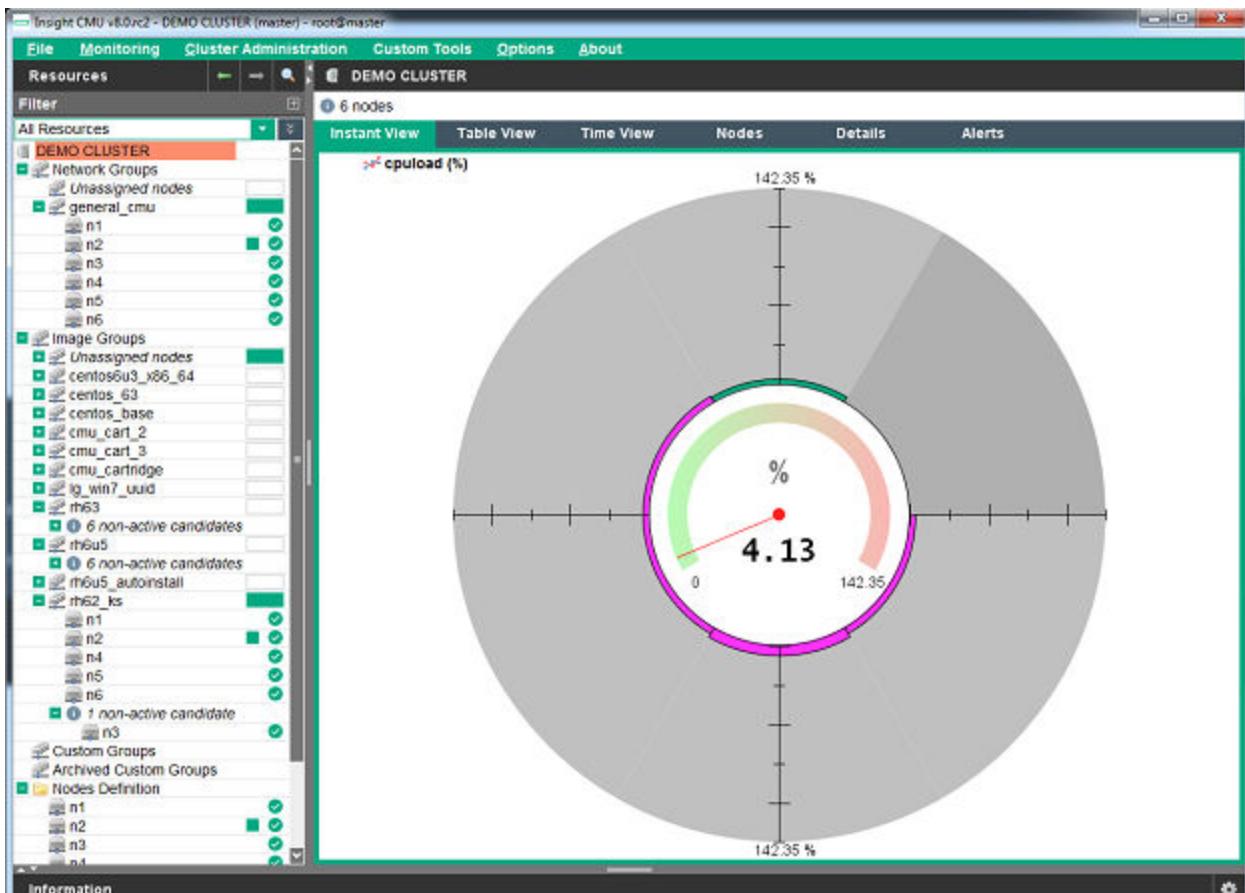
The following tabs are available in the right pane:

- **Instant View**
- **Bar Graph View**
- **Table View**
- **Time View**
- **Nodes**

- Details
- Alerts

## Global cluster view

The pie graphs in the global cluster view represent the cluster monitoring sensor value. To select the sensors being monitored, right-click an item in the right pane. A metric window displays. Select a metric and click **OK**.



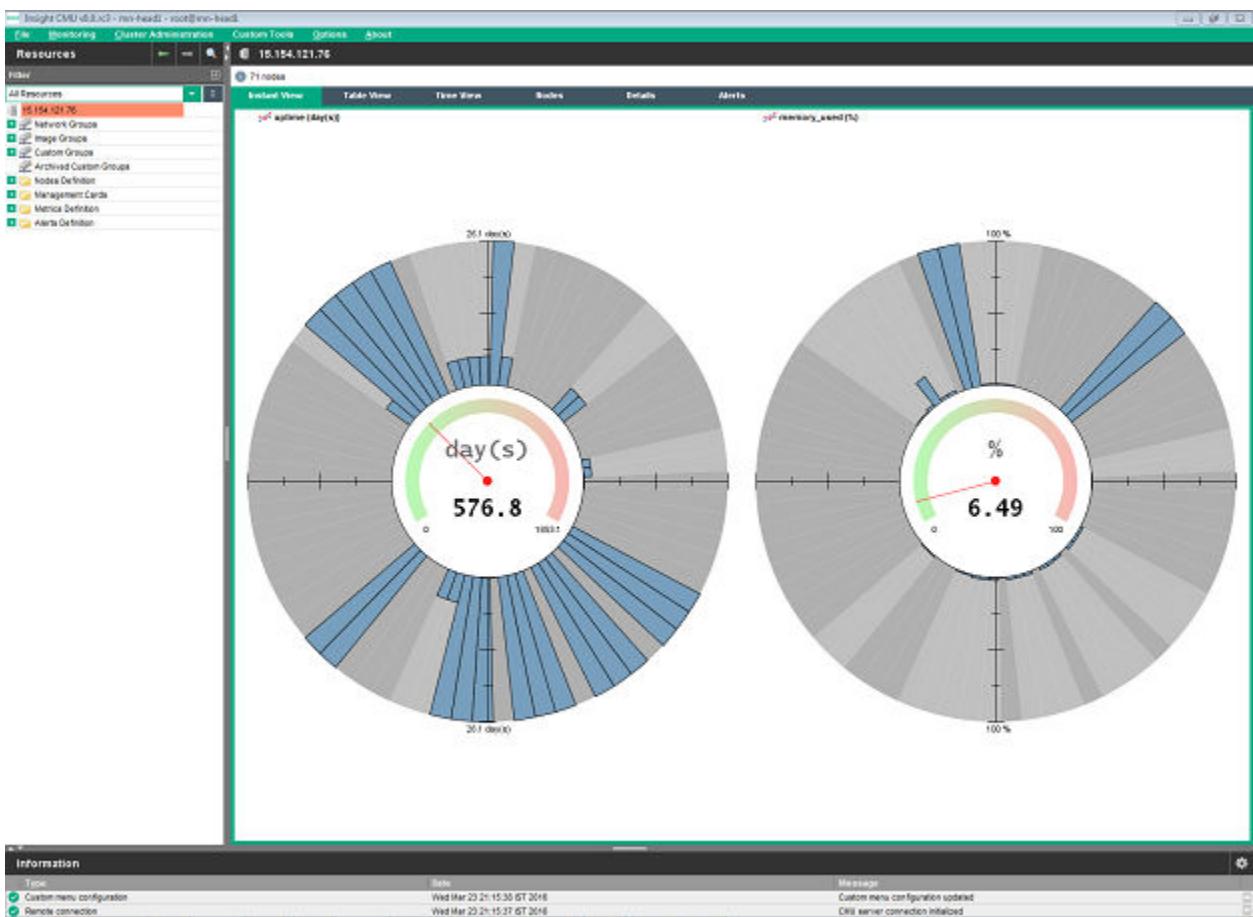
**Figure 5: Monitoring window**

Hover the mouse over a portion of the pie to display the name of the corresponding node, the status, and the value of the displayed sensor.

For a given metric, the internal circle of a pie graph represents zero. The external circle represents the maximum value. By default, the current value of the metric displays in blue. To change the default color, click **Options > Properties** in the top bar and select the monitor options. To change the color for a specific petal, click the petal. A gray-colored pie graph means that there is no activity on the node or that a metric is not correctly updated.

## Gauge widget

The gauge widget consists of the pie graphs in the right pane of the following figure:



**Figure 6: Memory used summary**

By default, the pie graphs display the sum value, or the average value, of the sensors. In the figure, the pie graph on the left shows the sum of all the days that all the nodes have been up. The pie graph on the right shows the average CPU load on each node.

To configure the gauge widget, take one of the following actions:

- Edit the following file:  
`/opt/clmgr/etc/ActionAndAlertsFile.txt`
- Click **Custom Tools > Edit > Monitoring Configuration** and update the information.

## Resource view

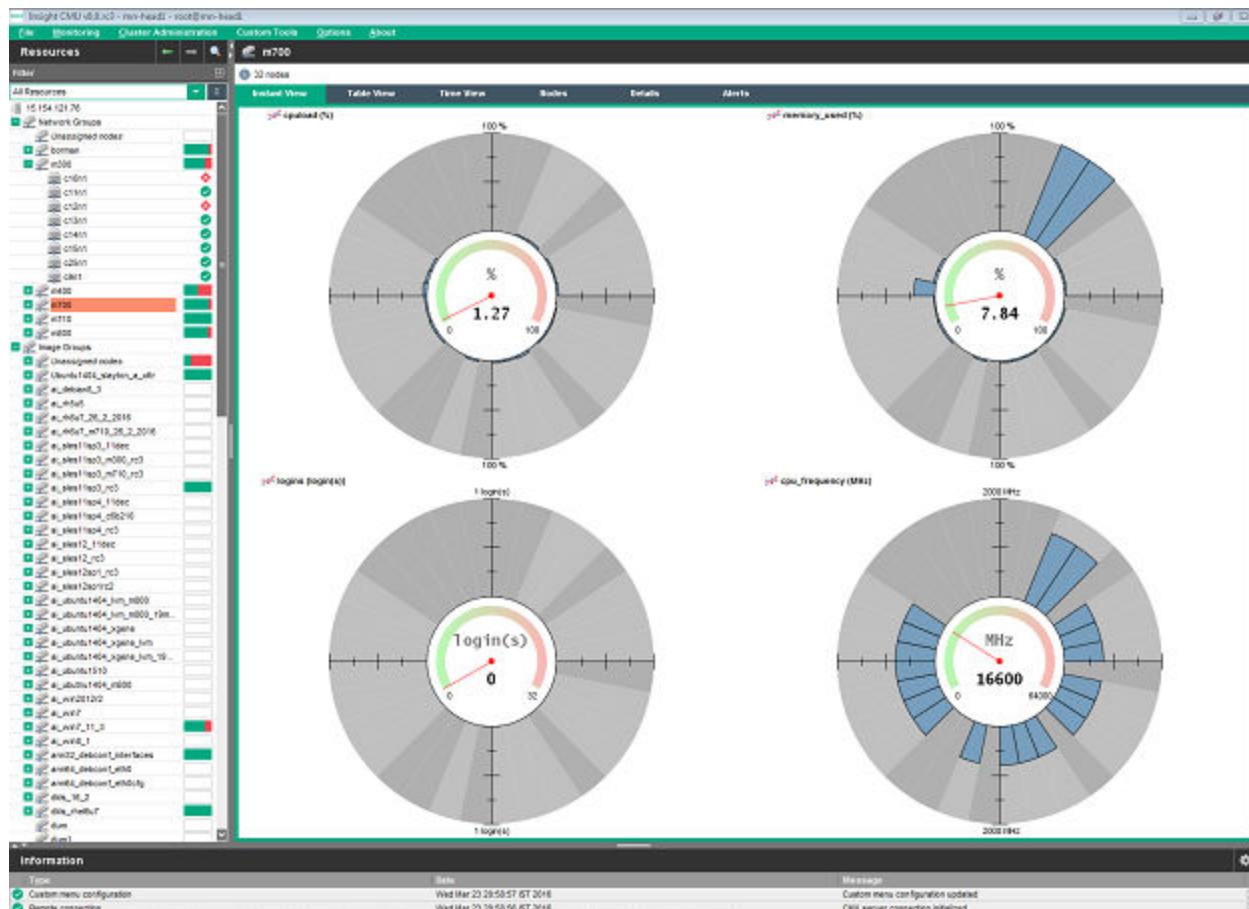
Click the resource in the left-frame tree to display the resource in the central frame.

Monitoring values can be displayed by:

- Global cluster
- A specific image group
- A specific network group
- A specific custom group

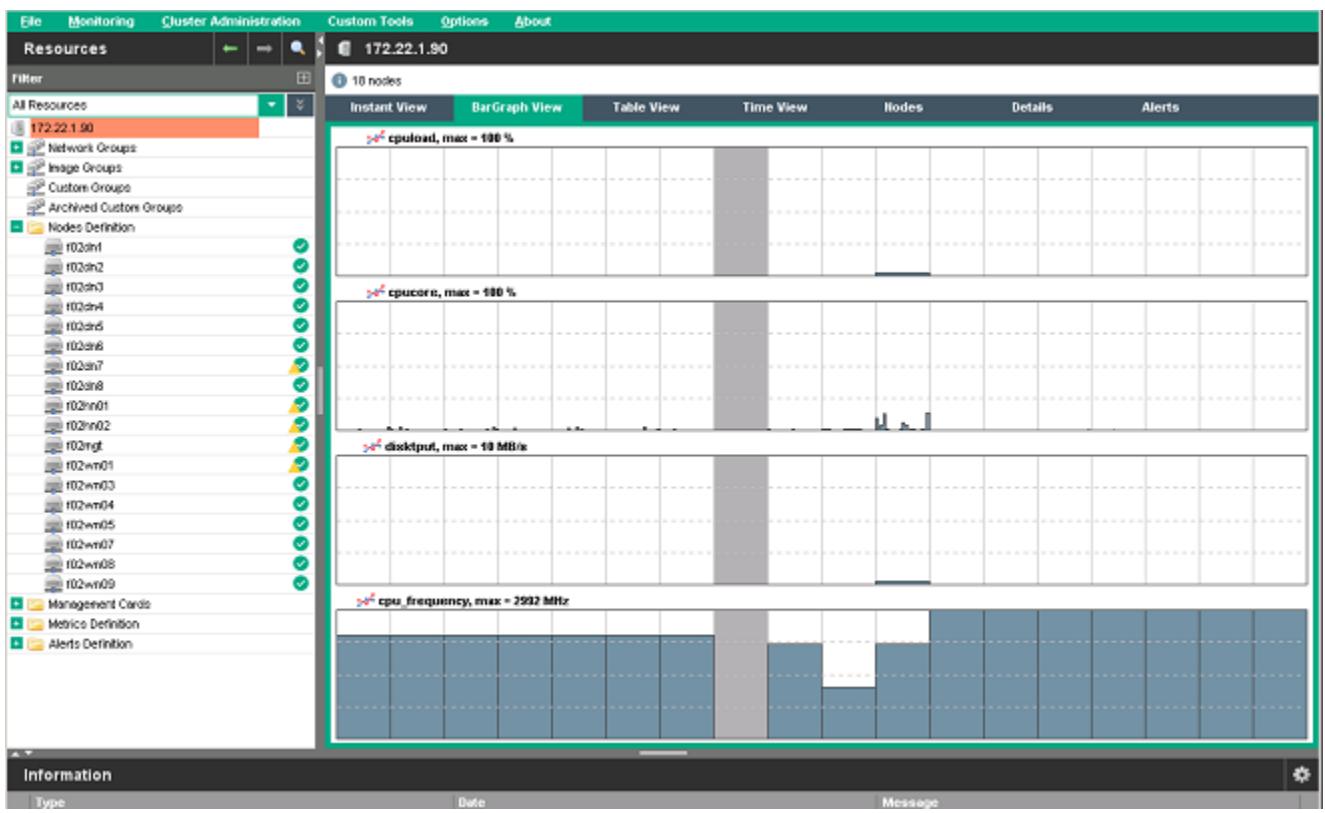
**NOTE:** You can display resource or node-specific monitoring metrics and alerts at the command line by using the /opt/clmgr/bin/cmu\_monstat command. For more information, see the cmu\_monstat manpage.

To see pie graphs that represent the monitored values, click the **Instant View** tab. To change the pie graph, right-click the central frame, select the metrics, and click **OK**.



**Figure 7: Resource view overview - instant view**

To see bar graphs that represent the monitored values, click the **BarGraph View** tab. To change the bar graphs, right-click the central frame, select the metrics, and click **OK**.



**Figure 8: Resource view - BarGraph view**

To view alerts raised for nodes in this group, select the **Alerts** tab in the central frame.

49 nodes					
	Instant View	Table View	Time View	Nodes	Details
					Alerts
	Date	Node	Message	Value	Unit
⚠ Mar 10, 2014		n04	Someone is connected	1	login(s)
⚠ Mar 10, 2014		n03	Someone is connected	1	login(s)
⚠ Mar 10, 2014		n02	Someone is connected	1	login(s)
⚠ Mar 10, 2014		n01	Someone is connected	1	login(s)

**Figure 9: Alert messages**

**NOTE:** You can define reactions to alerts in the /opt/clmgr/etc/ActionAndAlertsFile.txt file. For more information, see the following:

#### Customizing monitoring, alerting, and reactions on page 60

#### Detail mode

To display a table with sensor values, select the **Table View** tab in the central frame. Each table cell displays a status color representing the percentage of maximum value used. The following status colors are available.

- **Green** = The value is below 33% of the maximum value.
- **Orange** = The value is between 33% and 66% of the maximum value.
- **Red** = The value is above 66% of the maximum value.

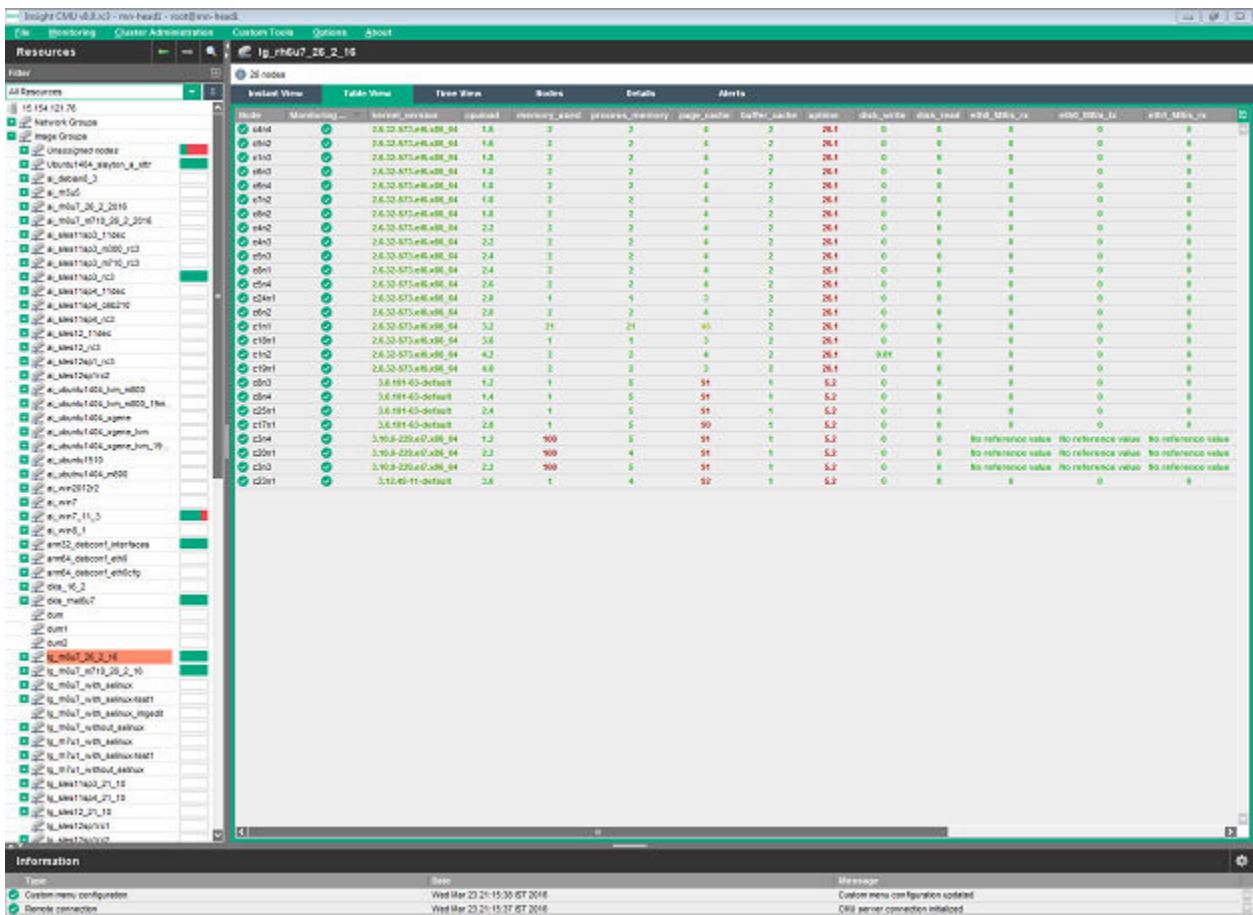
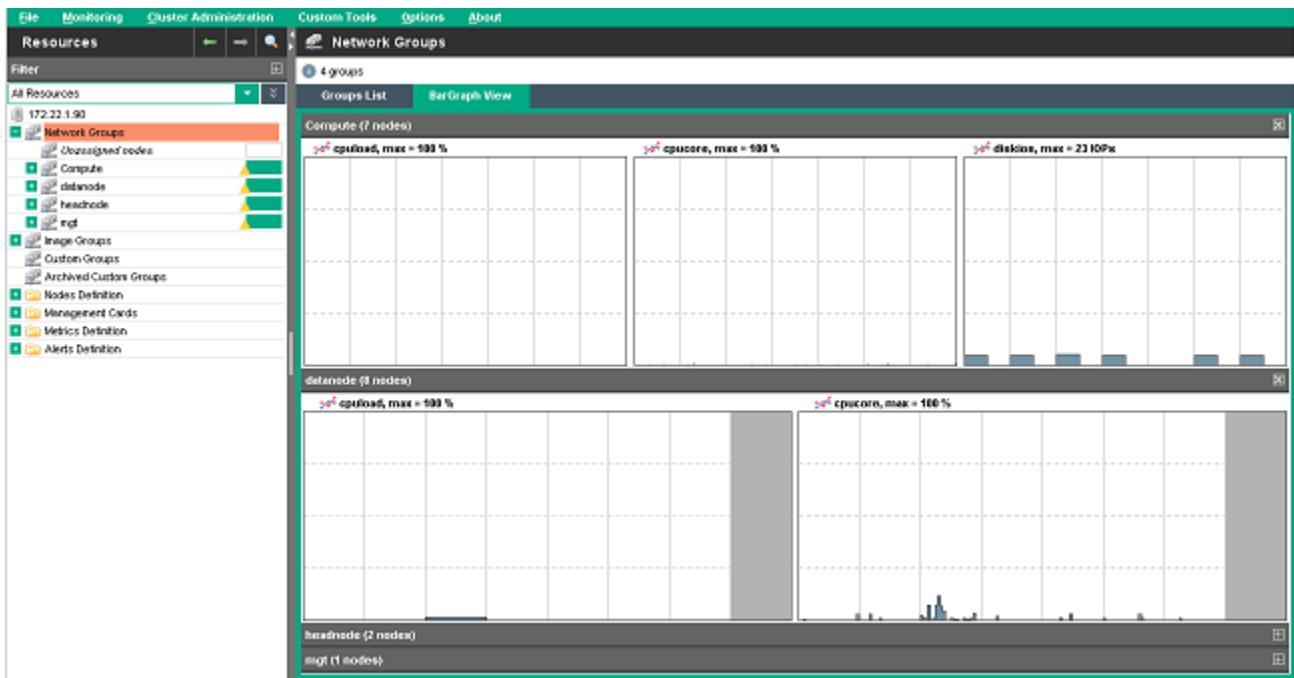


Figure 10: Table View details

### Group mode with bar graph view

Select a group in the left panel and click the **BarGraph View** tab to display multiple bar graphs. Each bar graph view contains the node associated with the subgroups.



**Figure 11: Group mode with BarGraph view**

In the image above, the right panel displays the subgroups under the **Network Groups** category. In the right pane, the top section displays the **Compute** subgroup, which is defined to display 3 metrics. In the right pane, the bottom section displays the **datanode** subgroup, which is defined to display 2 metrics. To minimize a section, click the **X** button in the upper right corner of the section. To expand a minimized section, click the **+** button in the right corner of the section.

## Node view

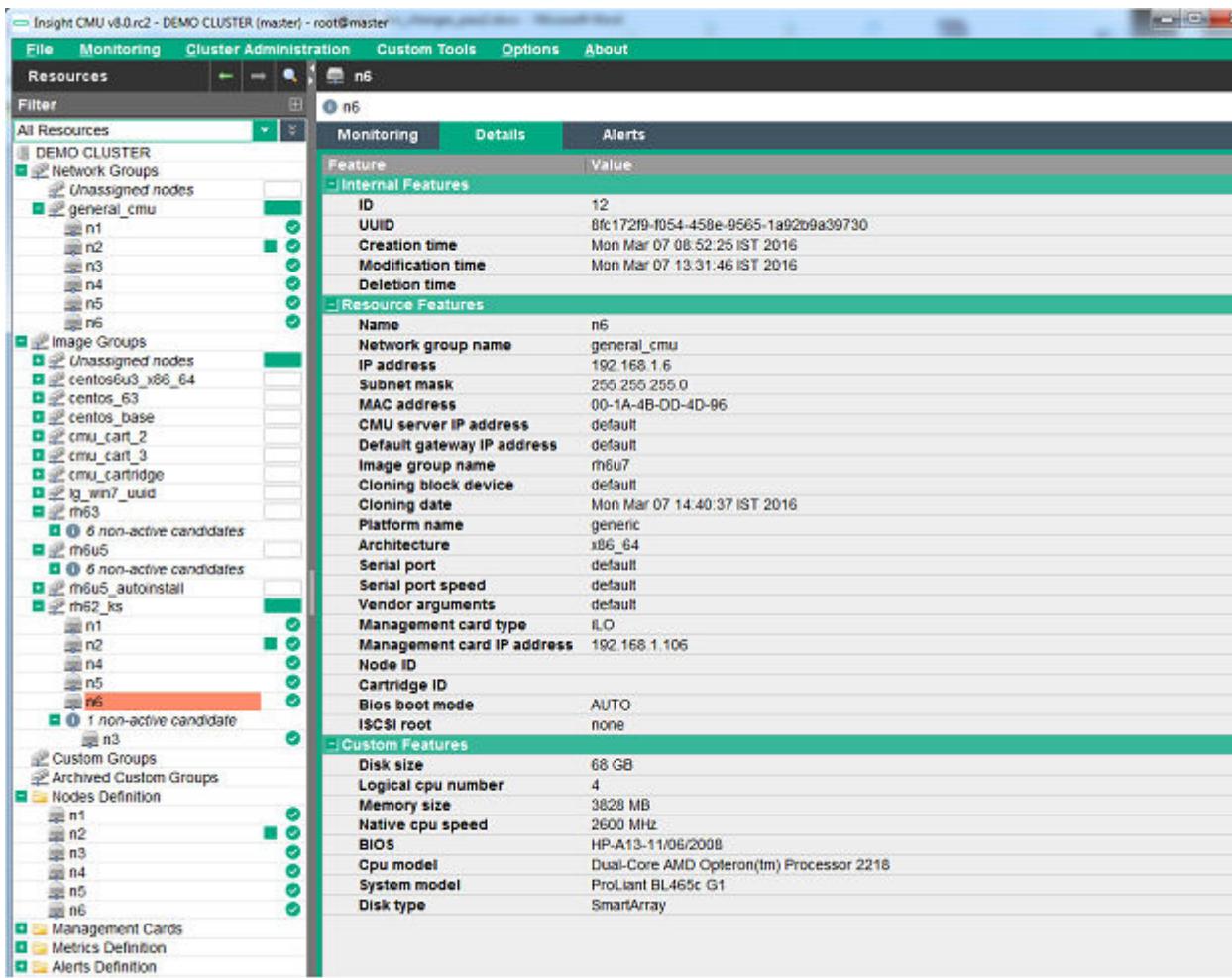
To display details for a specific node, select the node in the left-panel tree. This displays the Node view in the central frame.

The following tabs are available in Node view.

- **Monitoring** - Displays monitoring metric values for the node.
- **Details** - Displays static data for the node.

Some values are filled during the initial node discovery (scan node). For others, right-click the node in the left-panel tree and select **Update> Update Node Details**.

- **Alerts** - Contains the alerts currently raised for the specified node.



**Figure 12: Node details**

The title of the central frame displays the name of the node. The title is colored according to the state of the node.

The following tables display.

- Node Details table - Contains the static information from the cluster monitor.
- Information Retrieved table - Contains the current values of the sensors retrieved for the node.
- Alerts Raised table - Contains the alerts currently raised for the node.

## Time view

You can use the cluster manager to visualize cluster activity in time and in a scalable manner.

Assuming that the GUI client has enough memory and OpenGL capabilities, Time view extends the 2D flowers visualization to provide an evaluative 3D view of your cluster with the Z-axis representing the time. For system requirements, see the following:

### **Technical dependencies** on page 58.

Time view visualizes the last 24 intervals at the finest interval resolution and the previous 80 intervals at 6 times the interval resolution per ring. For the default interval time of 5 seconds, the last 2 minutes (24 x 5 = 120 seconds = 2 minutes) is displayed at 5 seconds resolution and the previous 40 minutes (80 x 6 x 5

= 2400 seconds = 40 minutes) is displayed at a 30 seconds (6 x 5) per ring resolution. Detailed values are still available for the compressed rings with the use of tooltip functionality.

The long-standing 2D flowers are still available in the **Instant View** panel.

### Tagging nodes

Nodes can be labeled with a color. This allows chosen nodes to be easily tracked through different views or partitions. This functionality is available from the **Instant View** and **Time View** tabs. Toggle through a predefined set of four colors by clicking a node. Colored nodes are shared between the Instant view and the Time view. This allows them to be efficiently located regardless of the chosen visualization.

### Adaptive stacking

Adaptive stacking is an efficient way to monitor your cluster over a long period of time. Adaptive stacking provides 210 intervals of data, without sacrificing the finest granularity provided by the monitoring engine. The first 24 rings using the monitoring interval progressively slide and consolidate into an intermediate ring, making room for new data. The intermediate ring is full when six rings are stacked in it. When this happens, the stacked rings slide and a new intermediate ring is created. The entire set of 210 intervals of history is displayed as 24 rings at the monitoring interval and 80 rings of 6 x interval time. Stacked rings are displayed darker than single rings to differentiate them. For the default monitoring time of 5 seconds, this represents 2 minutes (24 x 5) at the detail 5 seconds interval and 40 minutes at 30 seconds (6 x 5) intervals.

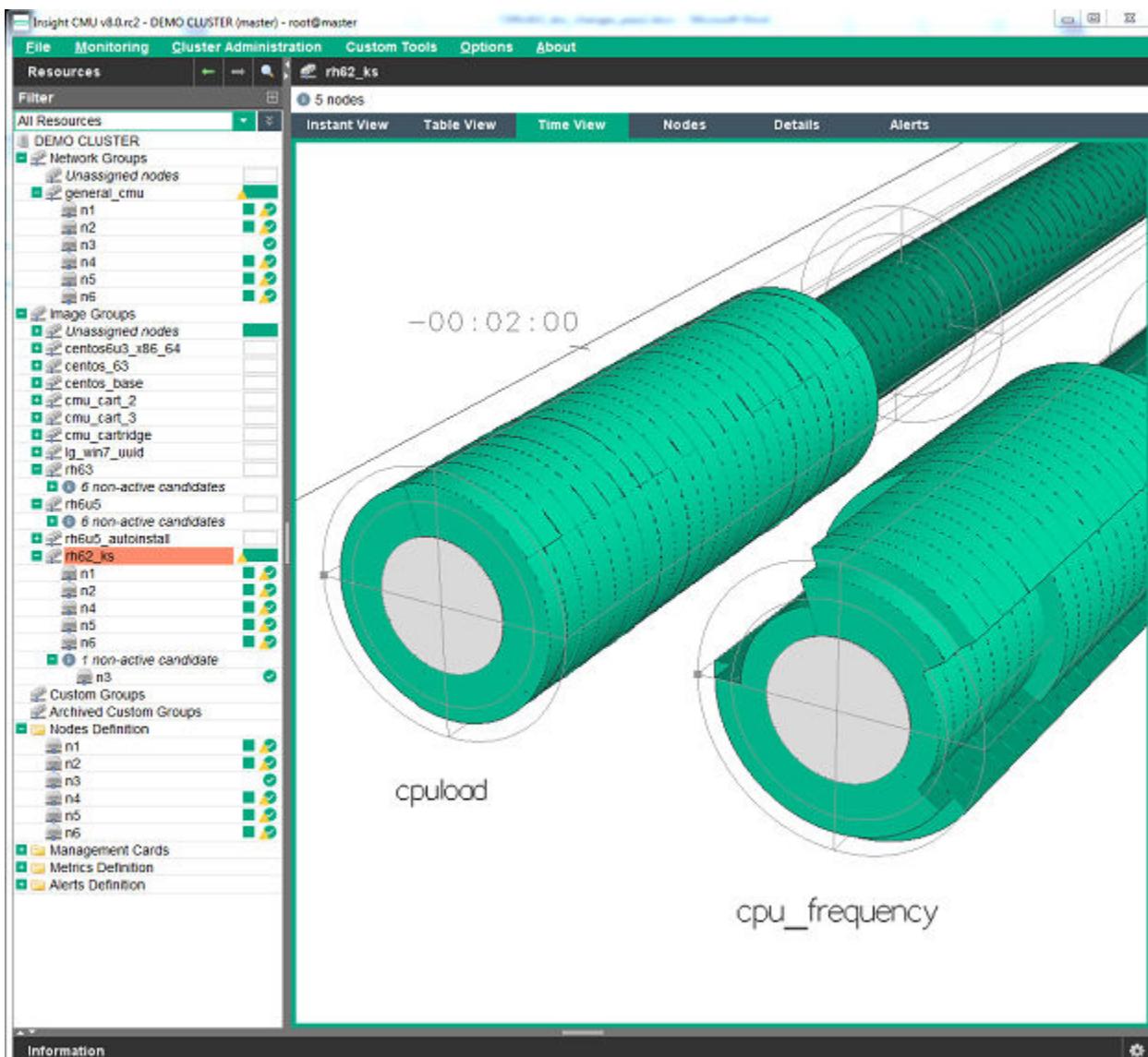


Figure 13: Time view

## Launching the Time View GUI

### Procedure

1. Select a group under one of the group categories.

That is, select one of the following:

- **Network Group**
- **Image Group**
- **Custom Group**

2. In the right panel, click the **Time View** tab.

Each selected metric is represented by a tube filled with rings. Each ring represents a snapshot of the metric value at a given time. A ring is composed of petals. Each petal represents a value for a given metric, at a given time, for a given node.

To interact with a node, right-click the node or hover over a 3D petal with your mouse to make a tool tip appear. The tool tip displays detailed values for the petal. Some functions on the **Time View** tab are inherited from 2D flowers. All node interaction is preserved from 2D to 3D.

## Bindings and options

### Mouse control

- Left-click a node - Marks the node from a set of four predefined colors
- Right-click a node - Opens the interactive menu for the node
- Right-click elsewhere - Opens the metrics selection menu

---

**NOTE:** Time view cannot display more than 10 metrics. For more information, see the following:

**Technical dependencies** on page 58.

---

Navigating within the 3D scene

- Left-click and drag - Translate the scene
- Right-click and drag - Rotate the scene
- Rotate the mousewheel - Rotate tubes on themselves
- Press the mousewheel and drag - Zoom

### Keyboard control

Keyboard shortcuts are available for some Time view options. The following shortcuts are also available in **Options > Properties**.

- K, k - Increase or decrease space between the tube and petals (**Radial offset**)
- L, l - Increase or decrease space between rings (**Z offset**)
- M, m - Increase or decrease space between petals (**Angular offset**)
- +, - - Increase or decrease petal outline width (**Petal outline width**)

### Custom cameras

To save a custom camera position, press Ctrl+1 to 5. Restore it later by pressing 1 to 5. (Custom camera position 1..5 options)

- e - Set perspective view
- z - Set history view
- s - Set front view

### Options

The following options are also available in **Options > Properties**.

- **Anti-aliasing level** - Set the smoothness of the line rendering. Higher levels are best, but not all graphic cards can support it, and it can reduce performance.
- **Petal pop-out speed** - The petal inflate speed for a new petal. When set to the maximum, petals directly appear fully inflated.
- **Activate ring sliding** - Enable or disable ring slide along the tube. Deactivating this option can improve low-performance conditions.
- **Draw petal outline** - Set to display the black outline surrounding each petal. Improves the readability in most cases.
- **Display metrics skeleton/name/cylinder** - Set to display the tube skeleton, name, or cylinder.

## Technical dependencies

Time View is a live history tool. This means that the GUI stores the history data (210 intervals of data in a circular buffer fashion) from the time it is started. The memory requirements of Time View on the end station running the GUI depend on the cluster size. A typical 500-node cluster can require 2GB to 3GB of RAM. Memory consumption does not impact the admin node. For larger clusters, the memory consumption can exceed 4GB requiring a 64-bit JVM on the GUI client side.

Because of the high memory and CPU/GPU consumption, Time View is limited to displaying 10 metrics at a time.

Hewlett Packard Enterprise recommends using OpenGL hardware acceleration for a higher quality experience, such as improved graphics and faster activation of anti-aliasing.

Time View has been tested using Java in many environments.

For more information, see the following:

[\*\*Troubleshooting Time View\*\*](#) on page 58

## Troubleshooting Time View

If **Time View** issues an `OutOfMemory` [...] error, increase the maximum HEAP memory usage of the GUI. To specify the memory consumption allowed for the JVM, set the `--Xmx` JVM argument when starting the CLI. In the GUI, edit `CMU_GUI_MB` (specified in MB) in `cmuserver.conf`.

---

(!) **IMPORTANT:** Setting this value too high might cause the cluster manager to issue `Unable to start JVM` messages on hosts with insufficient memory or hosts running a 32-bit JVM. Hewlett Packard Enterprise recommends a 64-bit JVM and requires it for large clusters.

---

If **Time View** stops running, a restart button displays below the **Time View** panel.

Some GPUs might not support anti-aliasing levels set to 8. In this case, the cluster manager displays the following:

- Black strips display on the left and right of **Time View**
- Cylinders display above the rings making the visualization inoperable

The preceding visual errors indicate a problem with the anti-aliasing level. If you observe either of the preceding errors, set anti-aliasing to a lower value, such as 4. If you continue to observe visual errors, set the value to 0.

## Archiving custom groups

Monitoring data for deleted custom groups can be archived and visualized later as history data.

## Procedure

1. Access the **Custom Group Management** window.
2. In the **Select a Custom Group** field, use the drop-down list to select the custom group to delete.
3. Click **Delete**.

A window displays asking if you would like to archive the selected custom group.

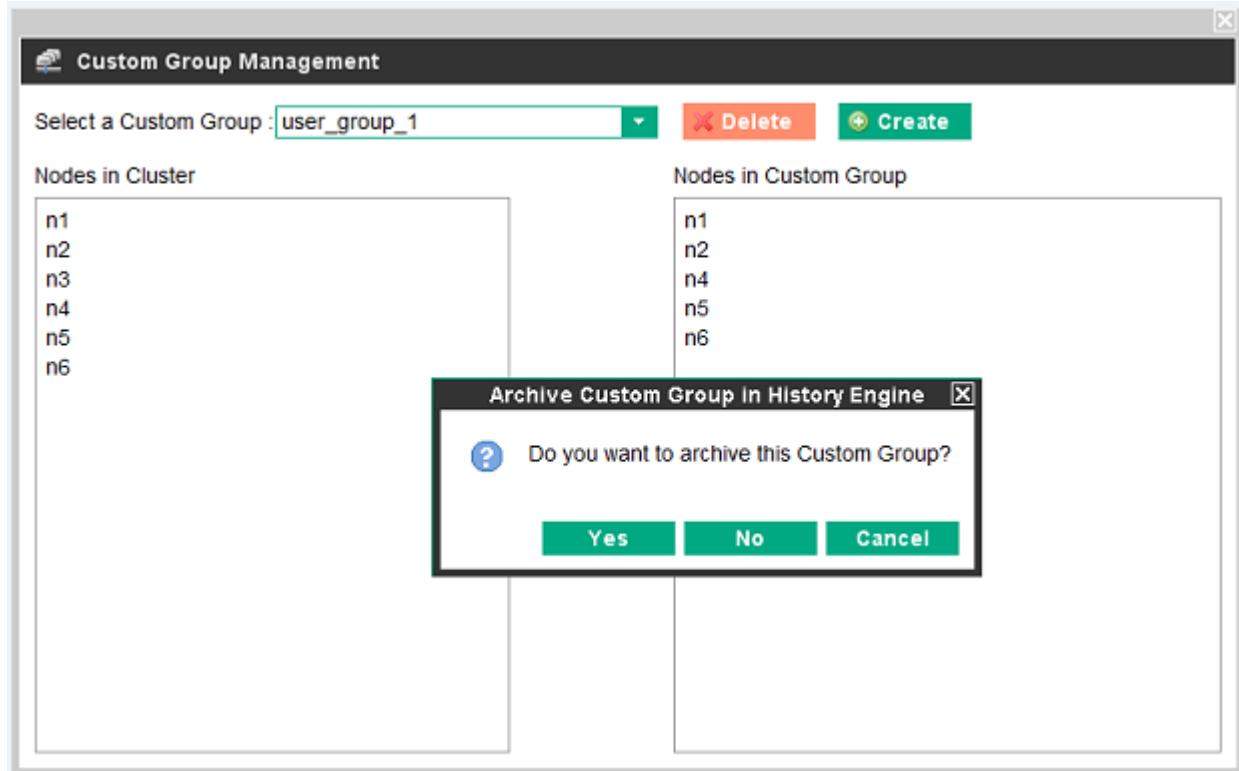


Figure 14: Archiving deleted custom groups

4. Click **Yes** to archive the group.

After the custom group is deleted, it displays in the **Archived Custom Groups** list in the left-frame tree.



Figure 15: Archived custom groups

---

**NOTE:** Custom groups can also be archived using the `cmu_del_custom_group` command. For more information, see the `cmu_del_custom_group` manpage.

---

## Visualizing history data

When selecting an archived custom group in the left-frame tree, a static **Time View** picture appears in the central frame. This picture shows the activity view of the custom group during its existence. All options available with **Time View** are also available when visualizing archived custom groups.

## Limitations for displaying archived custom groups

To display an archived custom group, the following conditions must be satisfied.

- Time must not exceed 24 hours.
- The number of nodes must not exceed 4096.
- The number of metrics must not exceed 100.
- The product of the three parameters above must not exceed 409600.

The table below displays examples of valid combinations of these three parameters.

**Table 2: Valid archived custom group parameters**

Nodes	Metrics	Hours	Nodes Metrics Hours
4096	10	10	409600
4096	5	20	409600
4096	100	1	409600
256	100	12	307200
2048	8	24	393216
1024	16	24	393216

- ① **IMPORTANT:** If the criteria above are not met, the archived custom group will not display. Instead, a warning message displays.

## Stopping monitoring

To close the monitoring GUI, click the **X** button in the upper right corner of the main monitoring window.

When the monitoring GUI closes, the monitoring engine does not automatically stop. To stop the monitoring engine on the cluster, on the toolbar, click the **Monitoring** tab and select **Stop Monitoring Engine**.

## Customizing monitoring, alerting, and reactions

### Action and alert files

Sensors, alerts, and alert reactions are described in the `/opt/clmgr/etc/ActionAndAlertsFile.txt` file.

The following is an example of the contents of this file:

```
#This is a CMU action and alerts description file  
#=====
```

```

#
# ACTIONS
#
#
#
#-----KERNEL VERSION, RELEASE, BIOS VERSIONS-----#
kernel_version      "kernel version"  9999999   string Instantaneous
release uname -r
#-----CPU-----#
#
#-- Native
cpuload           "% cpu load (raw)"      1      numerical
MeanOverTime     100      % awk '/cpu / {printf"%d\n", $2+$3+$4}' /proc/stat
#-- Collectl
#cpuload "% cpu load (normalized)" 1 numerical Instantaneous 100 % COLLECTL
(cputotals.user) + (cputotals.nice) + (cputotals.sys)
#cpuload "% cpu load (normalized)" 1 numerical Instantaneous 100 % COLLECTL
100 - (cputotals.idle)
#
#-----MEMORY-----#
#
#-- Native
#memory_used    "% memory used"      1      numerical
Instantaneous  100      % free | awk ' BEGIN { freemem=0;
totalmemory=0; } /cache:/ { freemem=$4; } /Mem:/ { totalmemory=$2; } END
{ printf "%d\n", (((totalmemory-freemem)*100)/totalmemory); } '
#
#
#
ALERTS
#
#
#cpu_freq_alert "CPU frequency is not nominal"  1      24      100
<      % sh -c "b=`cat /sys/devices/system/cpu/cpu0/cpufreq/
scaling_cur_freq`;a=`cat /sys/devices/system/cpu/cpu0/cpufreq/
cpuinfo_max_freq`;echo 100 \* \$b / \$a |bc"
login_alert      "Someone is connected"  3      24      0      >
login(s)        w -h | wc -l
root_fs_used    "The / filesystem is above 90% full"  4      24
90      >      % df / | awk '{ if (\$6=="/") print \$5}' | cut -f 1 -d
% -
#reboot_alert    "Node rebooted"      4      24      5      < rebooted awk
'{printf "%.1f\n",\$1/60}' /proc/uptime
# The line below allows to report MCE errors; be careful for possible false
positives
#mce_alert       "The kernel has logged MCE errors; please check /var/log/
mcelog" 5 60 1 > lines wc -l /var/log/mcelog |cut -f 1 -d ' '
#
#
ALERT_REACTIONS
#
#
#login_alert "Sending mail to root" ReactOnRaise echo -e "Alert
'CMU_ALERT_NAME' raised on node(s) CMU_ALERT_NODES. \n\nDetails:\n`/opt/
clmgr/bin/pdsh -w CMU_ALERT_NODES 'w -h'`" | mailx -s "CMU: Alert
'CMU_ALERT_NAME' raised." root
#

```

```

#root_fs_used "Sending mail to root" ReactOnRaise echo -e "Alert
'CMU_ALERT_NAME' raised on node(s) CMU_ALERT_NODES. \n\nDetails:\n`/opt/
clmgr/bin/pdsh -w CMU_ALERT_NODES 'df /'` | mailx -s "CMU: Alert
'CMU_ALERT_NAME' raised!" root
#
#reboot_alert "Sending mail to root" ReactOnRaise echo -e "Alert
'CMU_ALERT_NAME' raised on node(s) CMU_ALERT_NODES. \n\nDetails:\n`/opt/
clmgr/bin/pdsh -w CMU_ALERT_NODES 'uptime'` | mailx -s "CMU: Alert
'CMU_ALERT_NAME' raised." root
#

```

Each line corresponds to a sensor, alert, or an alert reaction. Lines prefixed with a pound character (#) are ignored. Lines cannot begin with a leading white space. Sensors are placed at the beginning of the file, between the `ACTIONS` and `ALERTS` tags. Each alert is in the middle of the file between the `ALERTS` and `ALERT_REACTIONS` tags, and each alert reaction is at the end of the file below the `ALERT_REACTIONS` tag.

Most sensors have both a native line and a commented `collectl` line. To use `collectl` for collecting monitoring data, enable it by removing the comment from the corresponding sensor line.

---

**NOTE:** Using `collectl` requires additional steps. For information, see the following:

**[Using collectl to gather monitoring data](#)** on page 66

---

## Actions

Each action contains the following fields:

- **Name**

The name of the sensor as it appears in the GUI. The name must only consist of letters.

- **Description**

A quote-contained string describing what the sensor is. This appears in the GUI.

- **Time multiple**

An integer value that determines when the sensors are monitored.

For example, if the monitoring has a default timer of 5 seconds:

- A time multiple of 1 means that the value is monitored every 5 seconds.
- A time multiple of 2 means that the value is monitored every 10 seconds.

For more information, see **[Changing the monitoring interval](#)** on page 89.

- **Data type**

This can be numerical or a string. A string sensor cannot be displayed in the pie graphs by the interface.

- **Measurement method**

This can be either Instantaneous or MeansOverTime.

Instantaneous returns the sensor value immediately. MeansOverTime returns the difference between the current value and the previous value divided by the time interval.

For example, if the sensors return values of 1, 100, 50, and 100 at 4 continuous time steps of 5 seconds:

- The Instantaneous option returns values of 1, 100, 50, and 100.
- The MeanOverTime option returns values of N/A, 19.8, -10, and 10.

- **Max value**

Used by the interface to create the pie graphs at the beginning. If a greater value is returned by a sensor, the maximum value is automatically updated in the interface.

- **Unit**

The unit of the sensor. The GUI uses this measurement.

- **Command**

The command to be executed by the script. This can be an executable or a shell command. The executable and the shell command must be available on the compute nodes.

## Alerts

Each alert contains the following fields:

- **Name**

The name of the sensor as it appears in the GUI. The name must only consist of letters.

- **Description**

A quote-contained string describing what the sensor is. This appears in the GUI.

- **Severity**

An integer from 1 to 5, where 1 is minor and 5 is fatal. This appears by the interface.

- **Time multiple**

An integer value that determines when the sensors are monitored.

For example, if the monitoring has a default timer of 5 seconds:

- A time multiple of 1 means that the value is monitored every 5 seconds.
- A time multiple of 2 means that the value is monitored every 10 seconds.

- **Threshold**

The threshold that must not be overcome by the sensor.

- **Operator**

The comparison operator between the sensor and the threshold. Only **>** is available.

- **Unit**

The unit of the sensor. The GUI uses this measurement.

- **Command**

The command to be executed by the script. This can be an executable or a shell command. The executable and the shell command must be available on compute nodes.

## Alert reactions

Each alert reaction contains the following fields:

- **Name(s)**

The names of one or more alerts from the **ALERTS** section. The reaction is associated with each of the alerts. If an alert is specified in more than one reaction, then only the first reaction is taken. The list of alert Names is white-space separated.

- **Description**

A quote-contained brief description of the reaction.

- **Condition**

The condition under which to perform the reaction.

- ReactOnRaise - Execute the reaction when the alert shows as raised and the previous state of the alert was lowered.
- ReactAlways - Execute the reaction when the alert shows as raised, subject to the alert's time multiple. For example, if the monitoring has a default timer of 5 seconds and the alert's time multiple is 6, the reaction will trigger every 30 seconds as long as the alert is raised.

- **Command**

The command to execute. This can be a single-line shell command, a shell script, or an executable file. Scripts and executable files must be available on the admin node.

The following keywords are supported within the **Command**. Each keyword is substituted globally (throughout the command line) using the defined values.

- CMU\_ALERT\_NAME

The name of the alert that caused the reaction.

- CMU\_ALERT\_LEVEL

The level of the alert.

- CMU.React\_MESSAGE

The text of the **Description** for this reaction.

- CMU\_ALERT\_NODES

A list of all of the nodes that raised the alert during the current monitoring pass. The list is condensed in the form provided by `cmu_condense_nodes`.

- CMU\_ALERT\_NODES\_EXPANDED  
The list displayed from CMU\_ALERT\_NODES, but expanded, ordered, and separated by commas.
- CMU\_ALERT\_VALUES  
The list of alert values. This list is comma-separated and ordered like the names of CMU\_ALERT\_NODES\_EXPANDED.
- CMU\_ALERT\_TIMES  
The time the alert was triggered on each node. This list is comma-separated and ordered like the names of CMU\_ALERT\_NODES\_EXPANDED.
- CMU\_ALERT\_SEQUENCE\_FILE  
The path to the sequence file containing the alerts and alert values from the monitoring pass that triggered the reaction. Analyze this file with the /opt/clmgr/bin/cmu\_monstat command.

---

**NOTE:** To protect the admin node from large numbers of concurrent reactions, a reaction will only launch on behalf of compute nodes that do not have previous instances of the reaction still running. Limit the command runtime of a reaction if the reaction is expected to be triggered frequently.

---

## Modifying sensors, alerts, and alert reactions

Several optional sensors, alerts, and alert reactions are commented in the ActionAndAlertsFile.txt file.

To modify a sensor, alert, or alert reaction, do the following.

- Comment a sensor, alert, or alert reaction to stop monitoring it.
- Uncomment a sensor, alert, or alert reaction to start monitoring it.
- Modify a sensor, alert, or alert reaction line to change its parameters.
- Add your own sensors, alerts, or alert reactions by adding a line to the **ACTIONS**, **ALERTS**, or **ALERT\_REACTIONS** section.

Modifications in the ActionAndAlertsFile.txt file are only applied when the monitoring daemons are restarted.

## Restarting the monitoring daemons

### Procedure

1. Change the ActionAndAlertsFile.txt file on the admin node.
2. Exit the GUI.
3. Enter the following command to stop the daemons:

```
# systemctl stop cmu
```

4. Enter the following command to start the daemons:

```
# systemctl start cmu
```

5. Start the GUI.

For information, see the following:

[Launching the GUI](#) on page 14

## Using `collectl` to gather monitoring data

The default method for specifying commands to collect monitoring data is described in the following:

[Actions](#) on page 62

This default method is referred to as **native mode**.

The cluster manager provides an alternative method that uses the `collectl` tool to gather monitoring data. Data appears using the same interface as the native mode.

### Installing and starting `collectl` on compute nodes

#### Procedure

1. Install the package on the compute nodes.

```
# mount /dev/cdrom /mnt
# cd /mnt/Tools/collectl
# rpm -ivh collectl-3.x.x-x.noarch.rpm
# scp collectl-x.x.x.src.tar.gz goldenNode1:/tmp
# ssh goldenNode1
# cd /tmp
# tar zxf collectl-x.x.x.src.tar.gz
# cd collectl-x.x.x
# ./INSTALL
```

2. If not already done, install the monitoring RPM on the compute nodes as described in the following:

[\(Conditional\) Installing the monitoring client](#) on page 45

3. Edit the `/etc/collectl.conf` file as follows.

```
DaemonCommands = -s+dcmnNE --import misc --export lexpr -A server -i5
```

---

**!** **IMPORTANT:** For diskless configurations, only use the `DaemonCommands` options provided in the example above. Do not use any option that causes disk I/O. For disk-based clusters, you can configure `collectl` to archive metrics locally.

---

4. Start `collectl`.

```
# /etc/init.d/collectl start
Starting collectl: [ OK ]
```

5. Configure `collectl` to start automatically.

```
# chkconfig --add collectl
collectl 0:off 1:off 2:on 3:on 4:on 5:on 6:off
```

## Modifying the ActionAndAlerts.txt file

The ActionAndAlerts.txt file contains definitions for using collectl monitoring. These lines are commented out so that the cluster manager works in native mode by default, without the use of collectl.

To switch to collectl monitoring, edit the following file:

```
/opt/clmgr/etc/ActionAndAlertsFile.txt
```

The edits to make are as follows:

- Insert a # character in column 1 of the line below the #- Native line.
- Remove the # character from the line below the #- Collectl line.

For example, the following file shows collectl monitoring enabled:

```
#-----CPU-----#
#- Native
#cpuload "% cpu load (raw)" 1 numerical MeanOverTime 100 % awk '/cpu / {printf "%d\n", $2+$3+$4}' /proc/stat
#- Collectl
cpuload "% cpu load (normalized)" 1 numerical Instantaneous 100 % COLLECTL (cputotals.user) + (cputotals.nice) + (cputotals.sys)
```

In the line below the #-Collectl line, the command field must start with the string COLLECTL in uppercase letters. The line continues with a series of collectl variables included in parentheses and connected with arithmetical operators. In this example, the cpuload metric reports the sum of cputotals.user, cputotals.nice, and cputotals.sys.

For a full list of available collectl variables, run the collectl command interactively, as follows:

```
# collectl -c 1 -s+C --export lexpr
```

The -c 1 option runs one shot only. The command output is the list of collectl variables and the current value.

```
waiting for 1 second sample...
sample.time 1217858718.002
cputotals.user 1
cputotals.nice 0
cputotals.sys 0
cputotals.wait 7
cputotals.irq 0
cputotals.soft 0
cputotals.steal 0
cputotals.idle 90
ctxint.ctx 239
ctxint.int 1073
ctxint.proc 4
ctxint.rung 152
disktotals.reads 0
disktotals.readkbs 0
disktotals.writes 11
disktotals.writekbs 80
nettots.kbin 4
nettots.pktin 49
nettots.kbout 6
nettots.pktout 17
cpuinfo.user.cpu0 0
cpuinfo.nice.cpu0 0
cpuinfo.sys.cpu0 0
cpuinfo.wait.cpu0 0
```

```
cpuinfo.irq.cpu0 0
cpuinfo.soft.cpu0 0
cpuinfo.steal.cpu0 0
cpuinfo.idle.cpu0 100
cpuinfo.intrpt.cpu0 0
cpuinfo.user.cpu1 0
cpuinfo.nice.cpu1 0
cpuinfo.sys.cpu1 0
cpuinfo.wait.cpu1 11
cpuinfo.irq.cpu1 0
cpuinfo.soft.cpu1 0
cpuinfo.steal.cpu1 0
cpuinfo.idle.cpu1 89
cpuinfo.intrpt.cpu1 0
cpuinfo.user.cpu2 4
cpuinfo.nice.cpu2 0
cpuinfo.sys.cpu2 2
cpuinfo.wait.cpu2 0
```

Use these variables to create the monitoring lines.

Native lines and `collectl` lines can be mixed in the `ActionAndAlertFile.txt` file.

For more information about using and fine tuning `collectl`, see the following:

<http://collectl.sourceforge.net/>.

## Installing and configuring colplot to plot collectl data

- 
- ① **IMPORTANT:** Due to the high-bandwidth nature of this setup, do not to use this option for diskless configurations.
- 

### Procedure

1. On the admin node, create an NFS export directory to store `collectl` data from compute nodes.

```
# mkdir /var/log/collectl
# vi /etc/exports
```

2. Add the following line:

```
/var/log/collectl *(rw,sync,no_all_squash,no_root_squash)
```

3. Refresh exports:

```
# exportfs -r
```

4. Install the `collectl-utils` package.

---

**NOTE:** To install `collectl-utils`, you must install the `gnuplot` RPM.

---

```
# mount /dev/cdrom /mnt
# cd /mnt/Tools/collectl
# cd /tmp
# tar zxf colplot-x.x.x.src.tar.gz
# cd colplot-x.x.x
# ./INSTALL
```

5. Copy `colplot` HTML files to the cluster manager web directory.

These files are located in different directories depending on whether your admin node runs RHEL or SLES.

- RHEL: # cp -a /var/www/html/colplot /opt/clmgr/www/colplot
- SLES: # cp -a /srv/www/htdocs/colplot /opt/clmgr/www/colplot

6. In the `/etc/colplot.conf` file, change the default `colplot` plot directory to point to the common `collectl` directory.

```
# vi /etc/colplot.conf
#PlotDir = /opt/hp/collectl/plotfiles
PlotDir = /var/log/collectl
```

7. If not already done, install the `collectl` RPM on the compute nodes.

```
# mount /dev/cdrom /mnt
# cd /mnt/Tools/collectl
# scp /mnt/Tools/collectl/collectl-x.x.x.src.tar.gz goldenNode1:/tmp
# ssh goldenNode1
# cd /tmp
# tar zxf collectl-x.x.x.src.tar.gz
# cd collectl-x.x.x
# ./INSTALL
```

8. If the `collectl` package is already installed, ensure that `collectl` is stopped.

```
# /etc/init.d/collectl stop
```

9. Import the common directory created on the administration server for `collectl`.

```
# mkdir /var/log/collectl
# vi /etc/fstab
X.X.X.X:/var/log/collectl /var/log/collectl nfs defaults 0 0
```

For `X.X.X.X`, specify the address of the admin node.

10. Modify the `collectl` configuration file to save data to be plotted in the common directory.

```
# vi /etc/collectl.conf
DaemonCommands = -s+dcmnNE --import misc --export lexpr -A server -i5 -f /var/log/collectl -P -oz -r
00:01,7
```

11. Restart `collectl`.

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

## Viewing plotted data

You can use a web browser to view the `collectl` plotted data gathered by `colplot`. To do this, use the address of the admin node followed by `/colplot`.

```
http://X.X.X.X/colplot
```

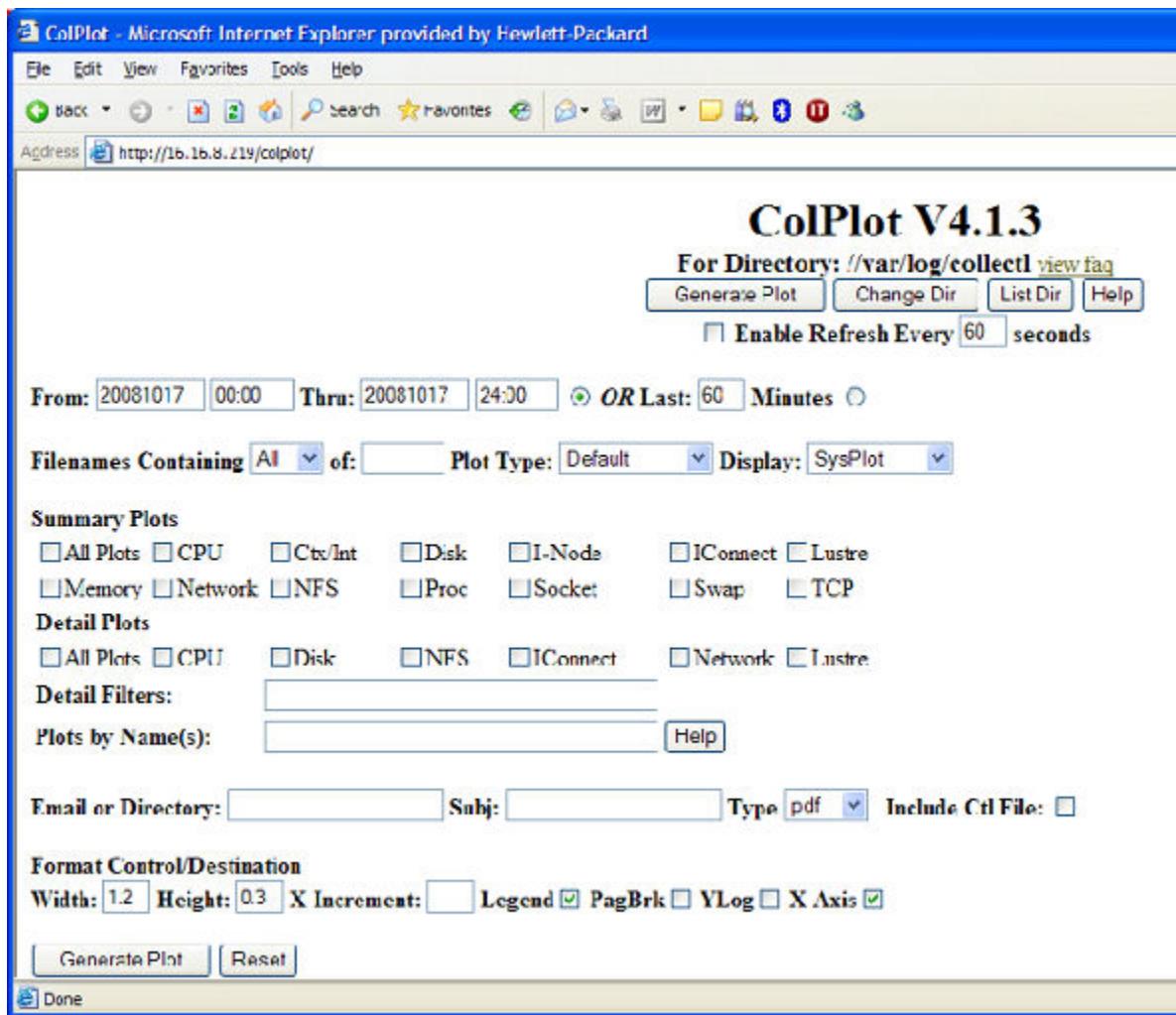
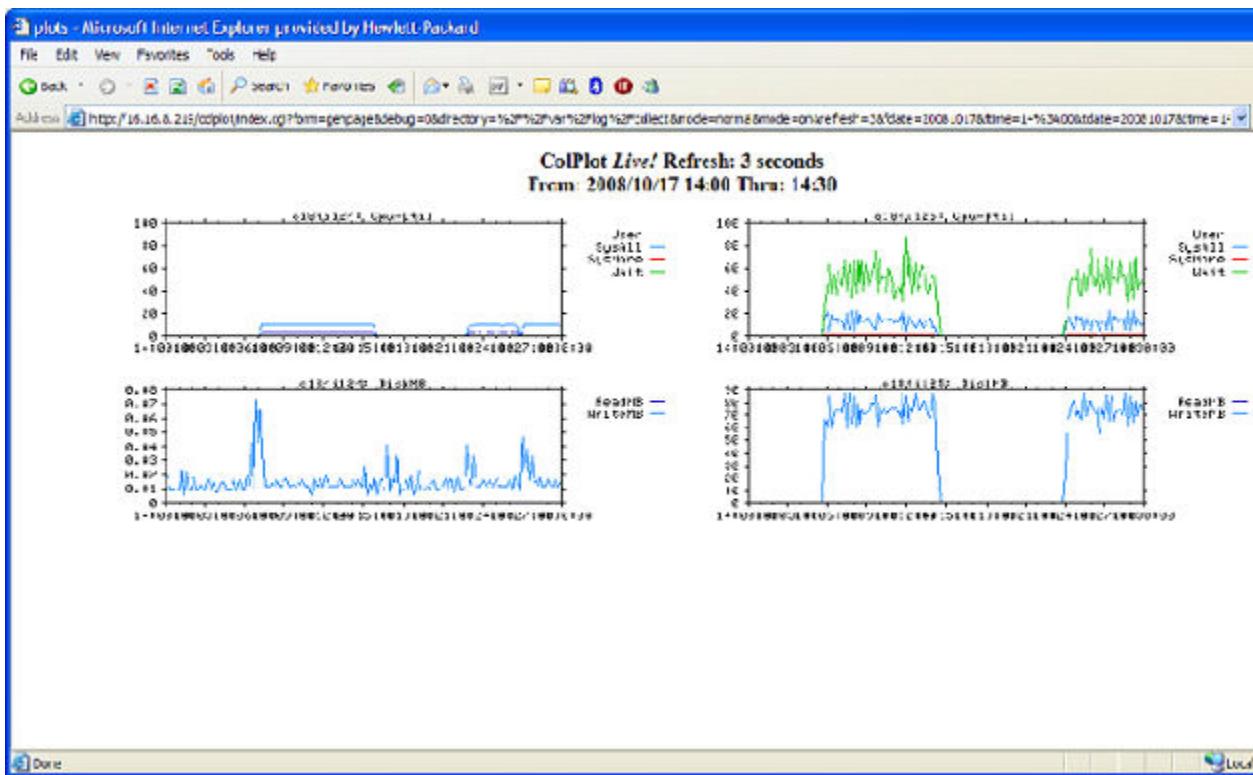


Figure 16: ColPlot window

Select the plotting options and click **Generate Plot**.



**Figure 17: ColPlot results**

## Monitoring NVIDIA GPUs

If your client nodes contain NVIDIA GPUs and are running version 270.xx.xx or later of the NVIDIA GPU driver, you can use the cluster manager to monitor the GPUs.

- ➊ **IMPORTANT:** If you are currently using the cluster manager to monitor NVIDIA GPUs, update the monitoring to use the latest GPU monitoring features. To update, deploy an updated set of monitoring images to your client nodes. To deploy the images, see the following:

**(Conditional) Installing the monitoring client** on page 45

When older GPU metrics exist, you must disable and then enable the metrics to install the latest metrics. To remove the NVIDIA GPU metrics and enable the new metrics, remove the previously existing NVIDIA GPU metrics after enabling the NVIDIA GPU monitoring.

## Installing the NVIDIA GPU driver

If not already done, install the NVIDIA GPU driver version 270.xx.xx or later on your client nodes. This can be done in two ways:

- Method 1:

Manually install the NVIDIA GPU driver on one of the client nodes, capture the client image, and deploy the remaining clients with the new image.

- Method 2:

Use the `/opt/clmgr/contrib/install_nvidia.pl` script to install the NVIDIA GPU driver on all running clients.

For more information, see the /opt/clmgr/contrib/install\_nvidia README file.

## Enabling NVIDIA GPU monitoring

To enable NVIDIA GPU monitoring, the /opt/clmgr/etc/ActionAndAlertsFile.txt file must be updated with entries for GPU monitoring.

### Procedure

1. In the GUI, click **Monitoring > Platform-specific metrics**.

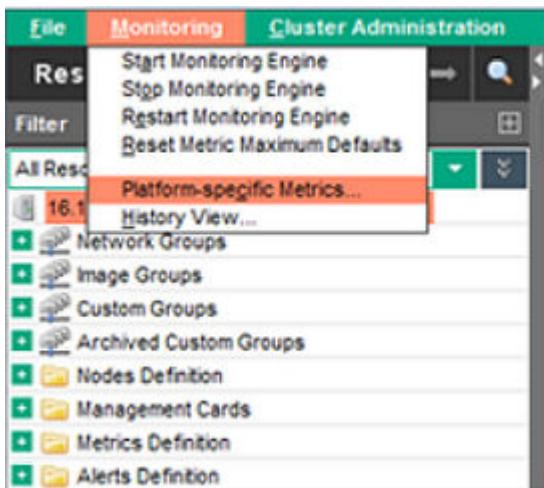


Figure 18: Monitoring metrics

A dialog displays for enabling and disabling platform-specific metrics.

2. From the metric category drop-down list, select **NVIDIA GPU** to add metrics for NVIDIA GPUs.

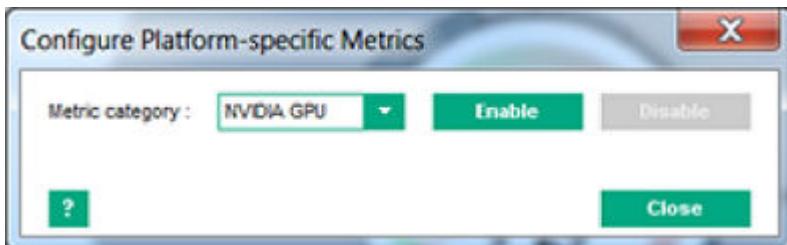


Figure 19: Configure platform-specific metrics

3. Click **Enable**.

4. Click **Close**.

A prompt displays requesting a monitoring restart. You can choose not to restart the monitoring at this time. To start gathering new NVIDIA GPU metrics, restart monitoring.

To view the new metrics in the GUI, select the metrics from the Monitoring sensors list as described in Global Cluster view.

---

**NOTE:** Not all metrics are supported by all NVIDIA GPUs and some lesser used metrics may be commented out within the `ActionAndAlertsFile.txt` file. To introduce/remove metrics from the Monitoring sensors list, you can uncomment/comment out the associated lines inside of the `ActionAndAlertsFile.txt` file as described in the following:

**Global cluster view** on page 48

The cluster manager dynamically determines if a client has working GPUs when monitoring is initially started after installation on the client. This monitoring process allows for configurations that have clients with GPUs and clients without GPUs. If the GPUs are not working when monitoring is started (or GPUs are added at a later date), redeploy monitoring to the client and restart monitoring to ensure that the GPUs are recognized.

To redeploy the monitoring client, see the following:

**(Conditional) Installing the monitoring client** on page 45

---

## Removing NVIDIA monitoring metrics

### Procedure

1. In the GUI, click **Monitoring > Platform-specific Metrics**.
2. Select **NVIDIA GPU** from the metric category drop-down list.
3. Click **Disable**.
4. Click **Close**.
5. Restart monitoring.

## Monitoring cluster manager alerts in HPE Systems Insight Manager

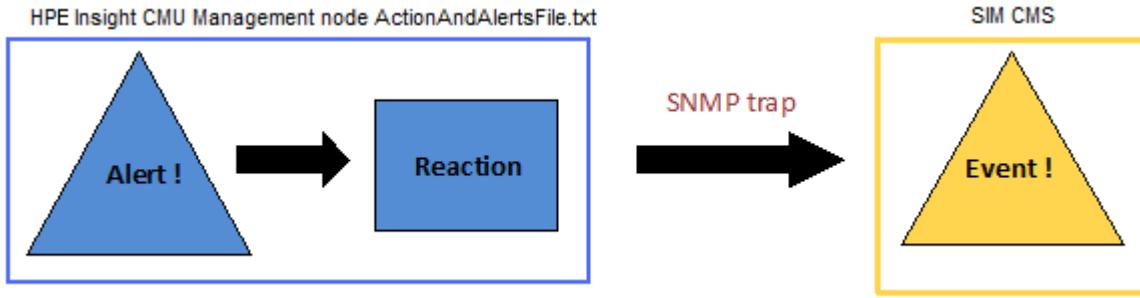
- (!) **IMPORTANT:** The information in this topic assumes that you have knowledge of HPE Systems Insight Manager (SIM) and Simple Network Management Protocol (SNMP).
- 

Using SIM, you can create an environment to monitor cluster manager alerts. This can be accomplished many ways. This topic offers one possible model. You can use this example as an outline for creating a model that works for your environment.

Cluster manager alerts are similar to SIM events. However, alerts and events are defined and responded to differently in each product. In the cluster manager, alerts are defined using the `ActionAndAlertFile.txt` file. An alert is raised when the result of an alert command exceeds a defined threshold relative to its declared operator. When this occurs, the alert is displayed in the GUI.

To convey the result of an alert to the SIM Central Management Server (CMS), you can use the cluster manager alert reaction feature and one of the SIM supported event protocols like SNMP traps. Create an alert reaction for the cluster manager alert you want to convey. For the alert command, provide a command or script that sends the selected SNMP trap to the SIM CMS.

HPE OpenView NNM, the perl `SNMP_util` CPAN module, and the Net-SNMP Open Source package are commonly used to send SNMP traps. All cluster manager client alerts are handled through the admin node. Cluster manager alert reaction keywords, such as `CMU_ALERT_NODES`, can be used to convey the names of the nodes that raised the alert through the SNMP trap.



**Figure 20: Cluster manager alert converted to SIM event**

To create a complete model for conveying cluster manager alerts to SIM, you might choose to create your own SNMP Management Information Base (MIB) to handle the alerts you define.

For information about alerts, see the following:

[Alerts](#) on page 63

For information about alert reactions, see the following:

[Alert reactions](#) on page 64

For information about how to configure SNMP with SIM, or how to compile and customize MIBs with SIM, see the following:

[Systems Insight Manager User Guide](#)

## Extended metric support

By default, the cluster manager runs pre-configured commands on every compute node to extract metric values and aggregates these values on the admin node for display with the GUI. The extended metric support allows users to gather metrics on the admin node with scripting or any other method and pass the data directly into the GUI for display.

The extended metric support consists of the following:

- The keyword EXTENDED, which is configured in the `ActionAndAlertsFile.txt` file to identify each extended metric.
- The `cmu_submit_extended_metrics` command found in the `/opt/clmgr/bin/` directory.

### Example of extended metric support

Suppose you want to configure the cluster manager to monitor workload scheduler information. Typically, this information can be viewed by executing a single command that displays status information for all of the compute nodes. Using extended metric support, you can create a script that periodically executes this command, parses the data into a simple format, and passes it to the GUI for monitoring.

The following example shows how to monitor the number of nodes allocated to jobs. In this example, the list of allocated nodes are gathered from SLURM, an open-source workload scheduler. Use the SLURM command to list the nodes that are currently allocated.

```
[root@cmumaster ~]# sinfo -t alloc -o "%N" -h
node[10-12,14,20-21,33-39,41-48,50-55]
[root@cmumaster ~]#
```

Use a cluster manager tool to expand names and create a space-separated list of allocated nodes:

```
[root@cmumaster ~]# sinfo -t alloc -o "%N" -h | /opt/clmgr/tools/
cmu_expand_names -s " "
node10 node11 node12 node14 node20 node33 node34 node35 node36 node37 node38
```

```
node39 node41 node42 node43 node44 node45 node46 node47 node48 node50 node51  
node52 node53 node54 node55  
[root@cmumaster ~]#
```

To apply this example to your workload scheduler, replace this SLURM command with the appropriate command from your workload scheduler.

To submit this data to the cluster manager:

```
/opt/clmgr/bin/cmu_submit_extended_metrics
```

The ‘help’ option describes how to submit data into the cluster manager.

```
[root@cmumaster ~]# /opt/clmgr/bin/cmu_submit_extended_metrics -h  
Usage: /opt/clmgr/bin/cmu_submit_extended_metrics -f <filename>
```

The *filename* must exist and contain per-node metric data in the following format:

```
BEGIN_NODE <nodelist>  
metric1_name metric1_value  
metric2_name metric2_value  
...  
metricN_Name metricN_value  
BEGIN_NODE <nodelist>  
metric1_name metric1_value  
metric2_name metric2_value  
...
```

The *nodelist* value is typically one node name, but it can be a space-separated list of node names if the subsequent metrics and values apply to a given list of nodes.

To obtain and submit this data, write a bash script:

```
[root@cmumaster ~]# cat ./allocated_nodes.sh  
#!/bin/bash  
  
CMU_EXPAND=/opt/clmgr/tools/cmu_expand_names  
CMU_SUBMIT=/opt/clmgr/bin/cmu_submit_extended_metrics  
CMU_NODES=/opt/clmgr/bin/cmu_show_nodes  
file=/tmp/alloc_nodes.txt  
  
alloc_nodes=`sinfo -t alloc -o "%N" -h | $CMU_EXPAND -s " " `  
  
# find the list of nodes that are unallocated  
all_nodes=`$CMU_NODES`  
free_nodes=""  
for n in $all_nodes; do  
    found=0  
    for a in $alloc_nodes; do  
        if [ $a = $n ]; then  
            found=1  
            break  
        fi  
    done  
    if [ $found = 0 ]; then  
        free_nodes="$free_nodes $n"  
    fi  
done  
  
# write the file and submit to CMU  
rm -f $file
```

```

echo "BEGIN_NODE $alloc_nodes" > $file
echo "allocated 1" >> $file
echo "BEGIN_NODE $free_nodes" >> $file
echo "allocated 0" >> $file

$CMU_SUBMIT -f $file

```

[root@cmumaster ~]#

The preceding script obtains and submits the `allocated` metric to the cluster manager.

The last step is to configure this new metric in the `ActionAndAlertsFile.txt` file:

```

allocated "nodes allocated to users" 2 numerical Instantaneous 1 alloc
EXTENDED /root/allocated_nodes.sh

```

The following table provides explanations of each field in the line example above:

**Table 3: Extended metric fields**

Field	Description	Example above
Name	Name of the extended metric.	allocated
Description	Brief description of the extended metric.	nodes allocated to users
Time multiple	The time-to-live setting. Multiply this number by 5 to determine the number of seconds that the extended metric data is considered valid after being received. If no new metric data is received after this time interval expires, the GUI marks the extended metric data as <b>Inactive Action</b> .	2
Data type	Description of the format of the extended metric data. This is either <code>numerical</code> or <code>string</code> .	numerical
Measurement method	<p>Either <code>Instantaneous</code> or <code>MeanOverTime</code>.</p> <p><code>Instantaneous</code> displays the latest value.</p> <p><code>MeanOverTime</code> displays the difference between the current value and the previous value divided by the time interval.</p>	Instantaneous
Max value	The value used by the GUI to initialize the metric pies. If this value is exceeded, then the scale of the metric pie adjusta to the new maximum value.	1
Unit	The unit of the extended metric.	alloc

*Table Continued*

Field	Description	Example above
EXTENDED keyword	Indicates that the metric is submitted by cmu_submit_extended_metrics.	EXTENDED
Script or command name	The script or command that collects, formats, and submits the metric to the cluster manager.	/root/allocated_nodes.sh

After you finish editing the `ActionAndAlertsFile.txt` file, restart monitoring and restart the GUI. These actions enable the modifications to take effect. The monitoring utility schedules the script to run according to the time multiple setting.

### Extended metric support on large clusters

On a large cluster, your data-gathering script might require additional time to complete. If your data-gathering script takes an unsatisfactory amount of time to gather, parse, and submit the data to the cluster manager, then determine that length of time and adjust the time multiple setting to ensure that enough time is allocated to complete the script. Otherwise, the cluster manager might display the metric as an **Inactive action** in the GUI. To determine the run time of the script, use the `time` command, as follows:

```
[root@cmumaster ~]# time ./allocated_nodes.sh
real    7.036s
[root@cmumaster ~]#
```

Then, divide the running time by 5 to get the time multiple. In this example:  $7/5=2$ .

---

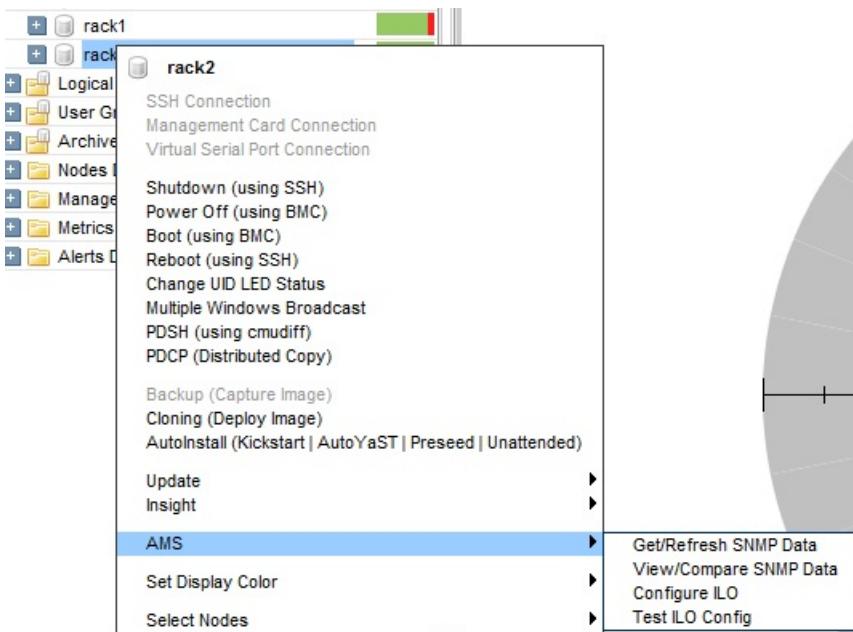
**NOTE:** Your data-gathering script might obtain, parse, and submit more than one metric. A typical example of this is gathering multiple temperature readings from a single source, such as through IPMI or from the Onboard Administrator of an HPE Blade enclosure. In this case, you only need to configure the script to run with one metric in the `ActionAndAlertsFile.txt` file. The other metrics gathered by this script can be configured in the `ActionAndAlertsFile.txt` file without any command after the EXTENDED keyword.

Several preconfigured scripts in `/opt/clmgr/contrib/` can gather and submit metrics to the cluster manager with the extended monitoring support. These scripts include `README` files that document how they work and how they can be configured in the cluster manager. Copy and modify these scripts freely to operate correctly on your cluster:

- The `cmu_IPMI_monitoring` script gathers IPMI metrics by querying the management card (also known as the baseboard management controller (BMC) for this information).
- The `cmu_OA_monitoring` script gathers power and temperature readings from the Onboard Administrators of Blade enclosures.
- The `cmu_get_ganglia_metrics` script gathers metrics from the ganglia monitoring daemons.

## Configuring iLO 4 AMS extended metric support for power and temperature

The cluster manager can gather server metrics from iLO 4 (or later) through the iLO Agentless Monitoring Support (AMS) and submit the metrics to the cluster manager through the EXTENDED monitoring support. To enable HPE AMS monitoring, update the `/opt/clmgr/etc/ActionsAndAlertsFile.txt` with entries for AMS monitoring. When enabled, an AMS menu item is added to the cluster manager GUI. By default, the AMS metrics gather power and temperature data. You can configure additional metrics by using the AMS menu item.



**Figure 21: Verify AMS submenu**

Client nodes require HPE iLO to be configured to provide AMS data. For more information, see the following:

#### **Configuring the HPE iLO SNMP port** on page 79

The following options are included in AMS with the HPE iLO:

- Configuring the iLO on each server with a public SNMP read-only port and enabling AMS.
- Requesting and displaying a full data report of all available iLO data.
- Configuring iLO SNMP data as metrics to be monitored.

#### **Procedure**

**1. Click Monitoring > Platform-specific Metrics.**

The **Configure Platform-specific Metrics** dialog box appears.

**2. In the Metric category drop-down list, select HPE AMS.**

**3. Click Enable.**

**4. Click Close.**

The GUI prompts you to restart the cluster manager. To start gathering new AMS metrics, restart the cluster manager. The cluster manager does not start gathering metrics until after you restart the cluster manager.

To view the new metrics, select the metrics from the Monitoring sensors list. For information about the list, see the following:

**Global cluster view** on page 48

## Removing AMS monitoring metrics

### Procedure

1. Click **Monitoring > Platform-specific metrics**.
2. In the metric category drop-down list, select **HPE Agentless Monitoring Service**.
3. Click **Disable**.
4. Restart monitoring.

**NOTE:** When AMS monitoring is disabled, the AMS metrics are no longer gathered. However, the AMS menu item remains in the GUI.

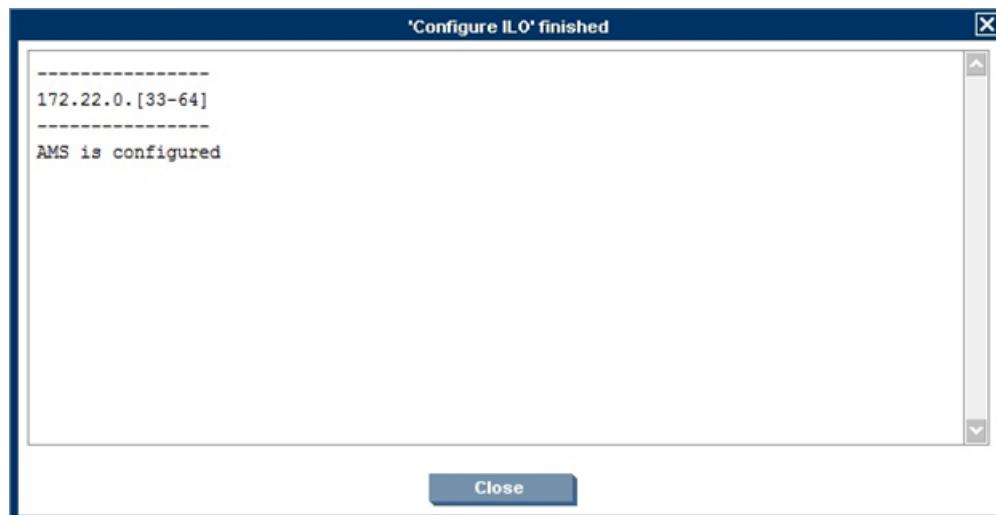
## Configuring the HPE iLO SNMP port

To configure the iLO SNMP port and enable AMS, select the servers in the left panel of the GUI and right-click to bring up the remote management menu. Then, select **AMS > Configure iLO**.



**Figure 22: Configure iLO SNMP port**

When this command completes, a summary displays the iLOs that were successfully configured.

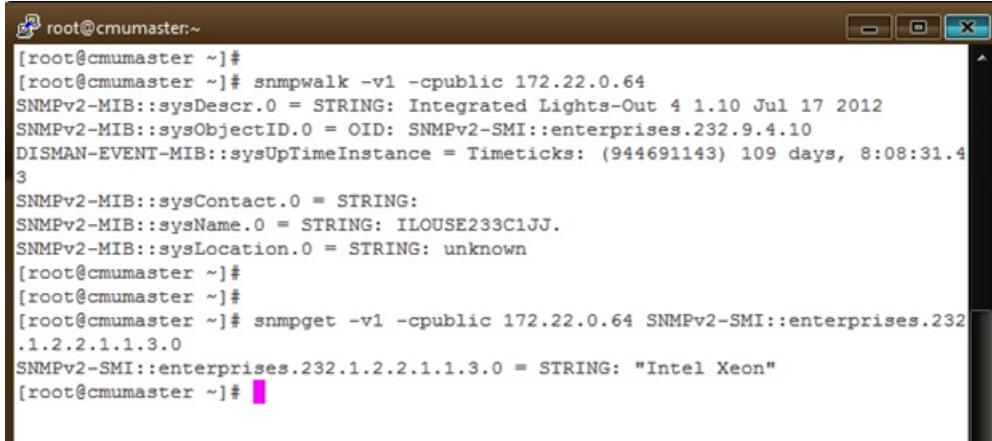


**Figure 23: Configure iLO finished**

To test which iLOs were configured with AMS, select the nodes. Then, select **AMS > Test iLO Config**.

## Accessing and viewing the HPE iLO data via SNMP

Enabling the AMS functionality in the iLO makes it possible to use an SNMP query to retrieve iLO data.

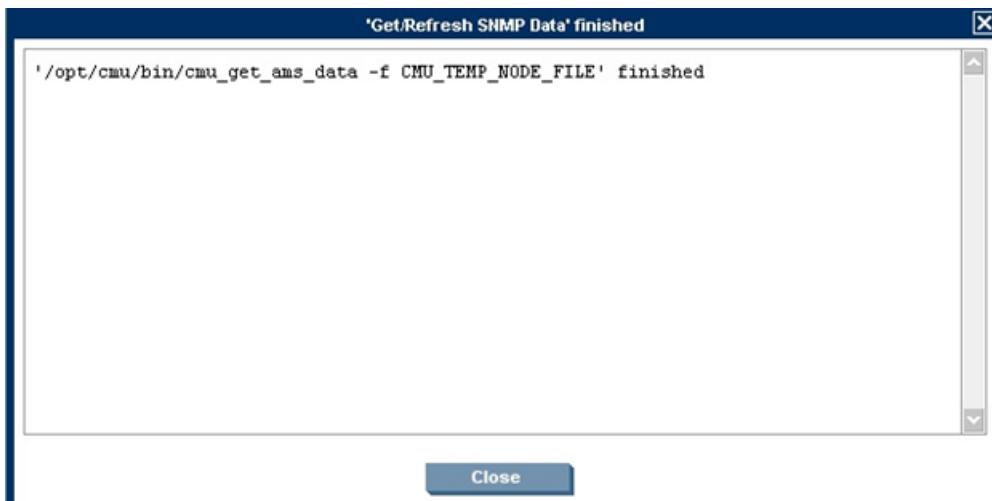


```
[root@cmumaster ~]# snmpwalk -v1 -cpublic 172.22.0.64
SNMPv2-MIB::sysDescr.0 = STRING: Integrated Lights-Out 4 1.10 Jul 17 2012
SNMPv2-MIB::sysObjectID.0 = OID: SNMPv2-SMI::enterprises.232.9.4.10
DISMAN-EVENT-MIB::sysUpTimeInstance = Timeticks: (944691143) 109 days, 8:08:31.4
3
SNMPv2-MIB::sysContact.0 = STRING:
SNMPv2-MIB::sysName.0 = STRING: ILOUSE233C1JJ.
SNMPv2-MIB::sysLocation.0 = STRING: unknown
[root@cmumaster ~]#
[root@cmumaster ~]#
[root@cmumaster ~]# snmpget -v1 -cpublic 172.22.0.64 SNMPv2-SMI::enterprises.232
.1.2.2.1.1.3.0
SNMPv2-SMI::enterprises.232.1.2.2.1.1.3.0 = STRING: "Intel Xeon"
[root@cmumaster ~]#
```

**Figure 24: SNMP query**

The published HPE management information bases (MIBs) define the SNMP strings. These MIBs are available on the internet. The cluster manager includes a subset of these MIBs in `/opt/clmgr/snmp_mibs/`. The cluster manager uses the MIBs to translate the SNMP OID strings into human-readable strings when gathering and viewing the data through the **AMS** menu options in the GUI.

To request a complete set of SNMP data from one or more iLOs, select the nodes in the GUI. Then, select **AMS > Get/Refresh SNMP Data**. When the process completes, the following window displays.



**Figure 25: Get/Refresh SNMP data**

To view the data, select the nodes in the GUI and click **AMS > View/Compare SNMP Data**.

```

responses: 1 ( node39 ), no data: 0
reference: node39
ignored: <none>
output: 2458 lines

| SNMPv2-SMI::enterprises.232.1.1.1.0 .cpqSeMibRevMajor.0 = INTEGER: 1
| SNMPv2-SMI::enterprises.232.1.1.2.0 .cpqSeMibRevMinor.0 = INTEGER: 31
| SNMPv2-SMI::enterprises.232.1.1.3.0 .cpqSeMibCondition.0 = ok(2)
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.0 .cpqSeCpuUnitIndex.0 = INTEGER: 0
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.1 .cpqSeCpuUnitIndex.1 = INTEGER: 1
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.2.0 .cpqSeCpuSlot.0 = INTEGER: 0
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.2.1 .cpqSeCpuSlot.1 = INTEGER: 0
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.3.0 .cpqSeCpuName.0 = STRING: "Intel Xeon"
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.3.1 .cpqSeCpuName.1 = STRING: "Intel Xeon"
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.4.0 .cpqSeCpuSpeed.0 = INTEGER: 1800
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.4.1 .cpqSeCpuSpeed.1 = INTEGER: 1800
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.5.0 .cpqSeCpuStep.0 = INTEGER: 7
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.5.1 .cpqSeCpuStep.1 = INTEGER: 7
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.6.0 .cpqSeCpuStatus.0 = ok(2)
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.6.1 .cpqSeCpuStatus.1 = ok(2)
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.7.0 .cpqSeCpuExtSpeed.0 = INTEGER: 100
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.7.1 .cpqSeCpuExtSpeed.1 = INTEGER: 100
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.8.0 .cpqSeCpuDesigner.0 = intel(2)
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.8.1 .cpqSeCpuDesigner.1 = intel(2)
| SNMPv2-SMI::enterprises.232.1.2.2.1.1.9.0 .cpqSeCpuSocketNumber.0 = INTEGER: 1

```

**Figure 26: View/Compare SNMP data**

Before displaying in the window, the data is piped through the `CMU_Diff` filter, in case you selected more than one node to compare the data. The first column is the SNMP OID string. The second column is the MIB definition for the corresponding SNMP OID string, and the third column is the value of the SNMP OID string.

**NOTE:** In some cases, the SNMP OID string value is translated to a human-readable string based on the definitions provided in the MIB.

The cluster manager translates the SNMP OID string and value into the strings defined by the MIB to make the data easier to read and understand.

## Configuring HPE iLO SNMP metrics

Many of the SNMP data values are static, but some are volatile and important to monitor, such as current temperature and power usage. The cluster manager provides a tool to query a set of pre-configured SNMP OID strings and display the strings in the GUI. The pre-configured SNMP OID strings and their corresponding metric names are in `/opt/clmgr/etc/cmu_ams_metrics`.

To configure additional SNMP OID strings for monitoring, you can add the strings to the file with their corresponding metric name in the cluster manager.

```

root@cmumaster:~# cat /opt/cmu/etc/cmu_ams_metrics
#
# This file is part of the CMU AMS support.
# This file maps SNMP OIDs to CMU metric names.
#
# First column is the SNMP OID.
# Second column is the CMU metric name.
# The optional 'SUM' keyword in the third column
# is used to add the values of multiple SNMP OIDs
# into a single CMU metric.
#
SNMPv2-SMI::enterprises.232.6.2.6.8.1.4.0.1 amb1_temp
SNMPv2-SMI::enterprises.232.6.2.6.8.1.4.0.2 cpul_temp
SNMPv2-SMI::enterprises.232.6.2.6.8.1.4.0.3 cpu2_temp
SNMPv2-SMI::enterprises.232.6.2.9.3.1.7.0.1 power1 SUM power
SNMPv2-SMI::enterprises.232.6.2.9.3.1.7.0.2 power2 SUM power
SNMPv2-SMI::enterprises.232.6.2.9.3.1.7.0.3 power3 SUM power
SNMPv2-SMI::enterprises.232.6.2.9.3.1.7.0.4 power4 SUM power

```

**Figure 27: cmu\_ams\_metrics**

The `/opt/clmgr/bin/cmu_get_ams_metrics` command gathers the data for the SNMP OID strings and submits it to the cluster manager. This command needs a file (`-f filename`) containing the list of nodes with the iLOs to be queried. Alternatively, you can request a query of all of the iLOs on all nodes in the cluster (`-a`).

Use the `-d` option on the `cmu_get_ams_metrics` command to display the data. This option can be used to confirm that the `cmu_get_ams_metrics` command is retrieving the correct SNMP data from the given nodes. Otherwise, the default action is to submit the data to the cluster manager. This may not work if the metrics are not yet configured in the cluster manager.

```

root@cmumaster:~#
[root@cmumaster ~]# /opt/cmu/bin/cmu_get_ams_metrics -h
usage : /opt/cmu/bin/cmu_get_ams_metrics -h
        /opt/cmu/bin/cmu_get_ams_metrics [-d] -a|-f <node_file> [-m mfile] [-s s
ecs]

-d      : display metric data
-a      : use all nodes in CMU
-f <f>: file containing list of nodes
-m <m>: file containing SNMP OID and corresponding metric name
        default metric file is /opt/cmu/etc/cmu_ams_metrics
-s <s>: number of seconds to sleep between reruns
[root@cmumaster ~]#
[root@cmumaster ~]# echo node33 > /tmp/node
[root@cmumaster ~]#
[root@cmumaster ~]# /opt/cmu/bin/cmu_get_ams_metrics -d -f /tmp/node
BEGIN_NODE node33
amb1_temp 29
cpul_temp 40
cpu2_temp 40
power 68

```

**Figure 28: cmu\_get\_ams\_metrics**

The last step is to configure the SNMP metrics. Add the following lines to the `/opt/clmgr/etc/ActionAndAlertsFile.txt` file.

```
amb1_temp "ambient temp" 4 numerical Instantaneous 60 Celsius EXTENDED /opt/
clmgr/bin/cmu_get_ams_metrics -f /opt/clmgr/etc/cmu_gen8_nodes
```

```
cpu1_temp "CPU 1 temp" 4 numerical Instantaneous 60 Celsius EXTENDED  
cpu2_temp "CPU 2 temp" 4 numerical Instantaneous 60 Celsius EXTENDED  
power "power usage" 4 numerical Instantaneous 100 watts EXTENDED
```

The `cmu_get_ams_metrics` command is only added to the `amb1_temp` metric because it only needs to get invoked once per monitoring cycle, and it provides the cluster manager with all 4 pre-configured SNMP-based metrics.

In this example, `cmu_get_ams_metrics` is invoked with the `-f /opt/clmgr/etc/cmu_gen8_nodes` option. If all of the nodes in your cluster have iLO 4 or later and are configured to support SNMP port queries, then you can replace this option with `-a`. Otherwise, create a file that contains a list of the nodes that support iLO SNMP queries, and provide that file to this command.

After you are finished configuring the `ActionAndAlertsFile.txt` file, restart the monitoring and restart the GUI. These new metrics then appear in the GUI display.

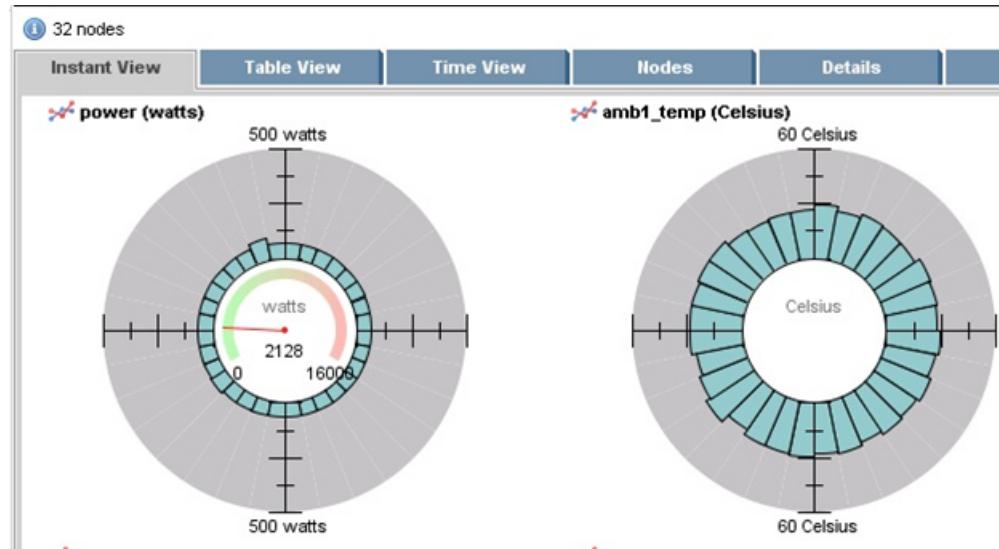


Figure 29: Instant view display

## Configuring HPE Moonshot power and temperature monitoring

You can use the cluster manager to monitor Moonshot server power and temperature. The `/opt/clmgr/tools/cmu_get_moonshot_metrics` command gathers and submits Moonshot-specific data to the cluster manager.

By default, this command queries all of the iLO Chassis Managers (iLOCMs) associated with the nodes that are configured in the cluster manager database and submits the power and temperature data to the cluster manager. Hewlett Packard Enterprise recommends running the command with the `-d` option, first, to verify that the data can be retrieved from the chassis. This option gathers and displays the data, rather than submitting it to the cluster manager.

After you have verified that the data can be gathered, update the `/opt/clmgr/etc/ActionAndAlertsFile.txt` file with the entries for Moonshot monitoring to enable Moonshot power and temperature monitoring.

```

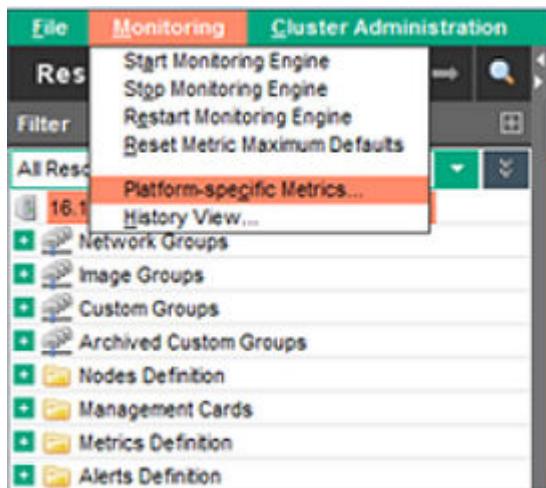
Last login: Thu Aug 28 00:20:19 2014 from 16.212.57.196
[root@cmutay1 ~]# /opt/cmu/tools/cmu_get_moonshot_metrics -h
usage : cmu_get_moonshot_metrics [-i <ILOCM IP>] [-d]
        cmu_get_moonshot_metrics [-i <ILOCM IP>] [-d]
        -d          Display data to stdout rather than submitting it to /opt/cmu/bin/cmu_submit_extended_metrics
        -h          Prints this help text
        -i <ILOCM IP>  Gather metrics from the specified ILOCM IP.
[root@cmutay1 ~]# /opt/cmu/tools/cmu_get_moonshot_metrics -d
BEGIN_NODE m700_c01-n1
power 6.51445
amb1_temp 46
cpu1_temp 58
BEGIN_NODE m700_c01-n2
power 6.51445
amb1_temp 46
cpu1_temp 52
BEGIN_NODE m700_c01-n3
power 6.51445
amb1_temp 46
cpu1_temp 53
BEGIN_NODE m700_c01-n4
power 6.51445
amb1_temp 46
cpu1_temp 58
BEGIN_NODE m700_c02-n1
power 6.5228
amb1_temp 42
cpu1_temp 52
BEGIN_NODE m700_c02-n2
power 6.5228

```

**Figure 30: Metrics data window**

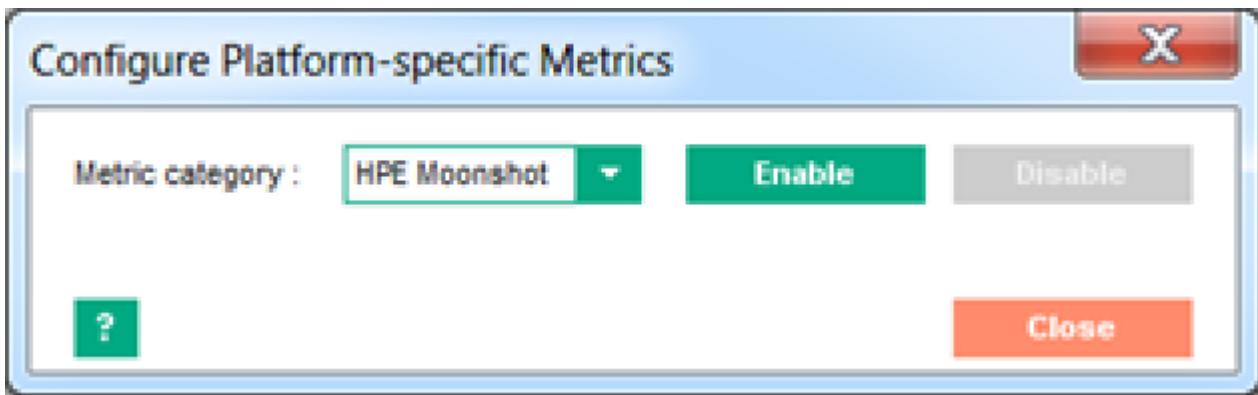
### Procedure

1. In the GUI, click **Monitoring > Platform-specific Metrics**.



**Figure 31: Monitoring metrics**

2. In the **Configure Platform-specific Metrics** dialog box, from the **Metric category** drop-down list, select **HPE Moonshot**.



**Figure 32: Configure platform-specific metrics**

3. Click **Enable**.

4. Click **Close**.

A prompt displays requesting you to restart CMU monitoring. You can choose not to restart monitoring at this time; however, the new Moonshot metrics will not be gathered until monitoring is restarted.

To view the new metrics in the GUI, select the metrics from the Monitoring sensors list as described in the following:

[Global cluster view](#) on page 48

## Removing Moonshot monitoring metrics

### Procedure

1. In the GUI, click **Monitoring > Platform-specific Metrics**.
2. In the metric category drop-down list, select **HPE Moonshot**.
3. Click **Disable**.
4. Restart monitoring.

## Metric arrays

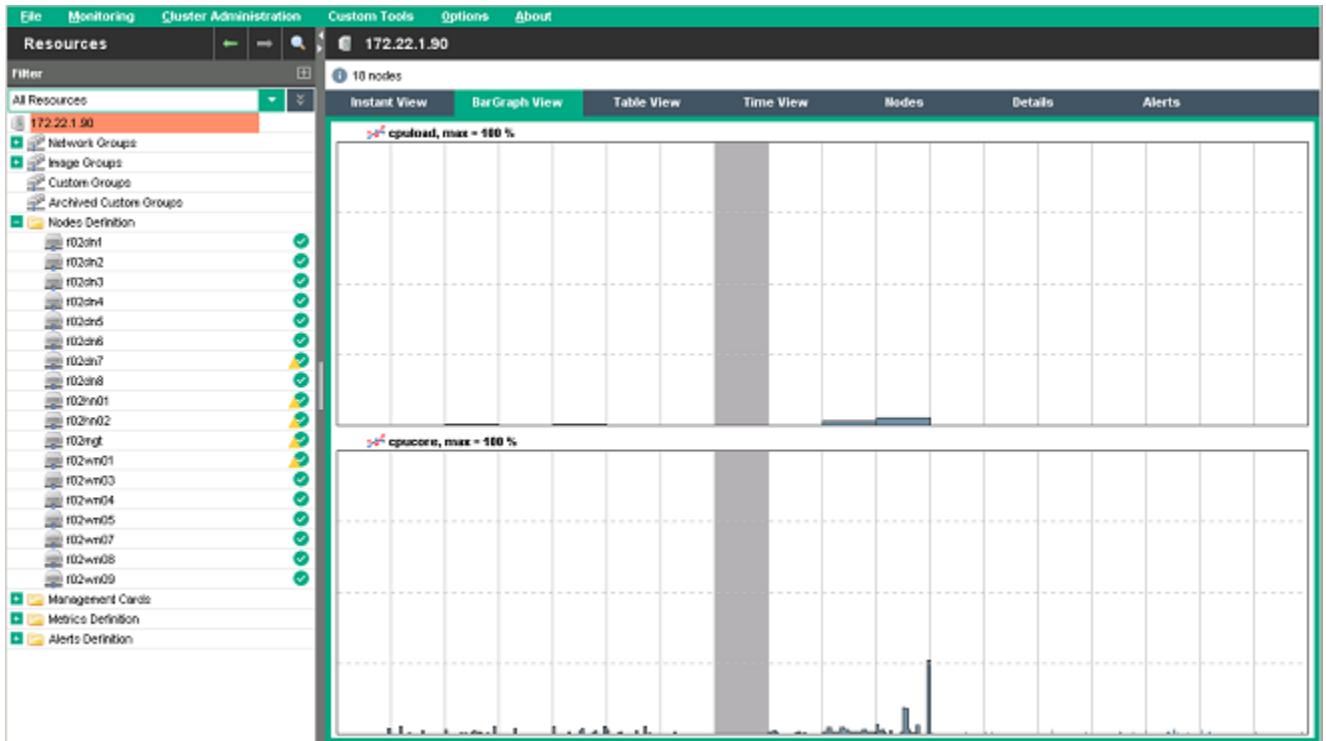
Metric arrays support multiple values for a metric on a single node. For example, on a node with 20 cores, the general `cpuload` metric shows the overall CPU utilization for the entire node.

Metric array support allows metrics to consist of an array of values. The GUI displays those values as single entities.



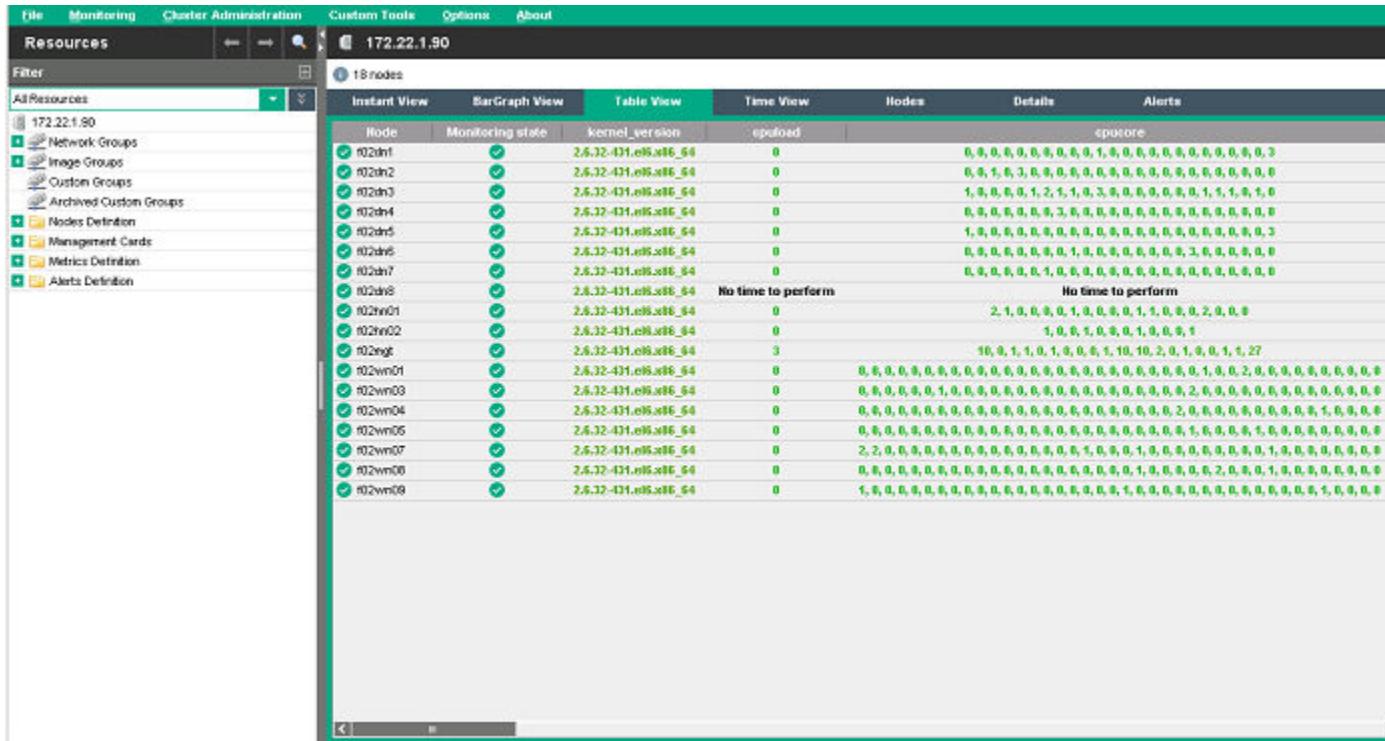
**Figure 33: Metric arrays instant view**

The figure above shows a single value for cpuload, but multiple values for cpucore. The cpuload metric is represented as a single petal, while cpucore has multiple bars within the petal. For very large systems, the detail bars in **Instant View** may not be visible or useful. To turn off the detail bars and show a single petal with the aggregation of the metric array values, go to **Options > Properties > Instant View** and uncheck **Display details in petals**. The resulting cpucore graphic looks very similar to cpuload.



**Figure 34: Metric arrays bar graph view**

The preceding figure shows the metric arrays **BarGraph View** with single bars per node for `cpuload` and multiple bars per node for `cputcore`.



**Figure 35: Metric arrays table view**

The preceding figure shows the metric arrays table view with a single value for `cpuload` and a comma-separated list of values for `cpucore`.

Hover the cursor over a node to display the details of the metric array.

cpucore (%)		
Node	Metric	00:00:00
f02hn02	cpucore[0]	2
	cpucore[1]	0
	cpucore[2]	0
	cpucore[3]	1
	cpucore[4]	0
	cpucore[5]	2
	cpucore[6]	0
	cpucore[7]	2
	cpucore[8]	4
	cpucore[9]	0
	cpucore[10]	0
	cpucore[11]	0

**Figure 36: Metric array tooltip**

**NOTE:** Metric arrays do not display in detail in **Time View**. The aggregation of the metric array values is used to determine the height of each petal in **Time View**. Hover the cursor over a node to display the detail values.

## Defining actions to return metric arrays

**The cpuload action is:**

```
cpuload "% cpu load (raw)" 1 numerical MeanOverTime 100 % awk '/cpu / {currcpu=$2+$3+$4;next} /cpu[0-9]+ / {numcpus++;next} END {printf "%d\n", currcpu/numcpus}' /proc/stat
```

**The awk command returns:**

```
[/opt/clmgr/etc ] # awk '/cpu / {currcpu=$2+$3+$4;next} /cpu[0-9]+ / {numcpus++;next} END {printf "%d\n", currcpu/numcpus}' /proc/stat  
59299544
```

59299544 is a single value.

**The cpcore action is:**

```
coreload "% cpu load/core (raw)" 1 numerical MeanOverTime 100 % awk '/cpu[0-9]+ / {printf "%d ",$2+$3+$4}' /proc/stat
```

**The awk command returns:**

```
[/opt/clmgr/etc ] # awk '/cpu[0-9]+ / {printf "%d ",$2+$3+$4}' /proc/stat  
105130692 51083122 28757733 21903178 17440549 15009018 12639859  
11613686 11213076 12860924 134539668 78196763 66838602 37706896  
31154883 53078625 35853133 31791896 185442487 243746309
```

This command returns 20 values for the 20 cores on the system.

For information about defining the actions to return metric or sensor values, see the following:

**Action and alert files** on page 60

### Defining metric arrays using collectl counters

The `collectl` tool does not provide arrays, but does provide multiple values. For information about configuring `collectl` to provide values to the cluster manager, see the following:

**Using collectl to gather monitoring data** on page 66

To show the names that `collectl` provides to the cluster manager, run `collectl` with the `--export lexpr` option.

```
# collectl -c 1 -s+C --export lexpr  
waiting for 1 second sample...  
sample.time 1217858718.002  
cputotals.user 1  
cputotals.nice 0  
cputotals.sys 0  
cputotals.wait 7  
cputotals.irq 0  
cputotals.soft 0  
cputotals.steal 0  
cputotals.idle 90  
ctxint.ctx 239  
ctxint.int 1073  
ctxint.proc 4  
ctxint.rung 152  
disktotals.reads 0  
disktotals.readkbs 0  
disktotals.writes 11  
disktotals.writekbs 80  
nettotals.kbin 4  
nettotals.pktin 49  
nettotals.kbout 6  
nettotals.pktout 17  
cpuinfo.user.cpu0 0  
cpuinfo.nice.cpu0 0
```

```

cpuinfo.sys.cpu0 0
cpuinfo.wait.cpu0 0
cpuinfo.irq.cpu0 0
cpuinfo.soft.cpu0 0
cpuinfo.steal.cpu0 0
cpuinfo.idle.cpu0 100
cpuinfo.intrpt.cpu0 0
cpuinfo.user.cpu1 0
cpuinfo.nice.cpu1 0
cpuinfo.sys.cpu1 0
cpuinfo.wait.cpu1 11
cpuinfo.irq.cpu1 0
cpuinfo.soft.cpu1 0
cpuinfo.steal.cpu1 0
cpuinfo.idle.cpu1 89
cpuinfo.intrpt.cpu1 0
cpuinfo.user.cpu2 4
cpuinfo.nice.cpu2 0
cpuinfo.sys.cpu2 2
cpuinfo.wait.cpu2 0

```

The preceding output shows a pattern for the various CPUs.

```

cpuinfo.wait.cpu0 0
cpuinfo.idle.cpu0 100
cpuinfo.wait.cpu1 11
cpuinfo.idle.cpu1 89
cpuinfo.wait.cpu2 0
cpuinfo.idle.cpu2 90

```

If you define `cputcore` as follows, it requests a search for all names with `cpuinfo.idle.cpu#` and `cpuinfo.wait.cpu#`, starting at `cpu0` until it encounters a name that does not exist.

```
cpucore "% cpu core" 1 numerical Instantaneous 100 % COLLECTL 100 - (cpuinfo.idle.cpu[0-]) - (cpuinfo.wait.cpu[0-])
```

### Collectl metric array rules

The `ActionAndAlerts.txt` file contains the metric array rules.

```

#Some lines below are setup to return multiple values. These should be returned as a space separated list.
# For collectl, patterns in the form [#-] or [ #-] or [a-] or [a-n] are acceptable for generating multiple
values.
# [#-] or [a-] - open ended patterns. Starts at low value provided, stops when expanded collectl variable
does not exist.
# EG: cpuinfo.idle.cpu[0-] will return array "cpuinfo.idle.cpu0 cpuinfo.idle.cpu1 cpuinfo.idle.cpu2 ..."
# if collectl is monitoring detail cpu. Will start at cpu0 and stop when it encounters a cputn which
is not provided by collectl.
# [ #-] or [a-n] - closed patterns. Starts at low value provided, stops at high value provided. Only returns
values actually provided by collectl.
# EG: netinfo.kbout.eth[0-9] will return array similar to "netinfo.kbout.eth0 netinfo.kbout.eth1 ...
netinfo.kbout.eth9" but only including
# the values that collectl is providing. IE: If only eth5 and eth6 are defined, only "netinfo.kbout.eth5
netinfo.kbout.eth6" will be returned.
#

```

## Changing the monitoring interval

You can set the monitoring interval in the `CMU_MONITORING_INTERVAL` variable in the following file:

```
/opt/clmgr/etc/cmuserver.conf

# monitoring interval in seconds
# default setting is 5 seconds for HPC systems
# can be set to higher frequency values down to 1 second
```

```

# this will not be supported when the cluster size is
# higher than defined into CMU_MAX_NODES_SMALL_MON_INTERVAL

#CMU_MONITORING_INTERVAL=5

# max cluster size triggering the support warning
# when monitoring interval CMU_MONITORING_INTERVAL is below 5 seconds
# changing this value is not advised
# as it may impact the performance of the cluster and/or
# the performance of the admin node.
# the default value is 180 nodes

CMU_MAX_NODES_SMALL_MON_INTERVAL=180

```

Using a smaller monitoring interval time updates the client more quickly and improves accuracy. However, keep in mind the following.

- The cluster manager consumes more CPU cycles because metrics are gathered more often. This can affect actual loads.
- Archives fill faster and, therefore, use more space.
- Network traffic is higher.

If the monitoring interval is changed, remember to change the `collectl` interval time to match the monitoring interval if the `collectl` tool is used. Also, review the **Time Multiple** values in the `ActionAndAlerts.txt` file. Using a time multiple of 2 for the default monitoring time of 5 seconds means that the value is gathered every 10 seconds. If the monitoring interval time is set to 10 seconds, then the values are gathered every 20 seconds.

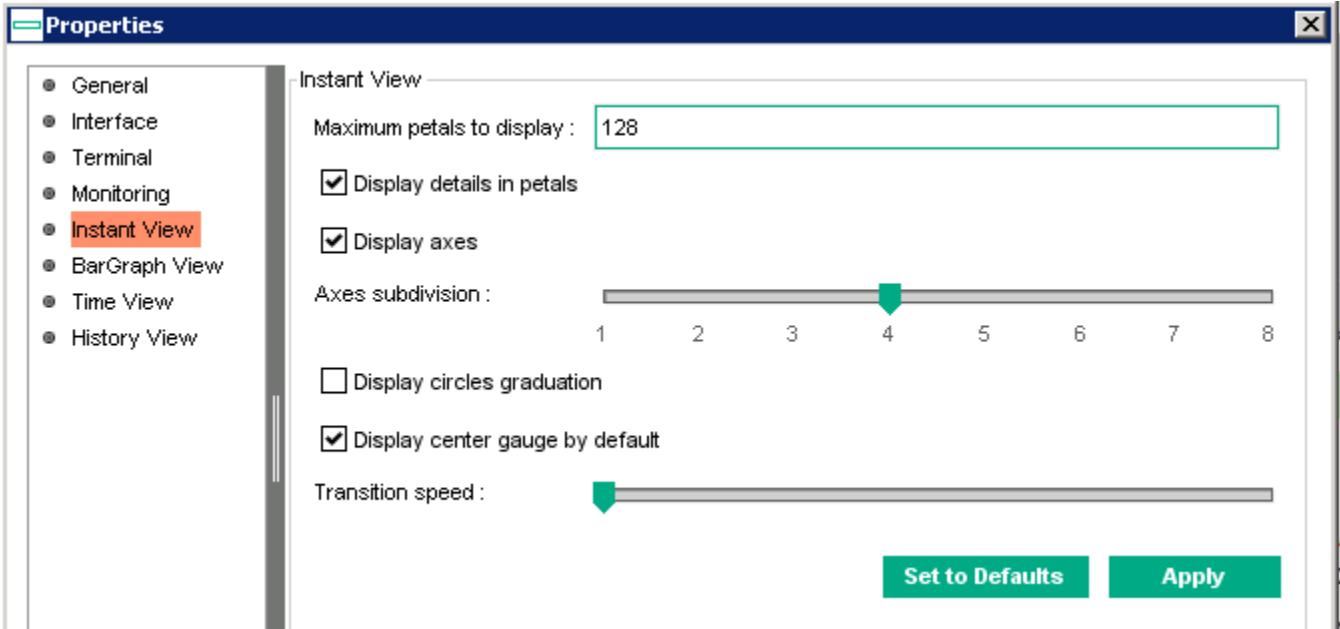
## Transitional display

The display interval is the time it takes for the cluster manager to update the display with new information.

On every interval, the GUI client is updated with the new values. These updates cause the petals or bars to jump from the current value to the new value. You can change the transition speed to allow a smooth change from the old values to the new values.

To change the transition speed in **Instant View**, click **Options > Properties > Instant View**, and slide the **Transition speed** needle to reflect the desired transition speed.

To change the transition speed in **Bar Graph View**, click **Options > Properties > Bar Graph View**, and slide the **Transition speed** needle to reflect the desired transition speed.



**Figure 37: Transition speed option**

To use the full interval for the transition, slide the needle to the far left. The bars reflect the new value only at the end of the interval.

To cause the display to transition from the old value to the new value during the first half of the interval, slide the needle to the middle, and the bars become fixed for the second half of the interval. To turn off the transition speed and cause the display to jump from one value to the next, slide the needle to the far right.

The following are the advantages and disadvantages of changing the transition speed:

- Advantages:
  - Improves readability by providing a smooth change from the old value to the new value.
  - Shows a general trend of activity, such as `cpuload` increasing.
  
- Disadvantages:
  - The current display is accurate only at the end of the interval. It takes an additional interval for the display to reflect the new value.
  - The client uses more CPU cycles on the client system because the screen display is updated several times during the interval.

# Managing a cluster

You can perform cluster management tasks on one or more nodes. The available tasks to perform depend on your privileges and the number of selected nodes.

## Administrator menu

The GUI has the following operator modes:

- Normal mode.

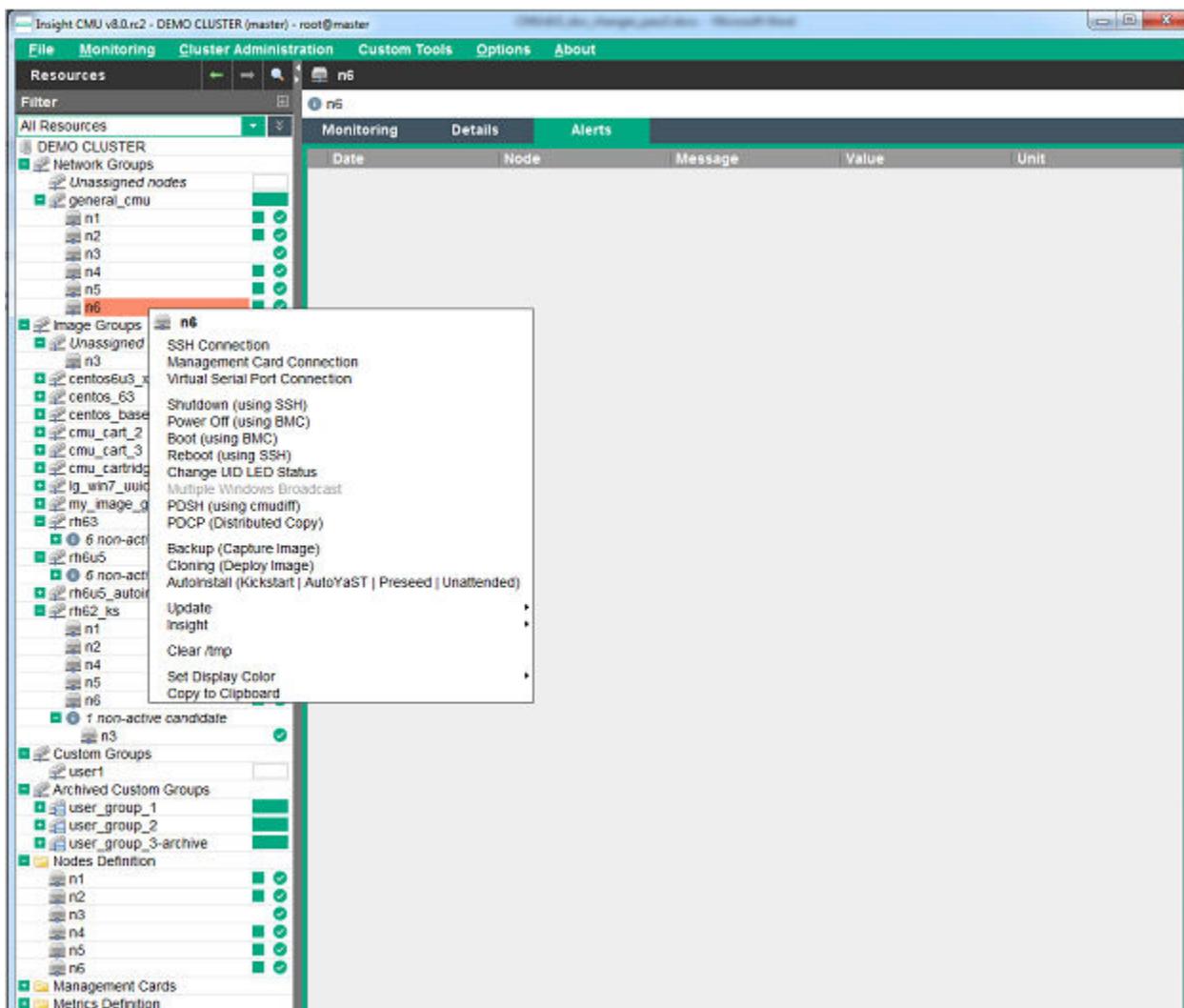
In normal mode, the GUI allows you to monitor node status and visualize static data. You cannot perform any other action on the cluster nodes in normal mode. This helps prevent unauthorized users from performing actions that can be harmful to a node.

- Administrator mode.

In administrator mode, you can perform actions on the cluster nodes.

With one or more nodes selected in the left panel, right-click to access a contextual menu. This menu allows you to perform actions on selected nodes.

The contextual menu is available in network group, image group, and custom group views. This menu is also accessible by right-clicking in the overview frame.



**Figure 38: Contextual menu for administrator mode**

For information about entering administrator mode, see the following:

[Launching the GUI](#) on page 14

## SSH connection

This option launches a secure shell session to the selected node with `ssh`.

---

**NOTE:** This option is available when only one node is selected.

## Changing the default gateway IP address of a node from the GUI

### Procedure

1. Click **Options > Enter Admin mode** and enter the cluster credentials.
2. Click **Cluster Administration > Node Management**.
3. In the left pane, select a node.

4. Right-click the node and select **Manage > Modify Node**.
5. In the **Modify Node Dialog** box, in the **Default Gateway IP Address** field, enter one of the following values and click **OK**:
  - The keyword `default`.
  - The keyword `cmumgt`.
  - The actual, numeric IP address of the gateway.

## Management card connection

This option launches a telnet or secure shell connection to the management card (also known as the baseboard management controller (BMC)) of the selected node. The management card (BMC) must be properly configured when the cluster manager is installed. If the node does not have a management card (BMC), this option is unavailable.

**NOTE:** This option is available when only one node is selected.

## Virtual serial port connection

This option launches a secure shell session to the management card (also known as the baseboard management controller (BMC)) of the selected node. Then, it automatically issues the appropriate command to open a virtual serial port on the management card (BMC).

**NOTE:** This option is available when only one node is selected.

## Shutdown

This option allows you to issue the shutdown command on the selected nodes. The shutdown command can be performed immediately or delayed for a specified time between 1 to 60 minutes. Enter text in the **Message** field to send a message to the users on the selected nodes.



**Figure 39: Halt dialog**

- ① **IMPORTANT:** Several ProLiant and SMP servers do not support HPE APM. If the nodes are not linked to a management card (also known as a baseboard management controller (BMC)), then do not use the shutdown command. Otherwise, the nodes might hang and require a manual shutdown. Instead, use the `reboot` command.

The cluster manager shuts down the selected nodes by using `rsh` or `ssh`. On the compute node, permission must be given to perform commands as superuser or root from the admin node. Otherwise, the shutdown command might not work properly.

## Power off

The **Power off (using BMC)** option allows you to power off the nodes that have a management card (also known as a baseboard management controller (BMC)). The selected nodes must have the same management card (BMC) password.

- (!) **IMPORTANT:** Use the shutdown command before powering off. Otherwise, the file systems might be damaged.



Figure 40: Power off dialog box

## Boot

The **Boot (using BMC)** option allows you to boot a collection of nodes on their own local disk or over the network. You must select the nodes to boot prior to running this command.

The boot procedure uses the management card (also known as the baseboard management controller (BMC)) of each node. The selected nodes must have the same management card (BMC) password. Enter the password for the management card (BMC) to boot the nodes.

- (!) **IMPORTANT:** If the nodes are booted, the boot procedure attempts a proper shutdown. If the shutdown fails, then the management card (BMC) resets. This can damage the file system. To avoid this risk, perform a shutdown operation before issuing a boot command.

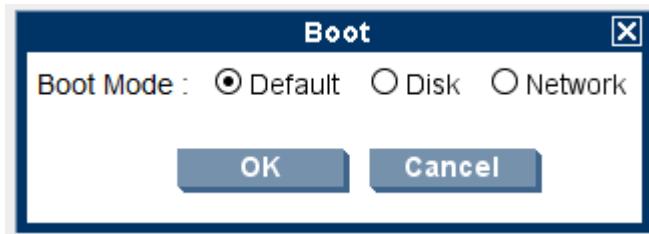


Figure 41: Boot dialog box

## Reboot

The **Reboot (using SSH)** option allows you to issue the `reboot` command on the selected nodes. You can specify whether to run the `reboot` command immediately or to delay the process for a specified time between 1 to 60 minutes. Enter text in the **Message** field to send a message to the users logged onto the selected nodes.

- (!) **IMPORTANT:** The `reboot` command is performed on nodes using `rsh` or `ssh`. On the compute node, permission must be given from the admin node to perform commands as the superuser or root. Otherwise, the `reboot` command might not work properly.



Figure 42: Reboot dialog box

## Change UID LED status

The **Change UID LED status** option changes the status of the locator LED on the selected nodes. If the switch is on, it turns on the LED on selected nodes.

**NOTE:** This option is available only if the node has an iLO management card (also known as a baseboard management controller (BMC)) properly registered in the cluster database and the system is equipped with a status LED.

## Multiple windows broadcast

The **Multiple windows broadcast** option launches a master console window and concurrent mirrored secure shell sessions embedded in a cluster manager terminal window on all selected nodes. All input entered on the master console window is broadcast to the secure shell sessions on the selected nodes, so a command only has to be entered once. To issue commands to a specific node, enter the input directly in the selected cluster manager terminal window for that node.

The following connections are available for multiple windows broadcast:

- A secure shell connection through the network when the network is up on selected nodes.
- Connection through the management card (also known as the baseboard management controller (BMC)), if selected nodes have a management card (BMC).
- Connection to the virtual serial port through the management card (BMC).

To improve the cluster manager terminal window display appearance, every window can be shifted in *x* and *y* from the previous one to fit on the screen. By default, the shift values are computed so the windows tile the screen and no window appears outside of the screen.

To paste the content of your clipboard in all terminals, click **Paste** in the master console.

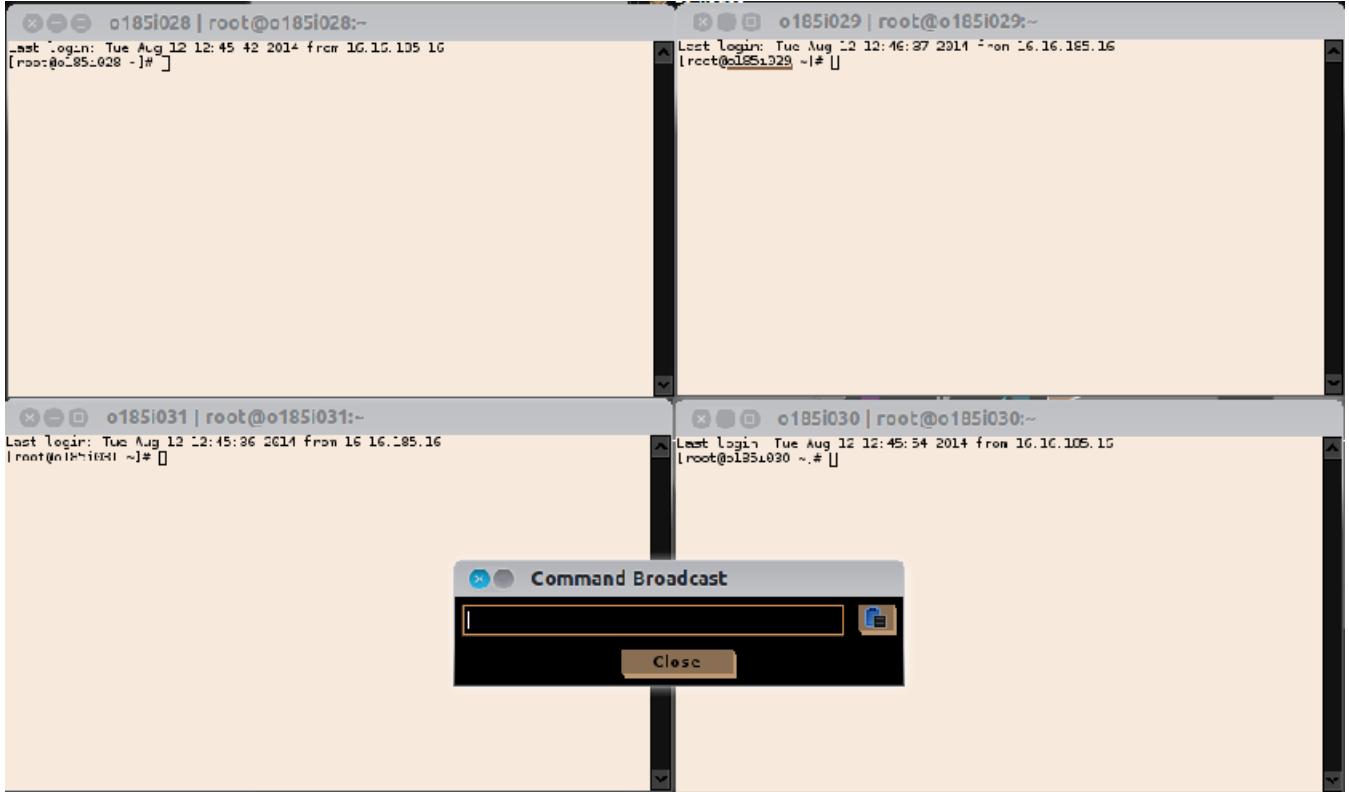


Figure 43: Multiple windows broadcast command

**NOTE:** Hewlett Packard Enterprise recommends limiting the multiple window command to 64 nodes at the same time.

## PDSH (using cmdiff) - single window pdsh

The **PDSH (using cmdiff)** option issues the `pdsh` command, which uses a single terminal to run commands on several nodes in parallel, simultaneously. The output of `pdsh` can be piped to a filter, which formats the output from all of the selected nodes in a synthetic form that omits repetitions of identical results. You can choose among three filtering options:

- `cmudiff [interactive]` (default)
- `cmudiff [non-interactive]`
- `dshbak`

```
o185i[024-043]
-----
cmudiff filter is <ON>
cmudiff comprehensive guide: type 'cmudiff [option]' to change cmudiff behaviour

--help          -h      : display complete cmudiff help
--reference    -r <name> : specify the reference node
--reorder       -R      : reorder lines in files to reduce differences
--max-population -M <value> : max population displayed by percentage of nodes
--replay        -P      : reprocess data from last command

type 'cmudiff' or 'dshbak' to toggle those filters on and off
anything else is a command passed to pdsh
cmu_pdsh> █
cmu_pdsh>
```

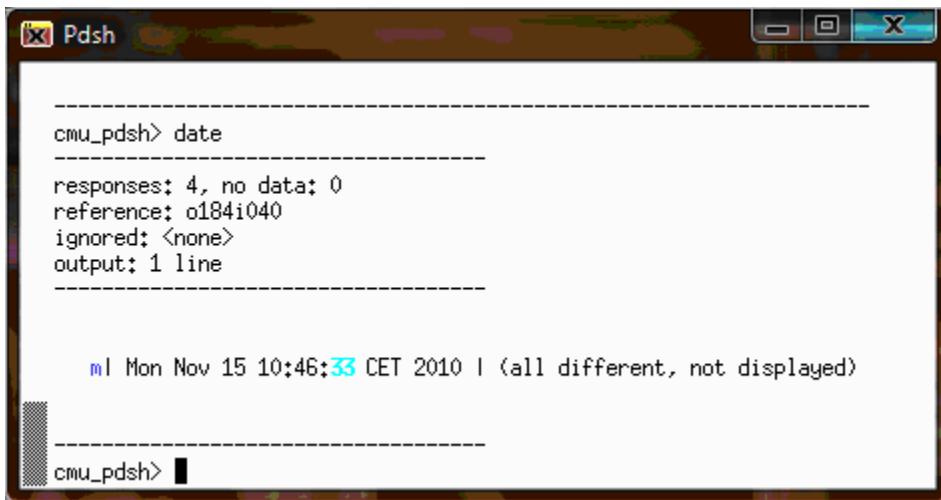
**Figure 44: pdsh window**

Use `dshbak` or `cmudiff` to toggle the filters on and off. These two filters are mutually exclusive, so you can perform the following:

- Filter with `cmudiff`
- Filter with `dshbak`
- Use no filter

#### Example 1: date command

The `cmudiff` output is two fields separated by dotted lines.



A screenshot of a terminal window titled "Pdsh". The window contains the following text:

```
-----  
cmu_pdsh> date  
-----  
responses: 4, no data: 0  
reference: o184i040  
ignored: <none>  
output: 1 line  
-----  
  
| Mon Nov 15 10:46:33 CET 2010 | (all different, not displayed)  
-----  
cmu_pdsh> █
```

The header displays the following information.

- The number of responses

In the example above, the number of responses is 4. This means that a response has been received from 4 compute nodes.

- The reference node

This is the node chosen by `cmudiff` as a reference. Differences in the output from this reference node are highlighted.

- The number of ignored lines
- The number of output lines

The output line displays below the header. In the figure above, the output is only 1 line. The `m` on the left indicates that the output from some compute node differs from the reference node. Some details about the output processing results display on the right.

Characters that differ from the reference node are highlighted in red. In the example above, the time drift in the `seconds` field differs.

Depending on the output length, the output of `cmudiff` can be piped to the `less` editor to enable scrolling through the output with arrows. Output editing is terminated by entering `q`.

#### Example 2: `dmidecode` command

This example uses `cmudiff` to detect BIOS firmware differences with the `dmidecode` command.

```
cmu_pdsh>dmidecode
```

```
responses: 5, no data: 0
reference: 0185i032
ignored: <none>
output: 824 lines
[[ use directional arrows to navigate, press 'q' to return ]]

| # dmidecode 2.9
| SMBIOS 2.6 present.
| 62 structures occupying 3513 bytes.
| Table at 0x000FCBD0.
|
| Handle 0x0000, DMI type 0, 24 bytes
| BIOS Information
|   Vendor: HP
|   Version: 034
|   Release Date: 11/25/2009
|   Address: 0xE0000
|   Runtime Size: 128 kB
|   ROM Size: 2048 kB
|   Characteristics:
|     ISA is supported
|     PCI is supported
(3 populations, not displayed)
(2 populations, not displayed)
(2 populations, not displayed)
/opt/cmu/tmp/cmu_diff_ZDXiFx/cmudiff_result lines 1-23/832 2%
```

**NOTE:** The window only displays a small portion of the output. Use the arrows to scroll up and down.

A difference is found in the BIOS release date. The comment `2 populations, not displayed` on the right suggests that two groups of nodes are present with two different BIOS release dates. One of the two populations might be a single node without a firmware upgrade.

To display the full list of `cmudiff` options, run `cmu_pdsh>cmudiff -h`.

```

cmu_pdsh> cmudiff -h
cmu_diff [hr:voR::dpm:M:i:c:Hfb:s:n] < data
--help           -h          : display this help
--reference     -r <name>   : the reference object
--verbose        -v          : verbose level (0 few feedback,1 process feedback, 2 all feedback)
--only-diff      -o          : only display lines with differences
--reorder        -R <window> : reorder lines in files to increase matching with reference file [not set by default since cmu version 7.0 ]
                                         : commands output such as 'dmidecode' (strict ordered format) can be analysed with -R0
                                         : while 'lsmod' usually gives better results with -R10 or -R50 [default is 10]
--expand-diff    -d          : display population of different lines
--no-perf         -p          : avoid printing perf and variable stuff (for regression suite purposes)
--match-threshold -m <value> : threshold (in %) over which lines are considered similar
--max-population -M <value> : maximum number of population displayed by block (with -d), in percentage of objects
--ignore          -i <name>   : ignore <name>
--cut             -c <value>   : cut output after <value> chars
--no-header       -H          : don't display the header
--no-footer       -F          : don't display the footer
--no-pager        -b          : do not put long outputs in less, just cat the content to stdout
--replay          -P          : get data from file instead of stdin
cmu_pdsh>
  --no-color      -n          : printing without colors
  --diff-color    <value>   : color code triplet of chars with differences (default: bold red on black = 1;3
3;40)
  --pop-color     <value>   : color code triplet of displayed population (default: blue = 34)
  --pop-diff-color <value>  : color code triplet of chars with differences in displayed population (default: red = 33)
cmu_pdsh>

```

Use the **-d** option to narrow the search to the failing nodes and display node populations.

```

cmu_pdsh> cmudiff -d
cmudiff filter is <ON>, with parameters -d
cmu_pdsh>
cmu_pdsh> dmidecode

```

```

responses: 5, no data: 0
reference: o185i032
ignored: <none>
output: 824 lines
[[ use directional arrows to navigate, press 'q' to return ]]

| # dmidecode 2.9
| SMBIOS 2.6 present.
| 62 structures occupying 3513 bytes.
| Table at 0x000FCB0.
|
| Handle 0x0000, DMI type 0, 24 bytes
| BIOS Information
|   Vendor: HP
|   Version: 034
|   Release Date: 11/25/2009
|   Address: 0xE0000
|   Runtime Size: 128 kB
|   ROM Size: 2048 kB
|   Characteristics:
|     ISA is supported
|     PCI is supported
(3 populations) o185i[039,043] are 9
(2 populations) o185i[040,042] are 9
(2 populations) o185i[040,042] are 8

```

/opt/cmu/tmp/cmudiff\_QT7JMD/cmudiff\_result lines 1-23/832 2%

The comment now says (2 populations) o185i[040,042] are 83% similar. This comment suggest that those two compute nodes have a different BIOS release date than all other nodes.

---

**NOTE:** An unresponsive node causes the answer from other nodes to be delayed until a timeout occurs from the unresponsive node. You can reduce this delay by setting the value in the `ConnectTimeout` in `.ssh/config` variable.

For example:

```
# vi /root/.ssh/config
Host *
StrictHostKeyChecking no
ConnectTimeout 1
```

---

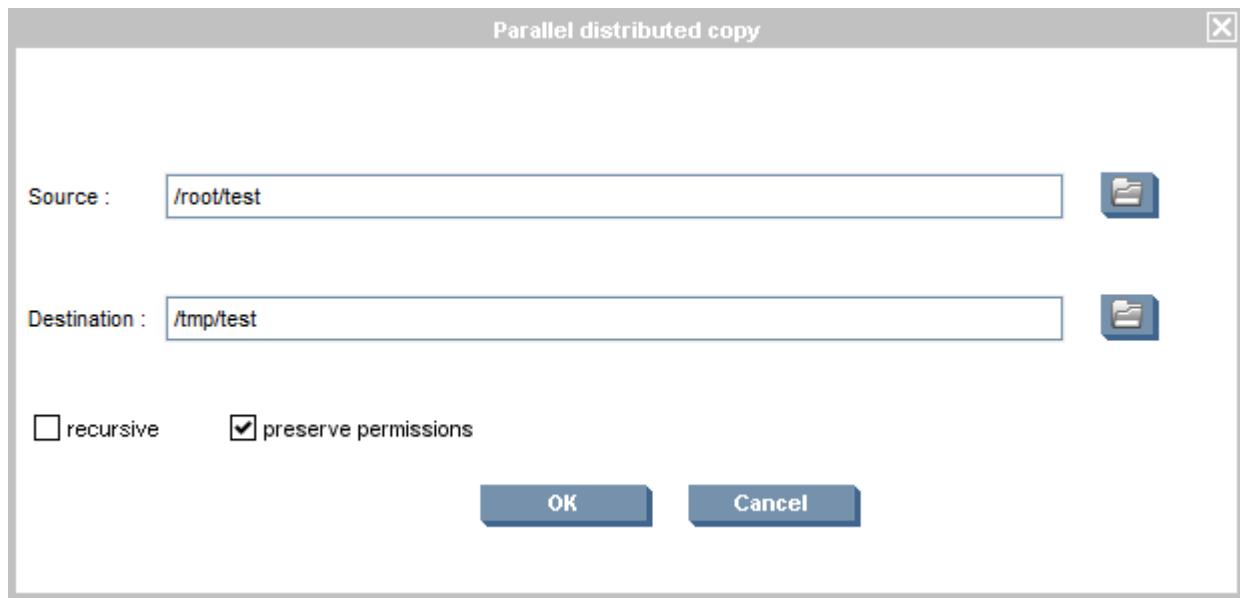
## PDCP (Distributed Copy) - parallel distributed copy

The **PDCP (Distributed copy)** option enables you to copy a file from the admin node to multiple nodes, simultaneously.

To copy the file:

1. Right-click the nodes on which to distribute the copy.
2. On the contextual menu, select **pdcp (distributed copy)**.

The following window displays.



**Figure 45: Parallel distributed copy window**

3. Make entries in the **Source** and **Destination** fields, and click **OK** to execute the distributed copy.

## Custom group management

A **custom group** is a set of nodes named by the cluster manager administrator. Each node can belong to several custom groups. Custom groups are not required for capture and deploy image operations. However, you can use the **Custom Group Management** window to add, delete, or rename a custom group. To perform tasks using the **Custom Group Management** option, click **Cluster Administration > Custom Group Management**.

## Adding custom groups

### Procedure

1. In the **Custom Group Management** window, click **Create**.

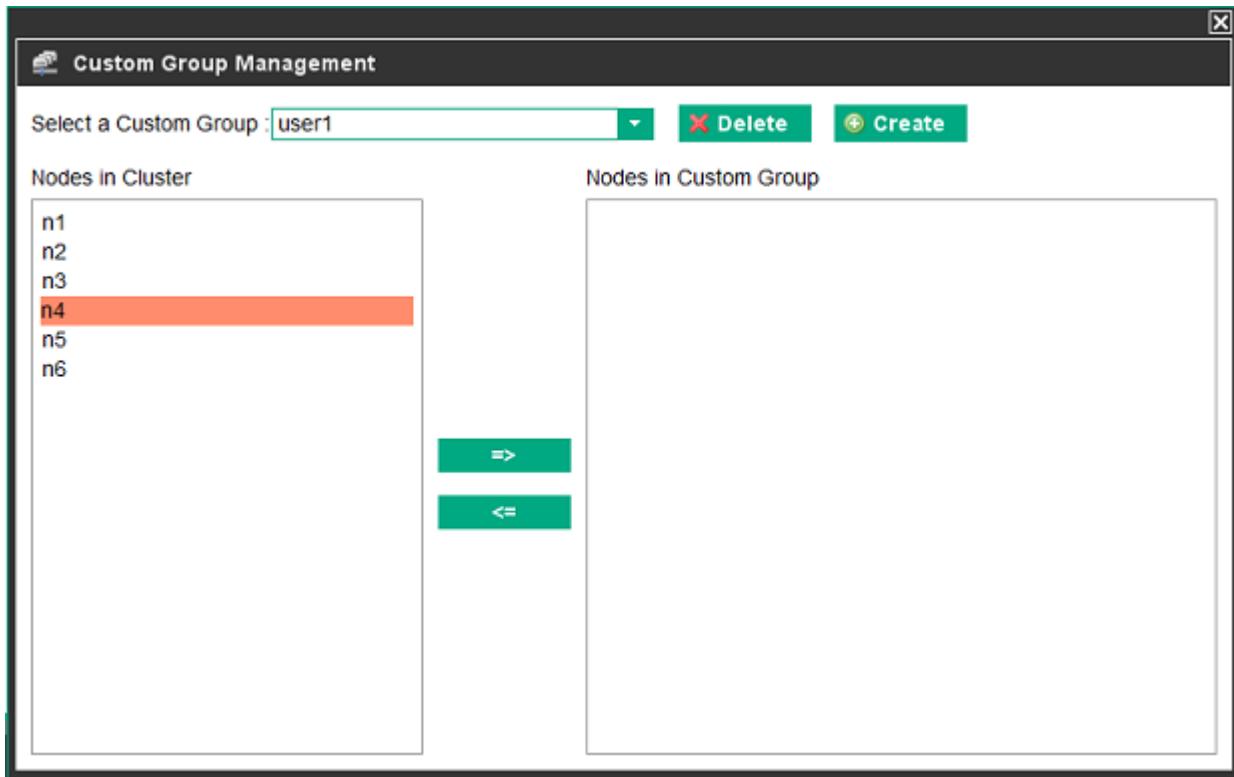


Figure 46: Custom group management

2. Enter a name for the custom group.
3. Click **OK**.

After the group is created, you can add nodes to the group. Select any number of nodes from the **Nodes in Cluster** section and use the arrows to move the nodes to the **Nodes in Custom Group** section.

## Deleting custom groups

### Procedure

1. In the **Custom Group Management** window, select the custom group to delete.
2. Click **Delete**.
3. Click **OK**.

## Firmware management

The cluster manager provides support for managing your firmware. You can view and compare BIOS settings and BIOS firmware versions across a set of chosen nodes. This functionality helps you confirm that your cluster is configured correctly and consistently. You can also run a firmware executable on a set

of nodes to upgrade your firmware to the latest version available. This ensures that your cluster hardware is performing efficiently and consistently. This firmware executable can be an online flash component, or a firmware RPM.

## Viewing and analyzing BIOS settings

Use the `conrep` tool to extract BIOS settings from each node to a file. The `conrep` tool is freely available from the [Hewlett Packard Enterprise Support Center](#) website. It can be found separately or packaged within the Smart-Start Scripting Tool Kit (SSSTK).

The `conrep` tool requires an XML file containing the information necessary to interpret the BIOS flash memory data on your server into human-readable text. The cluster manager is pre-configured with the most common XML file, but depending on your server type, this common XML file might not be compatible with your servers. If your servers require a special XML file, configure the `CMU_BIOS_SETTINGS_FILE` variable in `/opt/clmgr/etc/cmuserver.conf` to the full path and file name of the correct XML file.

The BIOS settings are extracted to a local file on each node. Display the contents of those files using the `cmu_dsh` command with the `CMU_Diff` filter. This allows the user to identify different settings across the set of chosen nodes.

The following items are located in `/opt/clmgr/tmp/conrep/` on each selected node:

- A copy of the `conrep` binary
- A copy of the `conrep` XML file
- The file containing the BIOS settings

The cluster manager provides the latest `conrep` kit available at release time. If a different version of `conrep` is required for the servers in your cluster, complete the procedure in this topic.

1. Download the appropriate version of `conrep` for your environment from the [Hewlett Packard Enterprise Support Center](#) website.
2. Copy the `conrep` binary and the `conrep.xml` file to a location on the admin node.
3. Configure the full path and file name of the new `conrep` binary.
  - a. Edit the `CMU_BIOS_SETTINGS_TOOL` variable in `/opt/clmgr/etc/cmuserver.conf` to point to the location of this new `conrep` binary.
  - b. Change the `CMU_BIOS_SETTINGS_FILE` variable to point to the location of the new `conrep.xml` file.

4. Select one or more nodes in the GUI.

5. Click **Show BIOS Settings**.

If an error occurs, additional software might be required on your compute nodes by the `conrep` binary.

- a. Log into one of the selected nodes.
- b. Change the directory to `/opt/clmgr/tmp/conrep`.
- c. To identify the missing library, run `conrep -h`.

When the `conrep -h` command runs correctly, then the **Show BIOS Settings** feature is enabled on this server.

## Checking BIOS versions

To check the BIOS version on sets of selected nodes, the cluster manager extracts the BIOS Vendor, Version, and Release Date fields from the output of `dmidecode` and concatenates them with hyphens to form a single string. These strings are aggregated with the `cmu_dsh` command and filtered using `dsh_bak` to provide a condensed display of the sets of nodes running common BIOS versions.

## Installing and upgrading firmware

Most servers provide the ability to upgrade firmware while the server is running by invoking an online ROM flash executable, or firmware executable. These are Linux executables that tests for the correct server type and installs a later version of the firmware. Newer firmware executables for Linux are packaged in an RPM format.

You can obtain this firmware executable or firmware RPM from the [Hewlett Packard Enterprise Support Center](#). Copy it to `/opt/clmgr/firmware/`. Then, you can use the cluster manager to select the nodes to upgrade, and these binaries are copied and executed in parallel. If a firmware RPM is given to the cluster manager, it will be unpacked and the firmware executable is copied to the selected nodes and executed in parallel. By default, the cluster manager executes these binaries with the `-s` option, which tells the binary to run in `script` mode. If necessary, you can change the arguments by editing the `CMU_FIRMWARE_EXECUTABLE_ARGUMENTS` variable in `/opt/clmgr/etc/cmuserver.conf`.

Hewlett Packard Enterprise recommends installing the firmware executable on one node, first, to test the operation. After the binary finishes executing, you must reboot the node for the new firmware to take effect. If this process is successful, then you can use the cluster manager to duplicate this process on a larger set of nodes.

## Requesting firmware versions

You can query the iLO (or the iLO CM on a Moonshot chassis) for a list of current firmware versions. This firmware version information is stored in the **Custom Features** table for each node, and can be viewed from the GUI or with the `/opt/clmgr/bin/cmu_show_features` command in the CLI.

The **Custom Features** table is located below the **Features** table. This table contains:

- The various firmware versions beginning with `[Firmware]`
- The node static data that was retrieved when the monitoring client was installed

---

**NOTE:** This feature is only available on servers with iLO4 or newer, and on Moonshot chassis with a recent version of the iLO CM firmware that supports Rest API queries.

---

### Procedure

1. Select a set of nodes from the left-panel tree.
2. Right-click the nodes and select **Update > Update Firmware Inventory (HPE Gen8+ hardware only)**.
3. When the command finishes, select a node from the left panel tree and click **Details** to view the firmware versions.

To request the firmware versions from the CLI, use the `/opt/clmgr/tools/cmu_get_fw_versions` command.

For example

```
# /opt/clmgr/tools/cmu_get_fw_versions -h
usage : /opt/clmgr/tools/cmu_get_fw_versions -h
        /opt/clmgr/tools/cmu_get_fw_versions -c <ILOCM IP address>
        /opt/clmgr/tools/cmu_get_fw_versions -n <nodename> | <noderange>
        /opt/clmgr/tools/cmu_get_fw_versions -n <nodename> -i <iLO IP address>
        /opt/clmgr/tools/cmu_get_fw_versions -f <nodefile>
#
```

To view firmware versions from the CLI, use the `/opt/clmgr/bin/cmu_show_features` command.

For example:

```
# /opt/clmgr/bin/cmu_show_features -n node35
# /opt/clmgr/tools/cmu_get_fw_versions -n node35
new feature(s) successfully added
# /opt/clmgr/bin/cmu_show_features -n node35
[Firmware] HP ProLiant System ROM = 06/02/2012
[Firmware] HP ProLiant System ROM - Capture = 04/04/2012
[Firmware] HP ProLiant System ROM Bootblock = 08/30/2011
[Firmware] SL-Chassis Firmware = 5.3
[Firmware] SL-Chassis Firmware Bootloader = 1.6
[Firmware] Server Platform Services (SPS) Firmware = 2.1.5.2B.4
[Firmware] System Programmable Logic Device = Version 0x2F
[Firmware] iLO = 2.03 Nov 07 2014
#
```

If you upgrade any firmware after querying the firmware versions, rerun the query command to refresh the **Custom Features** table with the latest version information.

## Saving user settings

You can save and restore preferences locally in the `cmu_gui_local_settings` file without cluster administrator privileges.

## CLI actions

### Starting a CLI interactive session

You can invoke the `cmucli` command at any time to start a CLI session in interactive mode.

```
# cmucli
```

### Basic commands

The following commands help you manage the cluster in command-line mode. The command-line mode is used interactively or in a script mode, giving a file name when invoking the command. Each command is executed on a specified set of nodes. Use a complete list of nodes, or a regular expression to specify the nodes. This interface includes administration, capture image, and deployment features.

- `cmu> exit`

Use this command to exit a CLI interactive session.

- **# cmucli *my\_path/my-file***

Use this command to start a noninteractive CLI session and execute commands from a file.

---

**NOTE:** The file must contain a set of valid cluster manager commands. The available commands and syntax are described in the following sections.

---

- **cmu> help**

Use this command to get help during a CLI session. This command displays all available commands in the CLI.

HELP COMMANDS

```
help administration | help database
help configuration | help other
or help <command name>
```

DATABASE COMMANDS

```
group | node
image_group | network_group
custom_group
```

ADMINISTRATION COMMANDS

```
boot | broadcast
halt | shutdown
locate | power
reboot | modify_password
capture | deploy
change_kernel | kickstart | autoinstall
pdcp | custom run
```

CONFIGURATION COMMANDS

```
add_node | add_{ks|ai}_image_group
add_image_group | add_network_group
add_custom_group | delete_node
delete_image_group | delete_network_group
delete_custom_group | add_to_image_group
add_to_network_group | add_to_custom_group
del_from_image_group | del_from_network_group
del_from_custom_group | change_active_image_group
probe_kernel | scan_macs
```

OTHER COMMANDS

```
Exit | cat
Date | echo
sh | vi
#
```

- **# help *command\_name***

Use this command to get help for a specific command. This command displays detailed information and the exact syntax of the command.

For example, to get more information about the `halt` command:

```

cmu> help halt
Delay can be anything in "now, 1, 5, 10, 15, 30, 60" minutes
    halt delay "mesg" all
        halt all nodes
    halt delay "mesg" node_1
        halt node node_1      halt delay "mesg" node_1 node_2
        halt nodes  node_1 node_2
    halt delay "mesg" node_1 - node_n
        halt all nodes between node_1 and node_n
    halt delay "mesg" node_*
        halt all nodes starting with node_ word
    halt delay "mesg" allbut node_1
        halt all nodes except node_1
    halt delay "mesg" all group_1
        halt nodes of image group group_1
    halt delay "mesg" all group_1 but node_exp
        halt nodes of image group group_1 except node_exp
    halt delay "mesg" all group_1 group_2
        halt nodes of group_1 and group_2
cmu>

```

- **cmu> groups**

Use this command to display the list of image groups in a cluster.

```
cmu> groups
```

```
list of group(s) with active nodes : default rhel7.4
```

```
list of available group(s) for image capture and deploy : ice-rhel7.4 lead-rhel7.4 rhel7.4 default
```

- **cmu> groups *group\_name***

Use this command to view the nodes in an image group and the attributes.

For example:

```
cmu>groups default
selected: o184i098 { 1 node }
```

- **cmu> nodes**

Use this command to display the list of nodes with the attribute name.

```
cmu> nodes
Machines.node.n1 = 192.168.1.1,255.255.255.0,00-19-
BB-3A-8A-60,rh6u7,192.168.1.101,ILO,x86_64,-1,-1,generic,default,default,de
fault,default,auto,default,none

Machines.node.n2 = 192.168.1.2,255.255.255.0,00-19-BB-3A-
A8-64,rh6u7,192.168.1.102,ILO,x86_64,-1,-1,generic,default,default,default,
default,auto,default,none

Machines.node.n3 = 192.168.1.3,255.255.255.0,00-1A-4B-DE-19-54,default,
192.168.1.103,ILO,x86_64,-1,-1,generic,default,default,default,auto
,default,default,none
cmu>
```

You can also use a subset of this command. The next section describes how to specify subset of nodes in the CLI.

### Management card (also known as baseboard management controller (BMC)) naming

Management card (BMC) naming differs between the GUI and the CLI. In the GUI, iLO displays as such. In the CLI, iLO displays as ILO.

## Specifying nodes

Using the CLI, you can execute a command on any number of nodes.

The `nodes` command:

- Displays node information
- Tests regular expressions for selecting nodes before executing a command

### Executing a command on one node

To execute a command on only one node, specify the name of the node. The following command executes on node o185i222:

```
cmu> node o185i222  
active node list selected: o185i222  
cmu>
```

### Executing a command on a list of nodes

To execute a command on multiple nodes, specify the names of each node.

```
cmu> boot o185i222 o185i233 o185i243  
active node list selected: o185i222 o185i233 o185i243  
cmu>
```

### Executing a command on a range of nodes

To execute a command on a range of nodes, specify the range using their attributes. Commands are executed on all nodes within the range.

```
cmu> boot o185i222 - o185i225  
active node list selected: o185i222 o185i223 o185i224 o185i225  
cmu>
```

---

**!** **IMPORTANT:** Spaces between the dash character (-) and node names are mandatory.

---

### Using wildcards

You can use the asterisk character (\*) as a wildcard to select all nodes with matching node names. For example:

```
cmu> nodes o185i22*  
active node list selected: o185i220 o185i221 o185i222 o185i223 o185i224 o185i225 o185i226 o185i227 o185i228 o185i229  
cmu>
```

### Complex list of nodes

A complex list of nodes can be specified using any combination of the above regular expressions.

---

**NOTE:** If a node is mentioned twice, the command is executed twice on the node.

---

To boot specific nodes:

```
cmu> boot o185i209 o185i217 o185i22* o185i232 - o185i235
active node list selected: o185i209 o185i217 o185i220 o185i221 o185i222
o185i223 o185i224 o185i225 o185i226 o185i227 o185i228 o185i229 o185i232
o185i233 o185i234 o185i235
cmu>
```

### Executing a command on all nodes

A command followed by the `all` option is executed on all nodes.

```
cmu> boot all
active node list selected: o185i192 o185i193 o185i194 o185i195 o185i196
o185i197 o185i198 o185i199 o185i200 o185i201 o185i202 o185i203 o185i204
o185i205 o185i206 o185i207 o185i208 o185i209 o185i210 o185i211 o185i212
o185i213 o185i214 o185i215 o185i216 o185i217 o185i218 o185i219 o185i220
o185i221 o185i222 o185i223 o185i224 o185i225 o185i226 o185i227 o185i228
o185i229 o185i230 o185i231 o185i232 o185i233 o185i234 o185i235 o185i236
o185i237 o185i238 o185i239 o185i240 o185i241 o185i242 o185i243 o185i244
o185i245 o185i246 o185i247 o185i248 o185i249 o185i250 o185i251 o185i252
o185i253 o185i254 o185i255
cmu>
```

### Excluding specific nodes from a command

Use the `allbut` option to select all nodes except specific ones. The `allbut` option can be followed by any combination of the above regular expressions.

```
cmu> boot allbut o185i2*
active node list selected: o185i192 o185i193 o185i194 o185i195 o185i196 o185i197 o185i198 o185i199
cmu>
```

### Executing a command on all nodes of an image group

Use the `all` option followed by a group name to select all active nodes of this group.

```
cmu> boot all default
active node list selected: o185i194 o185i202 o185i216 o185i222 o185i233 o185i243 o185i252 o185i253 o185i254
cmu>
```

### Executing a command on specific nodes of an image group

Use the `but` option to exclude active nodes of a group from the selection. Nodes to exclude can be specified with any combination of regular expressions.

```
cmu> boot all default but o185i222 - o185i252
active node list selected: o185i194 o185i202 o185i216 o185i253 o185i254
cmu>
```

## Administration and deployment commands

### Booting a set of nodes

You can boot any number of nodes in the cluster. Regular expressions to specify a list of nodes are accepted.

```
cmu> boot [net] cmu_nodes_regular_expression
```

The following boot modes are available:

- Normal mode boots nodes on the default device, generally the hard drive.
- Network mode fills the `dhcpd.conf` file to boot nodes over the network.

**Booting node o185i192 in normal mode on the hard drive:**

```
cmu> boot o185i192
active node list selected: o185i192
Entering normal Boot
Booting node o185i192
Not using ssh/rsh shutdown because node o185i192 is not pinging
Doing a powercycle of node o185i192 via lo100i (OFF then ON)
Boot order has been sent to every node
      Boot process finished !!
Please read /opt/clmgr/log/Boot-lo100i-Tue-08-Aug-at-12h36m23s.log to check errors
cmu>
```

**Booting node o185i192 in network mode:**

```
cmu> boot net o185i192
active node list selected: o185i192

Cleaning previous kickstart boot option
Copying /etc/hosts to /opt/clmgr/ntbt/rp/etc/
Entering Network boot
Writing dhcpd.conf
dhcpd.conf written successfully

Booting node o185i192
Not using ssh/rsh shutdown because node o185i192 is not pinging
Doing a powercycle of node o185i192 via lo100i (OFF then ON)
Waiting 1 second (cooldown)...

Waiting 5 minutes before removing the dhcpd.conf

      Boot process finished !!

Please read /opt/clmgr/log/Boot-lo100i-Tue-08-Aug-at-12h40m36s.log to check errors
cmu>
```

### Broadcasting commands to a set of nodes

The `broadcast` command launches the interactive single window broadcast utility. You can use the regular expressions described above to specify nodes. The command uses telnet and secure shell.

---

**NOTE:** This command has limited functionality. For more information on this command, see the following:

#### **Multiple windows broadcast** on page 96

---

```
cmu> broadcast cmu_nodes_regular_expression
```

To broadcast on all nodes of the cluster:

```
cmu> broadcast all
```

```
selected nodes: o185i192 o185i193 o185i194 o185i195 o185i196 o185i197
o185i198 o185i199 o185i200 o185i201 o185i202 o185i203 o185i204 o185i205
o185i206 o185i207 o185i208 o185i209 o185i210 o185i211 o185i212 o185i213
o185i214 o185i215 o185i216 o185i217 o185i218 o185i219 o185i220 o185i221
o185i222 o185i223 o185i224 o185i225 o185i226 o185i227 o185i228 o185i229
o185i230 o185i231 o185i232 o185i233 o185i234 o185i235 o185i236 o185i237
```

```
o185i238 o185i239 o185i240 o185i241 o185i242 o185i243 o185i244 o185i245  
o185i246 o185i247 o185i248 o185i249 o185i250 o185i251 o185i252 o185i253  
o185i254 o185i255
```

```
CMU pdsh interface  
dshbak filter is <ON>  
help|h|? to get help  
cmu_pdsh>
```

### Rebooting a set of nodes

The `reboot` command allows you to reboot any number of nodes in the cluster. You can use the regular expressions previously described.

```
cmu> reboot delay "message" cmu_nodes_regular_expression
```

For `delay`, specify the time period after which the specified nodes are rebooted. Specify one of the following:

- now
- 1
- 5
- 10
- 15
- 30
- 60

In the preceding list, the numeric values specify a number of minutes.

For `message`, specify a string to appear on each rebooted node. For example:

```
cmu> reboot 1 "Reboot for maintenance" o185i192  
active node list selected: o185i192
```

```
Rebooting the nodes  
..Return-->Command too long to execute<--:wait_for_command:  
.Please read /opt/clmgr/log/Reboot.log to check errors  
cmu>
```

### Halting a set of nodes

The `halt` command halts any number of nodes in the cluster. You can use the regular expressions previously described.

```
cmu> halt delay "message" cmu_nodes_regular_expression
```

For `delay`, specify the time period after which the specified nodes are halted. Specify one of the following:

- now
- 1
- 5
- 10
- 15

- 30
- 60

In the preceding list, the numeric values specify a number of minutes.

For *message*, specify a string to appear on each halted node. For example:

```
cmu> halt now "Halt for maintenance" o185i192
active node list selected: o185i192
```

```
Halting the nodes
..Please read /opt/clmgr/log/Halt.log to check errors
cmu>
```

### **Powering off a set of nodes**

The `power off` command powers off any number of nodes in the cluster. You can use the regular expressions previously described.

```
cmu> power off cmu_nodes_regular_expression
```

For example:

```
cmu> power off o185i192
active node list selected: o185i192

Please read /opt/clmgr/log/PowerOff.log for errors.
cmu>
```

### **Setting the locator LED on or off**

Use the `locate on|off` command to set the locator LED of any number of nodes on or off. You can use the regular expressions previously described.

```
cmu> locate on|off cmu_nodes_regular_expression
```

For example:

```
cmu> locate on o185i192
active node list selected: o185i192
cmu>

cmu> locate off o185i192
active node list selected: o185i192
cmu>
```

### **Deploying a set of nodes**

Use the `deploy_image` command to deploy an image to a node or a set of nodes. You can use the regular expressions previously described. Successfully deployed nodes are active in the image group associated with the image. Failed nodes are active in the default image group .

```
cmu> deploy_image "image_name" cmu_nodes_regular_expression
```

For example:

```
cmu> deploy_image "cluster" o185i195
active node list selected: o185i195

node list found in capture group cluster: o185i195
Save /opt/clmgr/log/cmucerbere.log file
Cleaning boot directory
Configuring the system
```

```

[INFO] CMU does not seem to be running
Copying the ssh settings
Rebuilding network-boot image
/opt/clmgr/tmp/GUI/config.txt was rewritten
Start deploying
Every 2.0s: /opt/clmgr/tools/logAnalyser.sh

Deploying started on 2006-08-08 at [17:07:55]
+-----+-----+
| NET BOOTING | |
+-----+-----+
| PARTITION & FORMAT DISKS | |
+-----+-----+
| GETTING DATA | |
+-----+-----+
| DEPLOYING ERROR | |
+-----+-----+
| DEPLOYED | o185i195 |
+-----+-----+

```

Deploy process finished on 2006-08-08 at [17:10:51]

Database report:

	deployed	error	unknown
ne1	1	0	0
sfs	0	0	0
ne2	0	0	0
ne3	0	0	0
ne4	0	0	0
test	0	0	0
Total	1	0	0

```

Detailed logs are in /opt/clmgr/log/cmucerbere.log and /opt/clmgr/log/cmucerbere-*.log
[INFO] CMU does not seem to be running
/opt/clmgr/tmp/GUI/config.txt was rewritten
cmu>

```

### Creating an image group

The `add_image_group` command creates an image group. Parameters are specified on one line.

```
cmu> add_image_group groupname "device"
```

For example:

```
cmu> add_image_group my_image_group "sda"
processing 1 image group ...
```

### Adding nodes to an image group

The `add_to_image_group` command adds nodes to an image group. Parameters are specified on one line.

```
cmu> add_to_image_group nodes to group_name
```

For example:

```
cmu> add_to_image_group o184i115 - o184i116 to my_image_group
selected: o184i115 { 2 nodes }
processing 2 nodes...
```

### Modifying the password for a management card (also known as a baseboard management controller (BMC))

The `modify_password` command modifies the management card (BMC) password in the cluster manager database.

```
cmu> modify_password ILO|ILOCM|lo100i
```

For example:

```
cmu> modify_password lo100i  
login> hpe  
password> service  
password successfully changed  
  
cmu>
```

- 
- (!) **IMPORTANT:** You can only change the password contained in the database. This command does not change the actual password of the management card (BMC). The password is echoed during this command.
- 

### Discovering MAC address for new nodes

The `scan_macs` command enables discovering the MAC address for new nodes. Enter parameters interactively.

For example:

```
cmu> scan_macs  
Enter the first nodename (ex. "n%i"): 1  
Enter the nodename prefix: n  
Enter the IP address of the first node: 192.168.1.1  
Enter the netmask: 255.255.0.0  
Enter the BMC type (ex. ILO, lo100i, ILOCM): ILO  
Enter the compute node NIC number attached to the admin network [1]:  
If you have a file of BMC IP addresses to scan enter the filename:  
Enter one or more BMC IPs to scan separated by commas or enter  
a starting IP if you'd like CMU to generate addresses to scan: 192.168.1.101  
Enter the total number of sequential addresses to scan from this IP [1]:  
Do you want to modify other default scanning behaviors? (y/[n]): n  
Discovered nodes are inserted into the CMU database by default.  
Enter a file name if you wish to write the node information to a file  
instead:  
SUMMARY  
First nodename is '1'  
The node prefix is 'n'  
The first IP address is '192.168.1.1'  
The netmask is '255.255.0.0'  
The BMC type is 'ILO'  
The ILO IP address list is '192.168.1.101'  
The compute node NIC attached to the admin network is '1'  
The admin node IP for the scanned nodes is 'default'  
The gateway IP for the scanned nodes is 'default'  
The BIOS boot mode is 'auto'  
The iscsi root string for the scanned nodes is 'none'  
Discovered nodes will be written to the CMU database.  
Is this correct? ([y]/n/q): y  
INFO: It looks like StrictHostKeyChecking is set to 'no' in /root/.ssh/  
config...  
Make sure you can ssh to all client nodes without providing a password or  
answering(yes/no) to a registration question or various CMU commands/systems
```

```
will fail to run.  
Scanning complete. 1 nodes added, 0 nodes updated.
```

## Administration utilities pdcp and pdsh

The cluster manager includes the open source software pdcp and pdsh.

### Example: pdcp

```
# /opt/clmgr/bin/pdcp -w cn0001,cn0002 source /tmp/dest
```

For *source*, specify a file on the admin node.

For *dest*, specify the name of a destination file. In the preceding command example, the destination file is copied to compute nodes `cn0001` and `cn0002`.

### Example: pdsh

```
# /opt/clmgr/bin/pdsh -w cn0001,cn0002 ls
```

```
cn0001: bin  
cn0001: inst-sys  
cn0002: anaconda-ks.cfg  
cn0002: CMU_CLONING_INFO  
cn0002: install.log.syslog  
cn0002: install.log
```

The `ls` command is executed on compute nodes `cn0001` and `cn0002`.

## Linux shell commands

The cluster manager provides a Linux shell API interface. Most functions provided from the GUI and CLI have their equivalent in the API interface. The API interface is easily called from a shell script.

For information about commands, see one of the following:

- [Manpages](#) on page 137
- The manpages themselves

# Advanced topics

## Enabling nonroot user access to commands and GUI actions

The root user has permission to run all cluster manager commands and perform all GUI actions. As the root user, you can configure the cluster manager in a way that enables nonroot users to perform the following actions:

- Run commands
- Log into the GUI and perform some GUI actions

The following topics explain how to configure access to the commands and GUI actions for nonroot users:

- [\*\*Enabling nonroot users to run commands based on pdsh\*\*](#) on page 116
- [\*\*Granting nonroot users the ability to run cluster manager commands\*\*](#) on page 117
- [\*\*Granting nonroot users the ability to use the GUI to manage the cluster\*\*](#) on page 118
- [\*\*Configuring the GUI to work with sudo\*\*](#) on page 120
- [\*\*Commands and GUI actions\*\*](#) on page 120
- [\*\*Customizing the cluster manager commands\*\*](#) on page 123

### Enabling nonroot users to run commands based on pdsh

The root user has access to all the commands in the following directories:

- /opt/clmgr/bin
- /opt/clmgr/tools

Certain commands in the preceding directories function only if the user has appropriate privileges. These commands include:

- pdsh
- pdcp
- rpdcp
- cmu\_diff
- cmu\_dsh (pdsh plus cmu\_diff)

The following procedure explains how to enable access to the pdsh commands for nonroot users.

## Procedure

1. Use your operating system documentation to add passwordless `ssh` access to the selected nonroot users.
2. Set `CMU_DIFF_TMP_DIR=/tmp` in nonroot user environments, as needed.

This step is required for `cmu_diff` and `cmu_dsh`.

## Granting nonroot users the ability to run cluster manager commands

The procedure in this topic explains how to enable nonroot users to run root-level cluster manager commands at the command line.

If you do not complete this procedure, a nonroot user cannot run various cluster manager commands. You can also use the `sudo` command to allow nonroot users to run commands and log incidents to log files.

For more information about editing the `sudoers` file, see your operating system documentation.

## Procedure

1. Use the `visudo` command to open the following file:

`/etc/sudoers`

2. (Optional) Create command aliases for groups of commands.

For example, assume that you want to create the following command groups:

- Power control
- Provisioning
- Remaining features

The following lines declare command groups and include the full path to each cluster command included in the group:

```
Cmnd_Alias CMU_POWER = /opt/clmgr/bin/cmu_halt,/opt/clmgr/bin/cmu_power,/opt/clmgr/bin/cmu_boot  
Cmnd_Alias CMU_IMAGE = /opt/clmgr/bin/cmu_backup,/opt/clmgr/bin/cmu_clone,/opt/clmgr/bin/cmu_autoinstall_node  
Cmnd_Alias CMU_ETC = /opt/clmgr/bin/cmu_console,/opt/clmgr/tools/cmu_cn_install,/opt/clmgr/bin/cmu_firmware_mgmt
```

For information about the full path to each command, see the following:

**Commands and GUI actions** on page 120

3. Determine which nonroot users are allowed to access root-level commands without being prompted for a password.

If you want to grant a nonroot user the ability to run commands, plan to include the `NOPASSWD` keyword in the `sudoers` for that user. A later step in this procedure includes examples that show the `NOPASSWD` keyword.

4. Add lines in the `sudoers` file for each user who needs nonroot access to cluster commands.

For example, the following lines specify permitted user actions and use command aliases from the examples in this procedure:

- The following line grants shutdown and reboot privileges to user `jsmith`:

```
jsmith ALL=(ALL) NOPASSWD: /opt/clmgr/bin/cmu_halt
```

In the preceding example, in the second field, rather than the keyword `ALL`, you can specify `localhost` or the host name of the admin node.

For example:

```
jsmith localhost=(ALL) NOPASSWD: /opt/clmgr/bin/cmu_halt
```

The `NOPASSWD` keyword specifies that user `jsmith` does not have to provide a password to run the `cmu_halt` command.

- The following line grants access to any site-customized GUI actions for user `kjones`:

```
kjones ALL=(ALL) NOPASSWD: /opt/clmgr/etc/cmu_custom_menu/custom_cmd
```

For `custom_cmd`, specify a custom command that you created in the following procedure:

**Customizing the cluster manager commands** on page 123

- The following line prompts `cjones` for a password and grants power control permissions to user `cjones`:

```
cjones ALL=(ALL) CMU_POWER
```

- The following line lets user `bstevens` control power and provisioning without providing a password:

```
bstevens ALL=(ALL) NOPASSWD: CMU_POWER, CMU_IMAGE
```

- The following line lets user `sbarney` run the full list of administrative control commands without a password:

```
sbarney ALL=(ALL) NOPASSWD: CMU_POWER, CMU_IMAGE, CMU_ETC
```

## 5. Verify that `Defaults requiretty` is commented out.

When you comment out this line, you avoid `tty` errors. Make sure that the line appears as follows:

```
# Defaults requiretty
```

## 6. Save and close the `/etc/sudoers` file.

## 7. (Optional) Enable nonroot users to administer the cluster by logging into the GUI.

The preceding steps enabling nonroot users to administer the GUI through commands. To enable nonroot users to administer the cluster by using the GUI, too, proceed to the following:

**Granting nonroot users the ability to use the GUI to manage the cluster** on page 118

## Granting nonroot users the ability to use the GUI to manage the cluster

The following procedure explains how to enable nonroot users to log into the GUI and use the GUI to manage the cluster.

## Prerequisites

[Granting nonroot users the ability to run cluster manager commands](#) on page 117

## Procedure

1. Use a text editor to open the following file:

```
/opt/clmgr/etc/admins
```

Within the file, notice that a pound character (#) in column 1 indicates a comment line.

2. Add a line to the file for each user who needs access to the GUI as a nonroot user.

Use the following format for each line:

```
username [ keyword[ keyword keyword ...]]
```

For *username*, specify the user login name of one user.

For *keyword*, specify zero or more keywords that represent a command or a set of commands. The *keywords* have the following purpose:

- If the *username* is not followed by any keywords, that user has permission to perform all actions from the GUI.
- If the *username* is followed by one or more keywords, that user has permission to perform only the actions that are related to the specified keywords. Use a space to separate individual keywords.

For information about user actions, see the following:

[Commands and GUI actions](#) on page 120

The keywords you can specify are as follows:

- Node management keywords:
  - NODE\_ADD - Permission to add nodes to the cluster
  - NODE MODIFY - Permission to modify current node settings
  - NODE\_DELETE - Permission to delete nodes from the database
- Network group management keywords:
  - NETWORK\_GROUP\_ADD - Permission to add network groups
  - NETWORK\_GROUP MODIFY - Permission to add/delete nodes to/from a network group
  - NETWORK\_GROUP\_DELETE - Permission to delete network groups
- Image group management keywords:

- IMAGE\_GROUP\_ADD - Permission to add image groups
- IMAGE\_GROUP MODIFY - Permission to add/remove nodes to/from image groups
- IMAGE\_GROUP\_DELETE - Permission to delete image groups
- IMAGE\_GROUP\_MAKE\_NODE\_ACTIVE - Permission to change the active image group of a node
- Custom group management keywords:
  - CUSTOM\_GROUP\_ADD - Permission to new custom groups
  - CUSTOM\_GROUP MODIFY - Permission to add/remove nodes to/from custom groups
  - CUSTOM\_GROUP\_DELETE - Permission to delete custom groups

**3.** Save and close the file.

## Configuring the GUI to work with sudo

### Procedure

1. Enter the following command to determine the path to the `sudo` binary on the admin node:

```
# which sudo
```

2. Open the following file in a text editor:

```
/opt/clmgr/etc/cmuserver.conf
```

3. Search in the file for `CMU_SUDO`.

4. Set `CMU_SUDO` to the path to the `sudo` binary on the admin node.

For example set the path to the following:

```
/usr/bin/sudo
```

5. Save and close the file.

6. Enter the following commands to stop and start the `cmu` service:

```
# systemctl stop cmu
# systemctl start cmu
```

7. Close the GUI and restart the GUI.

## Commands and GUI actions

The following tables map GUI features to corresponding commands. The column headings are as follows:

- The column 2 heading, **Underlying command (def: /opt/clmgr/bin/)**, shows the command that runs when a user clicks a specific set of actions in the GUI.
- The column 3 heading, **DB keyword**, shows the privilege that exists in the internal cluster database for a specific GUI action. For example, for **Node Management > Add Node**, no underlying command runs. However, the action requires the `NODE_ADD` privilege in the internal cluster database.

**Table 4: Features and controls**

GUI action (top menu bar)	Underlying command (def: /opt/clmgr/bin/ )	DB keyword
<b>Cluster Administration &gt; Node Management &gt; Add Node</b>		NODE_ADD
<b>Cluster Administration &gt; Node Management &gt; Delete Node</b>		NODE_DELETE
<b>Cluster Administration &gt; Node Management &gt; Modify Node</b>		NODE_MODIFY
<b>Cluster Administration &gt; Node Management &gt; Scan Node</b>	cmu_scan_macs	NODE_ADD
<b>Cluster Administration &gt; Node Management &gt; Change (Active) Image Group</b>		IMAGE_GROUP_MAKE_NODE_ACTIVE
<b>Cluster Administration &gt; Node Management &gt; Import Nodes</b>		NODE_ADD
<b>Cluster Administration &gt; Node Management &gt; Export Nodes</b>		
<b>Cluster Administration &gt; Network Management &gt; Add</b>		NETWORK_ADD
<b>Cluster Administration &gt; Network Management &gt; Delete</b>		NETWORK_DELETE
<b>Cluster Administration &gt; Network Management &gt; Modify</b>		NETWORK_MODIFY
<b>Cluster Administration &gt; Network Group Management &gt; Create</b>		NETWORK_GROUP_ADD
<b>Cluster Administration &gt; Network Group Management &gt; Delete</b>		NETWORK_GROUP_DELETE
<b>Cluster Administration &gt; Network Group Management &gt; Manage Nodes</b>		NETWORK_GROUP MODIFY
<b>Cluster Administration &gt; Image Group Management &gt; Create (disk-based)</b>		IMAGE_GROUP_ADD
<b>Cluster Administration &gt; Image Group Management &gt; Create Autoinstall</b>	cmu_autoinstall_node	IMAGE_GROUP_ADD
<b>Cluster Administration &gt; Image Group Management &gt; Create (diskless)</b>	cmu_add_image_group	IMAGE_GROUP_ADD
<b>Cluster Administration &gt; Image Group Management &gt; Rename</b>	cmu_rename_image_group	
<b>Cluster Administration &gt; Image Group Management &gt; Delete</b>	cmu_del_image_group	IMAGE_GROUP_DELETE

*Table Continued*

GUI action (top menu bar)	Underlying command (def: /opt/clmgr/bin/ )	DB keyword
<b>Cluster Administration &gt; Image Group Management &gt; Manage Nodes (disk-based and Autoinstall)</b>		IMAGE_GROUP MODIFY
<b>Cluster Administration &gt; Image Group Management &gt; Manage Nodes (diskless)</b>	cmu_add_to_image_group up	IMAGE_GROUP MODIFY
<b>Cluster Administration &gt; Image Group Management &gt; Manage Nodes (diskless)</b>	cmu_del_from_image_group	IMAGE_GROUP MODIFY
<b>Cluster Administration &gt; Custom Group Management &gt; Create</b>		CUSTOM_GROUP_ADD
<b>Cluster Administration &gt; Custom Group Management &gt; Delete</b>		CUSTOM_GROUP_DELETE
<b>Cluster Administration &gt; Custom Group Management &gt; Manage Nodes</b>		CUSTOM_GROUP MODIFY
<b>Monitoring &gt; Restart Monitoring Engine</b>	/opt/clmgr/tools/cmumonitoring	
<b>Monitoring &gt; Start Monitoring Engine</b>	/opt/clmgr/tools/cmumonitoring	
<b>Monitoring &gt; Stop Monitoring Engine</b>	/opt/clmgr/tools/cmustopmonitoring	

The following table shows GUI actions and the commands that run to complete the GUI actions.

**Table 5: GUI features and controls**

GUI feature (right-click node selection)	Underlying command
<b>ssh to admin node/ssh Connection</b>	ssh (node access assumes user account on target node)
<b>Management Card Connection</b>	/opt/clmgr/bin/cmu_console
<b>Shutdown</b>	/opt/clmgr/bin/cmu_halt
<b>Power Off</b>	/opt/clmgr/bin/cmu_power
<b>Boot</b>	/opt/clmgr/bin/cmu_boot
<b>Reboot</b>	/opt/clmgr/bin/cmu_halt
<b>Change UID LED Status</b>	/opt/clmgr/bin/cmu_power
<b>Multi-Window Broadcast (Mgt Card/VSP)</b>	/opt/clmgr/bin/cmu_console
<b>Capture Image</b>	/opt/sgi/sbin/cinstallman

*Table Continued*

GUI feature (right-click node selection)	Underlying command
<b>Provision Image (Deploy)</b>	/opt/sgi/sbin/cinstallman /opt/clmgr/bin/cpower /opt/sgi/sbin/cimage
<b>AutoInstall (kickstart autoyast preseed)</b>	/opt/clmgr/bin/cmu_autoinstall_node
<b>Update &gt; Get Nodes Static Info</b>	/opt/clmgr/tools/cmu_cn_install
<b>Update &gt; Monitoring Client</b>	/opt/clmgr/tools/cmu_cn_install
<b>Update &gt; Rescan MAC</b>	/opt/clmgr/tools/cmu_rescan_mac
<b>Insight &gt; Show BIOS Settings</b>	/opt/clmgr/bin/cmu_firmware_mgmt
<b>Insight &gt; Show BIOS Version</b>	/opt/clmgr/bin/cmu_firmware_mgmt
<b>Insight &gt; Upgrade Firmware</b>	/opt/clmgr/bin/cmu_firmware_mgmt
<b>[Archived Custom Group] Rename</b>	/opt/clmgr/bin/cmurename_archived_custom_group
<b>[Archived Custom Group] Permanently delete</b>	/opt/clmgr/bin/cmudel_archived_custom_groups

## Customizing the cluster manager commands

You can add site-specific menu options to the cluster manager GUI and to the cluster manager command set. You can run the commands you add to the command set at the system prompt. When you add a custom GUI option, the corresponding command is also available in the following directory:

/opt/clmgr/cmuchi

### Procedure

1. Open the following file:

/opt/clmgr/etc/cmu\_custom\_menu

2. Add custom menu options for your site to the `cmu_custom_menu` file.

The `cmu_custom_menu` file contains comments that explain how to edit the file. The file includes instructions that explain how to add commands and GUI options. The file provides commented, ready-to-use examples.

As explained in the file, you can use the custom menu keyword `CMU_SUDO` in the `/opt/clmgr/etc/cmu_custom_menu` file to apply `sudo` support to a command.

3. Save and close the `cmu_custom_menu` file.

4. Use the `cmu_custom_run` command, and review the output, to confirm the new command that you created.

Example 1. The following command displays the help text for the `cmu_custom_run` command:

```
# cmu_custom_run
```

Example 2. The following command displays the list of custom commands. Confirm that the list includes the new command you created in this procedure.

```
# ./cmu_custom_run -l
Title-----| Command
Clear /tmp | env
WCOLL=CMU_TEMP_NODE_FILE /opt/clmgr/bin/pdsh -S 'rm -rf /tmp/*' |
Uptime Martha /tmp | env
WCOLL=CMU_TEMP_NODE_FILE /opt/clmgr/bin/pdsh -S 'uptime' |
HPE iLO4+ Agentless Management Service|Get/Refresh SNMP Data | /opt/
clmgr/bin/cmu_get_ams_data -f CMU_TEMP_NODE_FILE |
HPE iLO4+ Agentless Management Service|Configure ILO | /opt/
clmgr/bin/cmu_config_ams -f CMU_TEMP_NODE_FILE |
HPE iLO4+ Agentless Management Service|Test ILO Config | /opt/
clmgr/bin/cmu_config_ams -t -f CMU_TEMP_NODE_FILE
```

5. On the admin node, create a text file that lists the names of the nodes upon which you want the new command to run.
6. Enter the `cmu_custom_run` command, in the following format, to run the new command:

```
cmu_custom_run -t "command_title" -f file
```

For `command_title`, specify the name for the new command.

For `file`, specify the full path and name of the file that lists the nodes upon which you want this command to run.

For example, the following command runs on the nodes in `/tmp/nodelist`:

```
# cmu_custom_run -t "HPE iLO4+ Agentless Management Service|Test ILO Config" -f /tmp/nodelist
executing "/opt/clmgr/bin/cmu_config_ams -t -f /tmp/nodelist"...
-----
100.117.20.168
-----
AMS is configured
```

7. (Conditional) Edit the `/etc/sudoers` file.

Complete this step if the customized command requires root permission and you want to control access by nonroot users.

## Customizing netboot kernel arguments on flat compute nodes

HPCM includes a diskless netboot environment. Within this environment, you can PXE boot one or more flat compute nodes when you need to perform the following types of actions:

- Firmware upgrades
- Disk RAID setups
- Other non-operating-system, hardware-level configuration actions

The PXE boot process includes a standard bootloader and a boot configuration file. In a custom network environment, you might need to edit the boot configuration file. You can customize the boot file for the entire system or for only a node or subset of nodes.

The following procedure explains how to edit the standard cluster configuration.

## Procedure

1. Log into the admin node as the root user.
2. Boot the cluster over the network.

After the first network boot attempt, you can access the boot configuration file template in the following location:

```
/opt/clmgr/etc/bootopts/pcmdefault
```

This template file contains the minimum kernel boot arguments required to boot the netboot kernel in a standard cluster configuration.

3. Back up the `pcmdefault` file to another location for safekeeping.
4. (Conditional) Create a configuration file that is specific to an individual node, or create multiple configuration files that are specific to groups of individual nodes.

Complete this step if you want to customize the network boot process for one node or for a group of nodes.

To make node-specific changes, copy the default file to a file with a node-specific name or to a file with a name that includes a hexadecimal representation of an IP address.

Example 1. To create a customizable boot parameter file for node `login0`, enter the following command:

```
# cp /opt/clmgr/etc/bootopts/pcmdefault /opt/clmgr/etc/bootopts/login0
```

Example 2. To create a customizable boot parameter file for a specific node using the node IP address, enter the following command:

```
# cp /opt/clmgr/etc/bootopts/pcmdefault /opt/clmgr/etc/bootopts/IP_addr_in_hex
```

For *IP\_addr\_in\_hex*, enter one of the following:

- A full 8-character hex number that represents a specific IP address.

For example, the hex address for 192.168.0.1 is `C0A80001`, which breaks down as follows:

```
192 => C0  
168 => A8  
0 => 00  
1 => 01
```

The command line that represents this IP address is as follows:

```
# cp /opt/clmgr/etc/bootopts/pcmdefault /opt/clmgr/etc/bootopts/C0A80001
```

- A single character that represents a broad subnet of IP addresses.

For example, the hex address for 192.x.x.x is `C`. The command line to represent this range of IP addresses is as follows:

```
# cp /opt/clmgr/etc/bootopts/pcmdefault /opt/clmgr/etc/bootopts/C
```

Example 3. Assume that compute nodes 172.20.0.1 through 172.20.0.15 require the same boot file modification:

In this case, create the following file:

```
/opt/clmgr/etc/bootopts/AC14000
```

The hexadecimal IP address AC14000 includes IP addresses 172.20.0.1 through 172.20.0.15.

5. Edit the default file or edit the node-specific files you created.

When you edit the default file, `pcmdefault`, your changes apply to the entire cluster.

When you edit one of the files you created, your changes apply to only a subset of nodes.

For example, if four nodes (`login0`, `login1`, `login2`, and `login3`) require the same boot file modification, create and edit file `/opt/clmgr/etc/bootopts/login0`. Then, copy `/opt/clmgr/etc/bootopts/login0` to the following files:

```
/opt/clmgr/etc/bootopts/login1  
/opt/clmgr/etc/bootopts/login2  
/opt/clmgr/etc/bootopts/login3
```

## Remote hardware control API (flat clusters only)

The remote hardware control API enables you to integrate cluster manager power and UID control with any computer that has a remote power control capability. The `/opt/clmgr/bin/cmu_power` command interacts with this API to provide remote power and UID control.

The existing hardware APIs are:

- `ILO` - The most common method of interacting with HPE BL/DL/SL servers.
- `lo100i` - The legacy method of interacting with low-end servers.
- `None` - The fallback method of interacting with desktop computers, laptop computers, and servers that do not provide remote power control.
- `IPMI` - The industry standard method, useful with third-party hardware or Hewlett Packard Enterprise hardware that does not contain iLO.
- `ILOCM` - The method for integration with ProLiant Moonshot Chassis.

The hardware API consists of a collection of programs that reside in `/opt/clmgr/hardware/HW_TYPE/`, where `HW_TYPE` refers to the name of the hardware API. For example, the iLO API programs reside in the `/opt/clmgr/hardware/ILO/` directory.

The name of the API programs in the hardware API directory must conform to the following format:

`cmu_HW_TYPE_power_action`

The following values are the five basic required actions. Specify one of the following for `action`:

- `off` - Remove power from the server.
- `on` - Apply power to the server.
- `osoff` - Attempt a graceful shutdown of the OS before removing power.
- `uid_off` - Turn off the UID LED.
- `uid_on` - Turn on the UID LED.

---

**NOTE:** The cluster manager `boot` command is composed of the `osoff` action, followed by the `on` action.

---

The following additional actions are supported by the `/opt/clmgr/bin/cmu_power` command, but the cluster manager does not require them:

- `status` - Provide a power status for the given node, either `on` or `off`.
- `press` - Simulate a momentary press of the power button.

The following programs are required when implementing a new FOO hardware API:

```
/opt/clmgr/hardware/FOO/cmu_FOO_power_off
/opt/clmgr/hardware/FOO/cmu_FOO_power_on
/opt/clmgr/hardware/FOO/cmu_FOO_power_0soff
/opt/clmgr/hardware/FOO/cmu_FOO_power_uid_off
/opt/clmgr/hardware/FOO/cmu_FOO_power_uid_on
```

All of these programs are invoked with the following arguments:

```
PROGRAM -n nodename -i BMC_IP -e errfile
```

The arguments are as follows:

- For `nodename`, specify the host name of the target server.
- For `BMC_IP`, specify the IP address of the management card (also known as the baseboard management controller (BMC)) for the target server.
- For `errfile`, specify the name of a file to which the cluster manager can log errors.

After a new cluster manager hardware API is developed and tested, it must be added to the cluster manager as a valid hardware type. To do this, add the name of the new hardware API to the `CMU_VALID_HARDWARE_TYPES` variable in `/opt/clmgr/etc/cmuserver.conf`. The setting of this variable is `CMU_VALID_HARDWARE_TYPES=ILO:1o100i:ILOCM`.

To add the IPMI hardware API, add IPMI to the list of valid hardware types:

```
CMU_VALID_HARDWARE_TYPES=ILO:1o100i:ILOCM:IPMI
```

After this is done, then you can add nodes to the cluster that include this new type of management card (BMC).

## Support for ScaleMP

You can integrate the cluster manager to work with ScaleMP. To enable ScaleMP support, add the following variable and setting to the `/opt/clmgr/etc/cmuserver.conf` file:

```
CMU_vSMP_PREFIX=vSMP_
```

This setting configures the prefix that is used to identify cluster manager image group nodes that can be PXE booted into the virtual SMP environment. The images that are associated with the image groups can be created with normal cluster manager methods (such as Autoinstall and deploying), or made diskless. The ScaleMP support is activated when nodes that are active members of an image group named `vSMP*` are PXE booted.

If the `vSMP_*` image group is a diskless image, then those nodes PXE boot the diskless image into the virtual SMP environment.

If the `vSMP_*` image group is a disk-based image, then those nodes must be actively PXE booted by selecting `network boot` in the boot menu.

The first node listed as active in the `vSMP_*` image group is the primary node in the virtual SMP environment. The remaining nodes are the secondary nodes.

For more information, see the ScaleMP documentation.

# Ansible integration

You can configure the cluster manager to supply Dynamic Inventory information to Ansible, a popular IT automation tool, in the form of an Ansible-format `hosts` file. As you add, delete, or modify nodes and groups, this inventory information is updated automatically, with no manual refreshing necessary. Additionally, sample Ansible playbooks are included in `/opt/clmgr/contrib/ansible` and can be run from the command line or with the sample entries supplied in `/opt/clmgr/etc/cmu_custom_menu`.

For more information about this feature, click **Custom Tools > Ansible > About Ansible Integration**.

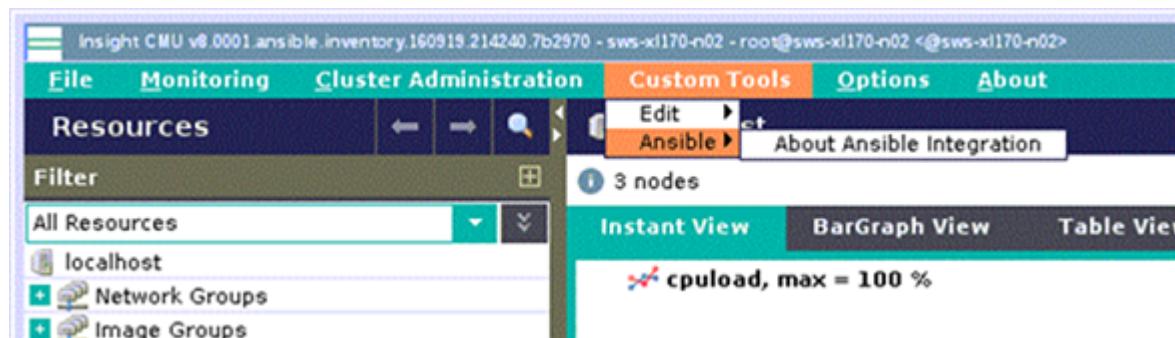


Figure 47: Ansible menu item

# Troubleshooting

Problems encountered while using the cluster manager can be classified as:

- Network boot problems that affect capturing and deploying images
- Administration command problems
- GUI-specific problems

## Retrieving cluster manager service status information

The `configure-cluster` command requires the `cmu` service to be running. The following procedure explains how to obtain the status of the `cmu` service.

### Procedure

1. Log into the admin node as the root user.

2. Enter the following command:

```
# /etc/init.d/cmu status
```

3. In the output, verify that the statuses for the `core` service and the `backend` service report configured.

The cluster manager must be properly configured before using the GUI.

---

**NOTE:** Do not use the `service` command or the `systemctl status` command to display the service statuses.

---

## Log files

Each cluster manager command logs information in a dedicated log file. All log files are available in `/opt/clmgr/log`.

### `cmuserver` log files

When using the GUI, all actions are sent to the `cmuserver` daemon running on the admin node.

The `/opt/clmgr/log/cmuserver-0.log` file on the admin node contains the current output of the `cmuserver` process. Results of the command `/etc/init.d/cmu` option are logged to `/var/log/cmuservice_hostname.log`. The `hostname` component is the host name of the admin node.

### Monitoring log files

The `/opt/clmgr/log/MainMonitoringDaemon_MGTXXX.log` file contains the output of the monitoring daemon running on the admin node.

The `/opt/clmgr/log/SmallMonitoringDaemon_NodeXXX.log` file contains the output of the monitoring daemon running on the compute node.

Designate one of the compute nodes as the secondary server for the network group. The `/opt/clmgr/log/SecondaryServerMonitoring_NodeXXX.log` file contains the output of the secondary server process.

# Network boot problems

Nodes might not boot in network mode due to one of the following causes:

- A hardware problem on the node
- A network problem with the network cable, switch port, or the NIC
- An incorrect MAC address in the cluster database
- The cluster configuration on the admin node is lost

## Troubleshooting switch problems

### Procedure

1. Verify that the admin node pings the management card (iLO or baseboard management controller (BMC)) and compute nodes.
2. Verify that the spanning tree is disabled on all ports connected to a node.
3. Verify that multicast IGMP snooping is disabled on the switch.

From the admin node, use the `telnet` command to log into the switch. The following examples show how to retrieve the status of IGMP snooping on various switches.

Example 1. On HPE FlexNetwork or HPE FlexFabric switches, use the `system-view` command and the `display igmp-snooping global` command, as follows:

```
neteng-050:~ # telnet mgmtsw
Trying switch_ip_addr...
Connected to mgmtsw0.
Escape character is '^]'.

*****
* Copyright (c) 2010-2018 Hewlett Packard Enterprise Development LP      *
* Without the owner's prior written consent,                            *
* no decompiling or reverse-engineering shall be allowed.           *
*****  
  
login: admin
Password:  
  
<mgmtsw0>system-view
System View: return to User View with Ctrl+Z.  
  
[mgmtsw0]display igmp-snooping global
IGMP snooping is not configured.
[mgmtsw0]quit
<mgmtsw0>quit
```

Example 2. On Edgecore or Ericsson-LG switches, use the `show ip igmp snooping` command, as follows:

```
# telnet mgmtsw5
Trying switch_ip_addr...
Connected to mgmtsw5.
```

```
Escape character is '^]'.
```

#### User Access Verification

```
Username: admin
```

```
Password:
```

```
CLI session with the ECS4610-50T is opened.  
To end the CLI session, enter [Exit].
```

```
mgmtsw5-0#show ip igmp snooping
```

IGMP Snooping	:	Disabled
Router Port Expire Time	:	300 s
Router Alert Check	:	Disabled
Router Port Mode	:	Forward
TCN Flood	:	Disabled
TCN Query Solicit	:	Disabled
Unregistered Data Flood	:	Enabled
Unsolicited Report Interval	:	400 s
Version Exclusive	:	Disabled
Version	:	3
Proxy Reporting	:	Disabled
Querier	:	Enabled

```
.
```

```
.
```

```
.
```

```
mgmtsw5-0#exit
```

Example 3. On Extreme Networks switches, use the **show igmp snooping** command, as follows:

```
# telnet mgmtsw2
```

```
Trying switch_ip_addr...
```

```
Connected to mgmtsw2.
```

```
Escape character is '^]'.  
  
telnet session telnet0 on /dev/ptyb0  
  
login: admin  
password:  
  
ExtremeXOS  
Copyright (C) 1996-2016 Extreme Networks. All rights reserved.  
This product is protected by one or more US patents listed at http://  
www.extremenetworks.com/patents along with their foreign counterparts.  
=====
```

```
=====  
  
Press the <tab> or '?' key at any time for completions.  
Remember to save your configuration changes.
```

```
Slot-1 mgmtsw2.9 # show igmp snooping  
Igmp Snooping Flag : forward-all-router
```

```

Igmp Snooping Flood-list      : none
Igmp Snooping Proxy          : Enable
Igmp Snooping Filters         : per-port

Vlan          Vid  Port   #Senders #Receivers Router Enable
-----
Default       1     0                  No
vlan0003      3     0                  No

Slot-1 mgmtsw2.2 # exit

```

## Troubleshooting network boot

### Procedure

1. Open a console to the management card (also known as the baseboard management controller (BMC)) of the node.
2. Select the node in the main GUI window and boot in network mode.
  - Does the node receive a DHCP response from the server?
    - If yes, check the IP address received by the node to verify that the correct server responded. If so, proceed to the next verification.
    - If no, shut down the other server. Verify the configuration of DHCP and PXE, and verify that spanning tree is not enabled on the switch connected to the node.

For more information on selecting which network sends the node its DHCP address, see the following:

[\*\*Customizing netboot kernel arguments on flat compute nodes\*\*](#) on page 124

- Can the node download its kernel?
    - If yes, proceed to the next verification.
    - If no, verify the `tftp` daemon configuration.
  - Can the node mount the root file system (/) with NFS?
    - If yes, the network image might be corrupted. Reinstall the cluster manager RPM.
    - If no, verify that the NFS server is started.
3. Verify that `/opt/clmgr/ntbt/rb` is exported to all nodes with NFS.

# Administration command problems

## Procedure

1. Verify that rsh, telnet, or ssh is configured on nodes.
2. Verify that the rsh or ssh root is enabled with the node password to all nodes of the cluster.
3. Verify that the database contains the correct IP address and host name.

# GUI problems

## GUI cannot be launched from browser

### Symptom

The GUI does not launch from the browser.

### Solution 1

#### Action

1. Clear the browser cache.
2. Clear the Java cache using the Java Control Panel applet.

Run the appropriate tool (for example, `jcontrol`) to access the Java Control Panel.

### Solution 2

#### Cause

Browser proxy settings are blocking the GUI launch.

#### Action

If you receive a certificate validation error while launching the GUI, check the network settings in the Java Control Panel applet. If it is set to use browser settings, browser proxy settings might be blocking the GUI launch. Try using "Direct connection" in the Java Control Panel. Run the appropriate tool (for example, `jcontrol`) to access the Java Control Panel.

### Solution 3

#### Action

1. If you receive a certificate expiration error while launching the GUI, add the admin node IP address to the exception list in the Java Control Panel applet and launch the GUI again.
  - a. **Control panel > java > Security > Exception Site list**
  - b. In the **Location** field, enter the IP address of the admin node.
  - c. Click **Add**.

A security warning for HTTP location displays.



- d. Click **OK**.

The site IP address is added to the exception list.

- e. Click **OK**.

## GUI cannot contact the remote cluster manager service

### Symptom

The GUI cannot contact the remote cluster manager service.

### Action

1. Enter the following command to verify that the cluster manager service is running properly on the admin node:

```
# /etc/init.d/cmu status
```

If the cluster manager service is not running properly, enter the following commands to stop and then start the service:

```
# systemctl stop cmu
# systemctl start cmu
```

2. Verify that the GUI on the client system is connected to the correct server.
3. Verify the `GENERAL_RMI_HOST` setting in the `cmu_gui_local_settings` file on the client system.
4. If `cmuserver` is running properly on the admin node, verify the following:
  - The firewall configuration on the admin node and the client system.
  - The RMI connections. Verify that RMI connections are allowed between the two hosts.

## GUI is running, but the monitoring sensors are not updated

### Symptom

The GUI is running, but the monitoring sensors are not been updated.

### Action

1. Verify that the cluster manager service is running properly on the admin node.

```
# /etc/init.d/cmu status
```

If the cluster manager service is not running properly, enter the following commands to stop and then start the cluster manager service:

```
# systemctl stop cmu
# systemctl start cmu
```

2. Verify that the host file of the nodes is properly configured. Each node must have access to the IP address of all other nodes in the cluster.
3. Verify that rsh or ssh is enabled between all nodes of the cluster and the admin node. All nodes must be able to execute commands as root for any other node without needing a password.
4. Verify that the cluster manager RPM is properly installed on all nodes.

The following commands return information that shows the RPM being properly installed:

- On the admin node:

```
admin:~ # rpm -q cmu
cmu-X.X.xxx.release.xxx.x86_64
```

- On non-admin nodes:

```
n1:~ # rpm -q cmu_cn
cmu_cn-X.X.xxx.release.xxx.x86_64
```

## Failed to validate certificate error displays

If the GUI is unable to start, the following Failed to validate certificate message displays:

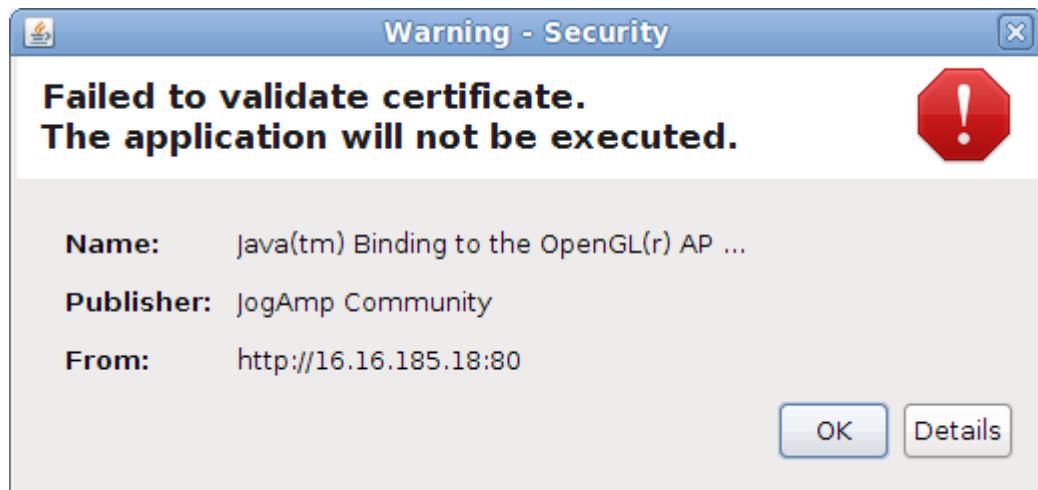


Figure 48: Certificate error

The detailed Java exception is as follows:

```
java.security.cert.CertPathValidatorException:  
java.security.InvalidKeyException: Wrong key usage
```

Change the Java setting value. The default value changed between Java 1.6u31 and Java 1.7u12.

If you are connected to the Internet, activate **Enable online certificate validation**. If you are not connected to the Internet, deactivate **Enable online certificate validation**.

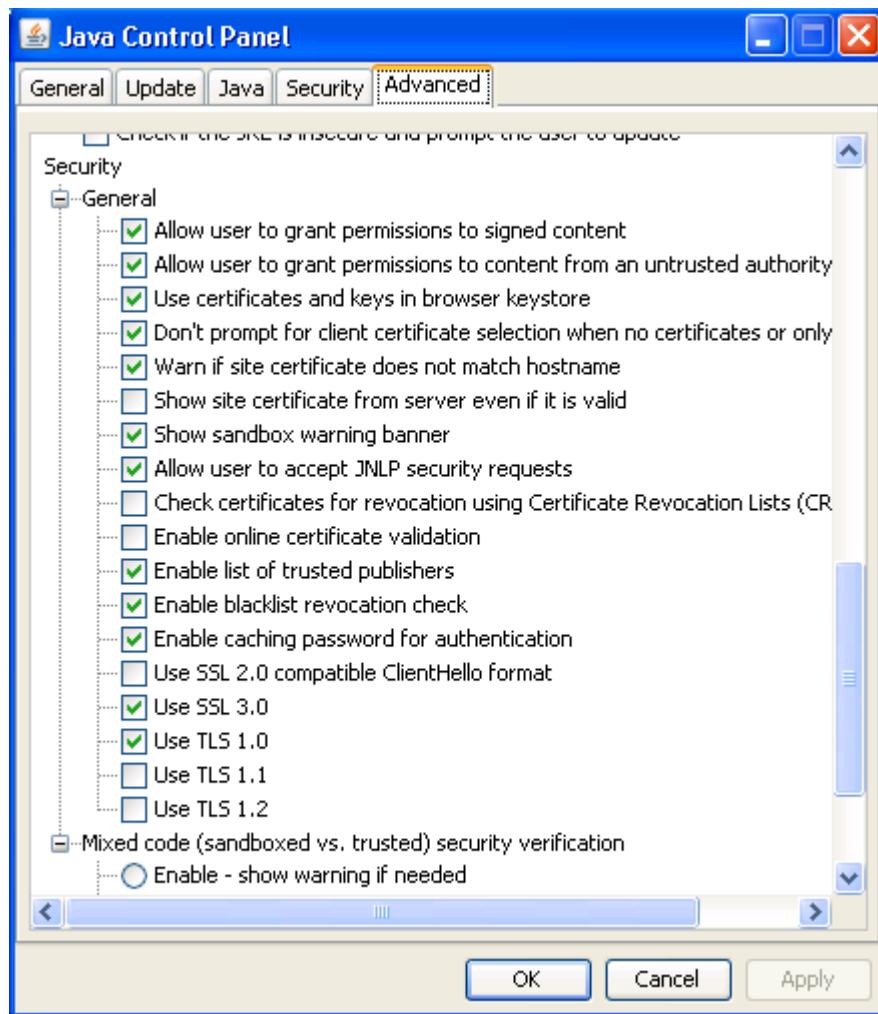


Figure 49: Java control panel

On Windows GUI client nodes, go to **System Preferences > Other > Java > Advanced > Enable online certificate validation**.

On Linux, run `javaws -viewer` in a shell, click the **Advanced** tab, then **Enable online certificate validation**.

---

 **TIP:** If you still encounter problems, try toggling the setting.

---

# Manpages

The cluster manager online manpages reside in the following directories:

- /opt/clmgr/man
- /opt/sgi/share/man

To retrieve a manpage, peruse the preceding directories for the command that interests you and use the `man(1)` command to display the output.

# Administering the cluster in the style of SGI Management Center

The chapters that follow explain how to administer the cluster using commands in the style of SGI Management Center.

# Configuring optional features on flat compute nodes

This chapter contains information about features you can configure on flat compute nodes. In many cases, your operating system distribution documentation explains the procedure in detail. The information in this manual supplements the distribution documentation and explains extra steps needed for the cluster.

## Configuring ICE compute nodes to use a flat compute node as a network address (NAT) gateway

Use the documentation from your software distribution to configure a NAT gateway on a flat compute node. During the configuration, update the static routes. The following procedure explains how to update the `sgi-static-routes.sh` file to set up a default route to the `ib0` IP address of a flat compute node.

### Procedure

1. Enter the following command to change to the directory where the static routes update script resides:

```
admin# cd /opt/sgi/share/per-host-customization/global
```

This directory contains several scripts. The system runs these scripts at startup, upon demand, or when you request. For example, the script runs when you use the `cimage` command with the `--push-rack` and `--customizations-only` options. The next few steps explain how to edit the `sgi-static-routes.sh` script file to point to the `ib0` IP address of a flat compute node.

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

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

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

Make sure that this line includes the IP address of `ib0` and the literal string `ib0`. The line might be correct in the file, but if necessary, edit the line. For this example, edit the line to remove the comment characters (#). The result is as follows:

```
if [ -d ${imagedir}${SLES_PATH} ]; then
    echo "default 10.148.0.2 - ib0 -" >>${imagedir}${SLES_PATH}routes
fi
if [ -d ${imagedir}${RHEL_PATH} ]; then
    echo "default via 10.148.0.2" >>${imagedir}${RHEL_PATH}route-ib0
fi
```

The `sgi-static-routes.sh` script customizes the network routing based upon the rack, the individual rack unit (IRU), and the slot of the ICE compute node. Some examples are available in the script.

4. Push the new image, as follows:

- Enter the following command to shut down and stop all the ICE compute nodes:  

```
admin# cpower node halt "r*i*n*"
```
- Enter the following command to propagate the changes:  

```
admin# cimage --push-rack image_name "r*"
```
- Enter the following command to power up all the ICE compute nodes:  

```
admin# cimage --push-rack image_name "r*"
```

When you power up the ICE compute nodes, the new default gateway route is present on all ICE compute nodes.

## Configuring a flat compute node as an NFS server

Use the documentation from your software distribution to configure an NFS server on a flat compute node. After you complete the NFS configuration, use the following procedure to create custom mount points for the ICE compute nodes.

### Procedure

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

```
/opt/sgi/share/per-host-customization/global
```

In the following steps, you complete the following actions:

- You add the new file system to the `sgi-fstab.sh` file.
- You ensure that the new file system mounts.

2. Use a text editor to open the following file on the admin node:

```
sgi-fstab.sh
```

3. Within file `sgi-fstab.sh`, add a line for file system mount point, and then save and close the file.

4. Enter the following command to shut down and stop all the ICE compute nodes:

```
admin node:~ # cpower node halt "r*i*n*"
```

5. Enter the following command to propagate the changes:

```
admin node:~ # cimage --push-rack image_name "r*"
```

6. Enter the following command to power up all the ICE compute nodes:

```
admin node:~ # cpower node on "r*i*n*"
```

When you power up the ICE compute nodes, the file system mounts on all ICE compute nodes.

## Configuring scratch disk space on system disks

By default, the cluster manager does not allocate disk space for customer use on the system disk of cluster nodes. The following topics describe how you can allocate disk space for customer use on the following types of nodes:

- [Configuring scratch disk space on leader and flat compute nodes](#) on page 141
- [Configuring scratch disk space on an admin node](#) on page 142

## Configuring scratch disk space on leader and flat compute nodes

The following procedure uses the `cinstallman` command to allocate scratch disk space on leader and flat compute nodes.

### Procedure

1. Use the `cinstallman` command in the following format to provide the primary specifications:

```
cinstallman --next-boot image --node nodes --force-disk dev --destroy-disk-label --root-disk-reserve nn
```

The variables are as follows:

Variable	Specification
<i>nodes</i>	The affected nodes. For example, use <code>n[1-5]</code> for hostnames <code>n1</code> through <code>n5</code> .
<i>dev</i>	The disk device name. For example, <code>/dev/sdz</code> .
<i>nn</i>	The disk space size in GiB.

When you reboot the affected nodes, a post-install script creates scratch disk space on the nodes with the following attributes:

- A single partition, partition number 61
  - Disk space of  $nn$ GiB
  - An ext4 file system
  - Mount point `/scratch` in the image
  - Persistent across installs
2. To customize the attributes of the disk space described in the previous step (for example, to customize the number of partitions), edit the following post-install script:  
`50all.create_filesystem_for_reserved_partition`  
The script is located in directory `/opt/clmgr/image/scripts/post-install`.  
The following notes and restrictions apply:
    - HPE expects you to modify this script.
    - If you add your own set of partitions, ensure that they start at 61.
  3. Reboot the desired nodes.

For example, the following command reboots flat compute nodes 1-5:

```
# cpower node reset "n[1-5]"
```

---

**NOTE:** You can also create the scratch disk space when you run the `discover` command to configure the node. The following `discover` command parameters create the same specifications as the `cinstallman` command parameters:

- `force_disk=dev`
- `destroy_disk_label=yes`
- `root_disk_reserve=nn` (in GiB)

For a description of node entries in the cluster definition file, see the following:

[\*\*HPE Performance Cluster Manager Installation Guide\*\*](#)

---

## Configuring scratch disk space on an admin node

You can allocate scratch disk space on the system disk of an admin node. To complete this task, add the following parameters to the kernel parameter list of the admin node installer:

- `destroy_disk_label=yes`
- `root_disk_reserve=size.`

For `size`, specify a size in GiB

As a result, the cluster manager creates the scratch disk space in partition 61. Notice that you must configure the scratch disk space. That is, you must create the file system, add the `fstab` entries, and complete other tasks.

## Configuring software RAID on cluster nodes

You can use the cluster manager to configure standard software RAID levels for admin, leader, and flat compute nodes. The RAID levels include 0, 1, 4, 5, 6, and 10. For any given RAID level, the RAID can be one of the following types:

- BIOS SW RAID
- Native MD

Each RAID scheme has its own distinct metadata. The cluster manager supports only BIOS SW RAID on admin nodes. On other managed nodes, the cluster manager supports both BIOS SW RAID and native MD. In general, configure RAID for a system disk as a one-time action for the life of the hardware.

The following topics describe how to configure RAID for different disk types:

- [\*\*Configuring software RAID on an admin node\*\*](#) on page 143
- [\*\*Configuring software RAID on leader nodes and flat compute nodes\*\*](#) on page 143

## Configuring software RAID on an admin node

The following procedure uses the cluster manager to configure software RAID on the admin node in an automated way.

### Procedure

1. Change the **BIOS SATA Configuration** mode from **AHCI** to **RAID**.
2. Add the following parameters to the kernel parameter list of the admin installer:

- `md_metadata=imsm`
- `md_raidlevel=n`

For *n*, specify an integer that represents the RAID level you want. The default is 1.

- (Optional) `force_disk="device1,device2,...,deviceN"`

If you use this parameter, specify the path to each disk device required for the RAID level you choose. For example:

`force_disk="/dev/sda,/dev/sdb,/dev/sdc"`

By default, the cluster manager configures RAID 1 on the first two empty disks.

---

**NOTE:** Disk identifiers listed with `force_disk`, such as `/dev/sda`, are used once and are automatically assembled by MD thereafter. These identifiers are not persistent; they could unexpectedly change from what you assumed.

Even though the identifiers are used only once, HPE recommends that you use `/dev/disk/by-id` or `/dev/disk/by-path` names for the disks instead. In this way, the following occur:

- The cluster manager operates on the exact disks you target
- You do not rely on device names that might not point where you expect

- 
3. Verify the new RAID volume.

To verify that the new RAID volume is the new boot device, stop at BIOS on a reboot and navigate to the **Boot** menu.

## Configuring software RAID on leader nodes and flat compute nodes

The following procedure explains how to use the `cinstallman` command to configure RAID on leader and flat compute nodes.

### Procedure

1. Use the `cinstallman` command in the following format to provide the primary specifications:

```
cinstallman --next-boot image --node nodes --md-metadata value --md-raidlevel n --force-disk dev
```

The variables are as follows:

---

Variable	Specification
<i>nodes</i>	The affected nodes. For example, use <code>n[1-5]</code> for hostnames <code>n1</code> through <code>n5</code> .
<i>value</i>	<p>The metadata type. Specify one of the following:</p> <ul style="list-style-type: none"> <li>• <code>imsm</code>, which specifies BIOS software RAID. The node must have the Intel Storage Manager or its equivalent for BIOS support.</li> <li>• <code>md</code>, which specifies native MD.</li> </ul>
<i>n</i>	An integer that signifies the RAID level you want to configure. The default is 1.
<i>dev</i>	<p>The disk device names. Use the following format: <code>"device1,device2,...,deviceN"</code>. For example: <code>"/dev/sda,/dev/sdb,/dev/sdc"</code> By default, the cluster manager configures RAID 1 on the first two empty disks.</p> <p><b>NOTE:</b> Disk identifiers listed with <code>--force_disk</code>, such as <code>/dev/sda</code>, are used once and are automatically assembled by MD thereafter. These identifiers are not persistent; they could unexpectedly change from what you assumed.</p> <p>Even though the identifiers are used only once, HPE recommends that you use <code>/dev/disk/by-id</code> or <code>/dev/disk/by-path</code> names for the disks instead. In this way, the following occur:</p> <ul style="list-style-type: none"> <li>• The cluster manager operates on the exact disks you target</li> <li>• You do not rely on device names that might not point where you expect</li> </ul>

---

## 2. Reboot the affected nodes.

For example, the following command reboots flat compute nodes 1-5:

```
cpower node reset "n[1-5]"
```

---

**NOTE:** You can also configure software RAID when you run the `discover` command to configure the node. The following `discover` command parameters create the same specifications as the `cinstallman` parameters:

- `md_metadata=imsm` or `md_metadata=md`
- `md_raidlevel=n`
- `force_disk=dev`

The `md_metadata` and `force_disk` options are destructive options designed to be one-time actions. The cluster manager database does not store these values. If you use the `discover` command to regenerate the cluster definition file, these options do not appear.

For a description of node entries for a `discover` cluster definition file, see the following:

**HPE Performance Cluster Manager Installation Guide**

---

## Configuring trusted boot nodes

The following topics explain how to configure cluster nodes to support the trusted platform module (TPM) or **trusted boot** feature:

- [Prerequisites and constraints](#) on page 145
- [Configuring trusted boot nodes at configuration time](#) on page 146
- [Configuring trusted boot nodes after the initial installation and configuration](#) on page 147
- [Building trusted boot node images](#) on page 147
- [Verifying that a node booted in trusted mode](#) on page 148
- [Managing trusted boot nodes](#) on page 149

## Prerequisites and constraints

The following are prerequisites for configuring trusted boot nodes:

- HPE supports trusted boot on for leader nodes and flat compute nodes.
- HPE does not support trusted boot on systems that boot in EFI mode.
- Trusted boot requires TPM-enabled hardware and TPM-enabled BIOS on the nodes that will be booting.
- Trusted boot requires the `tboot` package to be installed on the admin node. For example, use one of the following commands to install the `tboot` package:
  - On RHEL platforms, enter the following command:  
`admin:~# cinstallman --yum-image --node admin install tboot`
  - On SLES platforms, enter the following command:  
`admin:~# cinstallman --zypper-image --node admin install tboot`

- Trusted boot requires the following packages to be installed in each node image:

- tboot
- tpm-tools
- trousers

For information, see the following:

[\*\*Building trusted boot node images\*\*](#) on page 147

- Make sure that the `tcsd` service is enabled on the trusted node. For information about how to enable the `tcsd` service, see the following:

[\*\*Building trusted boot node images\*\*](#) on page 147

## Configuring trusted boot nodes at configuration time

To configure trusted boot nodes at configuration time, do the following:

- Ensure that the desired nodes meet the prerequisites and constraints.

For information, see the following:

[\*\*Prerequisites and constraints\*\*](#) on page 145

- Build a trusted boot node image.

For information, see the following:

[\*\*Building trusted boot node images\*\*](#) on page 147

- In the cluster definition file, use the `image=my-trusted-image` parameter to assign the trusted boot image to the desired nodes.

- In the cluster definition file, add the `trusted_boot=yes` parameter to enable trusted boot mode for all desired nodes.

- If you want the trusted nodes to boot from disk after initial installation, select one of two boot-from-disk modes. To configure the boot-from-disk modes, do one of the following:

- Add the `disk_bootloader=yes` parameter to the cluster definition file entries for the desired nodes.
- Change the BIOS boot order of the desired nodes.

For information about the boot-from-disk modes, see the following:

[\*\*Booting leaders or flat compute nodes from a local disk\*\*](#) on page 182

When you run the `discover` command to configure the cluster, the desired nodes are configured and put into trusted boot mode. For information about the `discover` command and its configuration file, see the following:

[\*\*HPE Performance Cluster Manager Installation Guide\*\*](#)

## Configuring trusted boot nodes after the initial installation and configuration

The following procedure explains how to configure trusted boot nodes after the initial installation and configuration session.

1. Ensure that the desired nodes meet the prerequisites and constraints.

For information, see the following:

[Prerequisites and constraints](#) on page 145

2. Build a trusted boot node image.

For information, see the following:

[Building trusted boot node images](#) on page 147

3. From the admin node, enter the following command to assign the trusted boot image to the chosen flat compute nodes:

```
# cinstallman --assign-image --node target_nodes --image new_image
```

4. Enter the following command to stage the image for the next network boot on the target nodes:

```
# cinstallman --next-boot image --node target_nodes
```

5. Enter the following command to enable trusted boot mode on the target nodes:

```
# cadmin --enable-tpm-boot --node target_nodes
```

6. (Optional) Enable boot-from-disk mode.

There are two possible boot-from-disk modes. Complete one of the following actions to enable boot-from disk on the target nodes:

- Use the `cadmin` command to enable a disk bootloader on the target nodes:

```
# cadmin --enable-disk-bootloader --node target_nodes
```

- Change the BIOS boot order on the target nodes.

For information about the two boot-from-disk modes, see the following:

[Booting leaders or flat compute nodes from a local disk](#) on page 182

7. Enter the following command to reboot the target nodes:

```
# cpower node reset target_nodes
```

When enabled on boot, GRUB 2 boots `tboot.gz` before booting the kernel and `initramfs`.

## Building trusted boot node images

The following procedure explains how to configure a flat compute node or leader node image as a trusted boot node. This procedure makes the required changes to a new node image in the following directory on the admin node:

`/opt/clmgr/image/images`

## Procedure

1. Use an `ssh` connection to log into the admin node as the root user.
2. Create a new flat compute node image based on the default image for your operating system.

```
# cinstallman --create-image --clone --source default_os --image new_image
```

For example, to target a flat compute node using RHEL 7.4, you might enter the following:

```
# cinstallman --create-image --clone --source rhel7.4 --image rhel7.4-tpm
```

3. Install the required packages in the new image.

For example:

- On RHEL platforms, enter the following command:

```
# cinstallman --yum-image --image new_image install tboot tpm-tools trousers
```

- On SLES platforms, enter the following command:

```
# cinstallman --zypper-image --image new_image install tboot tpm-tools trousers
```

4. Enter the following command to switch to the image environment:

```
# chroot /opt/clmgr/image/images/new_image
```

5. Enable the `tcsd` service in the image.

Use one of the following commands:

- For RHEL 7 (CentOS 7) or SLES 12 platforms, enter the following command:

```
IMG:~# echo 'enable tcsd.service' > \
/usr/lib/systemd/system-preset/80-enable-tcsd.preset
```

- For RHEL 6 (CentOS 6) or SLES11 platforms, enter the following command:

```
IMG:~# chkconfig tcsd on
```

6. Enter the following command to leave the `chroot` environment:

```
IMG:~# exit
```

7. Enter the following command to query the VCS history of the image:

```
# cinstallman --history --image new_image
```

8. Enter the following command to verify that the working copy has the expected file changes:

```
# cinstallman --changed --image new_image
```

9. Enter the following command to commit the working copy of the image:

```
# cinstallman --commit --image new_image --msg "Trusted Boot Config"
```

## Verifying that a node booted in trusted mode

To ascertain levels of trust, the trusted execution technology (TXT) uses a variety of measurements and special platform configuration registers (PCRs) to hold the measurements. The technology uses the term **measured boots** to refer to trusted boots.

To verify that a trusted boot has occurred, looking at the following measurement artifacts on the node:

- Examine the PCRs on the node:

```
# cat /sys/class/misc/tpm0/device/pcrs
```

- Examine the TXT statistics.

```
# /usr/sbin/txt-stat --heap    txt-stat.log
grep -i measure txt-stat.log
grep -i senter txt-stat.log
```

A quick way to verify that a measured boot occurred is to look for the text `TXT measured launch` in the statistics. This line appears as follows:

```
# /usr/sbin/txt-stat --heap | grep measure
TXT measured launch: TRUE
```

## Managing trusted boot nodes

On the admin node, you can use the following three commands to enable trusted boot, disable trusted boot, or show the trusted-boot setting for a specified node, respectively:

- `cadmin --enable-tpm-boot --node node_name`
- `cadmin --disable-tpm-boot --node node_name`
- `cadmin --show-tpm-boot --node node_name`

# System operation

This chapter includes the following topics:

- [Changing global cluster configuration settings](#) on page 150
- [discover command](#) on page 154
- [Managing slots](#) on page 157
- [Powering on and powering off cluster systems and cluster system components](#) on page 160
- [Power and energy management](#) on page 168
- [pdsh and pdcp commands](#) on page 168
- [Using the cadmin command, the administrative interface](#) on page 170
- [Console management](#) on page 179
- [Synchronizing system time](#) on page 180
- [Booting leaders or flat compute nodes from a local disk](#) on page 182
- [Changing the size of /tmp on ICE compute nodes](#) on page 183
- [Switching ICE compute nodes to a tmpfs root](#) on page 183
- [Configuring local storage space for swap and scratch disk space](#) on page 187
- [Using the cattr command to modify system attributes](#) on page 191
- [About disk quotas](#) on page 192
- [Creating custom partitions](#) on page 196
- [Backing up and restoring the system database](#) on page 199
- [Enabling EDNS](#) on page 201

## Changing global cluster configuration settings

This topic explains how to use the cluster configuration tool, `configure-cluster`, to enable optional features. The features you can enable depend on your hardware platform and your site requirements. When you use the cluster configuration tool, you use the menus to set systemwide, global values. The values you set apply to all nodes that you discover after you set the value, and the effects are as follows:

- When you configure a system for the first time, you run the cluster configuration tool before you run the `discover` command. All the nodes you discover receive the global values you set in the cluster configuration tool.
- You can add nodes or change global values on a production system. In these cases, you might need to use commands to reset values on older nodes that you had configured previously.

The `configure-cluster` and `discover` commands work in concert with the cluster definition file. This file defines the following:

- The roles of the various cluster nodes
- Global system attributes

- Data networks and their respective switches
- Management networks and their respective switches

The cluster definition file is a configuration file. The file provides a convenient and efficient method of specifying large-scale changes.

For an overview and examples of the cluster definition configuration file, see the following:

#### **HPE Performance Cluster Manager Installation Guide**

To preserve custom configuration changes across `update-configs` calls, see the following:

#### **Preserving custom configuration changes** on page 309

The following topics explain how to use `configure-cluster` or related commands to change global cluster configuration settings:

- [\*\*Changing the network time protocol \(NTP\) server\*\*](#) on page 151
- [\*\*Changing the site domain name service \(DNS\) server information\*\*](#) on page 151
- [\*\*Enabling or disabling a backup domain name service \(DNS\) server\*\*](#) on page 152
- [\*\*Configuring a redundant management network \(RMN\)\*\*](#) on page 152
- [\*\*Configuring the blademond rescan interval\*\*](#) on page 154

## **Changing the network time protocol (NTP) server**

The following procedure explains how to change or update your NTP server information in the cluster configuration database.

### **Procedure**

1. From the video graphics array (VGA) screen, or through an `ssh` connection, log into the admin node as the root user.
2. Enter the following command to start the cluster configuration tool:  
`# /opt/sgi/sbin/configure-cluster`
3. On the cluster configuration tool main menu, select **T Configure Time Client/Server (NTP)**, and select **OK**.
4. On the **This procedure will replace your ntp configuration file ...** screen, select **Yes**.
5. On the **A new ntp file has been put into position and includes server broadcast entries for the admin node cluster networks ...** screen, select **OK**.

## **Changing the site domain name service (DNS) server information**

The following procedure explains how to change or update your site DNS server information in the cluster configuration database.

## Procedure

1. From the VGA screen, or through an `ssh` connection, log into the admin node as the root user.
2. Enter the following command to start the cluster configuration tool:  
`# /opt/sgi/sbin/configure-cluster`
3. On the cluster configuration tool main menu, select **D Configure House DNS Resolvers**, and select **OK**.
4. On the **Enter up to three DNS resolvers IPs** screen, enter the IP addresses you want to configure.
5. Select **OK**.

## Enabling or disabling 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 ICE compute nodes can use a flat compute node as a secondary DNS server if the admin node is not available. You can configure a backup DNS only after you run the `discover` command to configure the cluster. This feature is optional.

The following examples show how to use commands to enable or disable a backup DNS.

- Example 1. To retrieve current DNS backup information, enter the following:

```
# /opt/sgi/sbin/backup-dns-setup --show-backup  
service0
```

- Example 2. To disable the backup DNS, enter the following:

```
# /opt/sgi/sbin/backup-dns-setup --delete-backup  
Shutting down name server BIND waiting for named to shut down (28s) done  
sys-admin: update-configs: updating SMC configuration files  
sys-admin: update-configs: -> dns  
.  
.  
.
```

- Example 3. To enable a backup DNS on `service0`, enter the following:

```
# /opt/sgi/sbin/backup-dns-setup --set-backup service0  
Shutting down name server BIND waiting for named to shut down (29s)  
done  
sys-admin: update-configs: updating SMC configuration files  
sys-admin: update-configs: -> dns  
.  
.  
.
```

To use the cluster configuration tool to enable or disable the backup DNS, see the following:

[HPE Performance Cluster Manager Installation Guide](#)

## Configuring a redundant management network (RMN)

By default, a cluster includes an RMN. An RMN is a secondary network from the nodes to the cluster network. When an RMN is enabled, the Linux bonding mode for leader nodes and compute nodes is 802.3ad link aggregation. The RMN has the following additional characteristics:

- The GigE switches are doubled in the system control network and stacked (using stacking cables).
- The links from the chassis management controllers (CMCs) are doubled.
- Some links from the admin node, leader nodes, and most compute nodes are doubled.
- Baseboard management controller (BMC) connections are not doubled. This design means that certain failures can cause temporary inaccessibility to the BMCs. During these failures, the host interfaces remain accessible.

The following are methods you can use to enable or disable an RMN:

- Use the `discover` command and its `redundant_mgmt_network` parameter.

Example 1. The following command disables the RMN on node `service0`:

```
# discover --node 0,xe500,redundant_mgmt_network=no
```

Example 2. The following command disables the RMN on rack leader 1:

```
admin:~ # discover --leader 1,redundant_mgmt_network=no
```

- Use the `cadmin` command and one of the following parameters:

- `--enable-redundant-mgmt-network`
- `--disable-redundant-mgmt-network`

If you use the `cadmin` command to change a compute node or a leader node, reboot the node to make your changes take effect.

Example 1. The following `cadmin` command enables the RMN on node `service0`:

```
# cadmin --enable-redundant-mgmt-network --node service0
```

Example 2. The following `cadmin` command enables the RMN on leader node `r1lead` and shows the required subsequent reboot:

To turn on the redundant management network on a leader node, perform the following command:

```
# cadmin --enable-redundant-mgmt-network --node r1lead
r1lead should now be rebooted.
# cpower leader reboot r1lead
```

- Use the cluster configuration tool.

You can use the cluster configuration tool to enable an RMN. When you use the cluster configuration tool to configure an RMN, the system enables an RMN for all nodes that you discover after you enable the setting. If you have existing nodes in the cluster without an RMN, those existing nodes are not changed.

For more information about the RMN, see the following:

#### **HPE Performance Cluster Manager Installation Guide**

The following procedure explains how to use the cluster configuration tool to configure an RMN.

## Procedure

1. From the VGA screen, or through an `ssh` connection, log into the admin node as the root user.
2. Enter the following command to start the cluster configuration tool:  
`# /opt/sgi/sbin/configure-cluster`
3. On the **Main Menu** screen, select **M Configure Redundant Management Network (optional)**, and select **OK**.
4. On the pop-up window that appears, select **Y yes** (default), and select **OK**.

## Configuring the `blademond` rescan interval

When enabled, the system checks every two minutes for changes to the number of ICE compute nodes in the system. If you remove or add an ICE compute node, the system automatically does the following:

- Detects the change
- Updates the system
- Integrates the change on the rack

By default, the interval between checks is set to 120, which is two minutes.

Use the following procedure to configure the `blademond` rescan interval from the cluster configuration tool.

## Procedure

1. From the VGA screen, or through an `ssh` connection, log into the admin node as the root user.
2. Enter the following command to start the cluster configuration tool:  
`# /opt/sgi/sbin/configure-cluster`
3. On the **Main Menu** screen, select **C Configure blademond rescan interval (optional)**, and select **OK**.
4. On the pop-up window that appears, accept the default of 120, which is two minutes, and select **OK**.  
Alternatively, enter a different value and select **OK**.

## discover command

The `discover` command configures nodes into a cluster system or configures nodes into a rack. It configures both the nodes and their associated BMC controllers. Generally, leader node numbering starts at one and compute node numbering starts at zero. The `discover` command also configures external InfiniBand switches and system management switches.

For information about `discover` command usage, see the following:

- The `discover(8)` manpage.
- The installation guide, which describes how to configure components in a cluster. The installation guide describes how to use the cluster definition file in conjunction with the `discover` command. Access the installation guide at the following link:

The following topics describe additional tasks you can perform with the `discover` command:

- [\*\*Using the generic hardware type\*\*](#) on page 155
- [\*\*Configuring a compute node to use a nondefault image\*\*](#) on page 155
- [\*\*Skipping a node while configuring\*\*](#) on page 156
- [\*\*Omitting unneeded switch configurations when reconfiguring\*\*](#) on page 156

## **Using the generic hardware type**

You can use the `discover` command to configure a cluster component with a hardware type of `generic`. Use the `generic` type for the following hardware:

- A component that you want to configure as part of the cluster
- A component that has only one IP address associated with it
- A component that is to be treated by the cluster manager as an unmanaged cluster component

One likely use is for Ethernet switches that extend the management network in large configurations. When the `generic` hardware type is used for external management switches on large systems, observe the following guidelines:

- HPE recommends that the management switches are the first hardware discovered in the system.
- HPE recommends that the management switches both start with their power cords unplugged. This state is analogous to how the system discovers leader nodes and compute nodes.
- If your site does not want the cluster manager to assign low numbers to the switches. In this case, explicitly give the external switches high numbers for node numbers.
- As an option, give these switches an alternative host name. Use the `hostname1` option or use the `cadmin` command after the `discover` command runs.

The following example configures two external switches into the cluster:

```
admin:~ # discover --nodeset 98,2,generic
```

## **Configuring a compute node to use a nondefault image**

The following example configures compute node 0 and uses `service-myimage` instead of the default image:

```
admin:~ # /opt/sgi/sbin/discover --node 0,image=service-myimage
```

---

**NOTE:** For information about how to direct a compute node to image itself with a custom image later, without rerunning the `discover` command, see the following:

[\*\*Pushing images from the admin node to the targeted nodes\*\*](#) on page 217

---

## **Skipping a node while configuring**

Assume that you want to configure rack 1, rack 4, and the first compute node. You want to ignore MAC address 00:04:23:d6:03:1c. The command is as follows:

```
admin:~ # /opt/sgi/sbin/discover --ignoremac 00:04:23:d6:03:1c --leader 1 --leader 4 --node 0
```

Alternatively, you can use the `--skip-provision` option of the `discover` command to configure a node yet not provision it.

For more information, see the following:

[\*\*HPE Performance Cluster Manager Installation Guide\*\*](#)

## **Omitting unneeded switch configurations when reconfiguring**

By default, the `discover` command performs top-level switch configuration operations each time it runs. You can direct the `discover` command to omit the switch configuration. If you want to add nodes on a system that is configured, skipping the switch configuration process saves time. By omitting the switch configuration steps, the `discover` command completes its work in less time.

The following procedure explains how to configure a new node for a cluster and skip the switch configuration steps.

### **Procedure**

**1.** Log into the system as root.

**2.** Enter the following command to retrieve the switch configuration status:

```
# cadmin --show-discover-skip-switchconfig
```

The output from this command is one of the following:

- no. The `discover` command is set to perform the switch configuration processing when it runs.
- yes. The `discover` command is set to suppress switch configuration processing when it runs.

**3.** (Conditional) Reset the `discover` command behavior.

Complete this step if both of following conditions are true:

- The previous step returned no.
- Your goal is to suppress switch configuration processing when the `discover` command runs.

Enter the following command:

```
# cadmin --disable-discover-switchconfig
```

Conversely, to enable switch processing, enter the following command:

```
# cadmin --enable-discover-switchconfig
```

**4.** Enter the following `discover` command to configure the new node and omit switch configuration:

```
# cattr set discover_skip_switchconfig yes
```

**5.** Enter the following command to configure the additional node:

```
# discover --rack 1 --configfile myconfigfile
```

6. Enter the following command to set the `discover_skip_switch` value to yes in the database:

```
# cattr set discover_skip_switchconfig yes
```

7. Enter the following command to show the value in the database:

```
# cattr list -g discover_skip_switchconfig
```

## Managing slots

The following topics explain how to manage multiple slots:

- [Retrieving slot information](#) on page 157
- [Booting from a different slot](#) on page 157
- [Cloning a slot](#) on page 158
- [Customizing slot labels](#) on page 159
- [Modifying boot options](#) on page 160

### Retrieving slot information

The following procedure explains the following:

- How to figure out which slot is booted
- How to retrieve information about the slots that are configured currently

#### Procedure

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

2. Enter the following command to verify the current boot slot:

```
# cadmin --show-current-root
admin node currently booted on slot: 1
```

3. Enter the following command to retrieve information about the slots available to be booted:

```
# cadmin --show-root-labels
CD slot 1: CM 1.0 / sles12sp3: Production
    slot 2: CM 1.0 / sles12sp3: Backup for slot 1
CD slot 3: CM 1.0 / sles12sp3: Chris's slot
    slot 4: CM 1.0 / rhel7.4: Do not destroy until June 30 2019
    slot 5: CM 1.0 / rhel7.4: (none)
```

### Booting from a different slot

If you configured more than one slot, you can boot from the boot partition of any slot. The following procedure explains how to change the system to boot from a different slot.

#### Procedure

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

2. Change the default slot.

You can specify the new slot now, or you can specify the new slot during the reboot. This step explains how to change the boot slot now. Enter the `cadmin` command in the following format:

```
cadmin --set-default-root --slot num
```

For *num*, specify the new boot slot number. *num* can be an integer from 1 to 10, inclusive.

For example, to specify a boot from slot 2, enter the following:

```
admin:~ # cadmin --set-default-root --slot 2
```

For information about the operating systems installed in each slot, see the following:

#### [Retrieving slot information](#) on page 157

3. Enter the following command to shut down the entire system:

```
# cpower system shutdown
```

4. Enter the following command to reboot the admin node:

```
# reboot
```

5. Connect to the system console to monitor the reboot.

Optionally, select a nondefault slot from which you want to boot.

During the reboot, the system displays a screen that shows all the available slots and highlights the current boot slot. To select a different boot slot, use the arrow keys to select a new slot and press **Enter**.

If you do not select a new slot, the system boots from the highlighted slot after approximately 10 seconds.

6. Log in as the root user again.

7. Enter the following command to reboot all the rack leaders and compute nodes:

```
# cpower system on
```

If the IP addresses are configured differently within different slots, the `cpower` command might not be able communicate with the BMCs immediately after you reboot the admin node. If you have trouble connecting to the rack leaders and compute node BMCs after you change slots, wait a few minutes. Issue the `cpower` command again. The wait enables the nodes to obtain new IP addresses.

## Cloning a slot

You can clone, or copy, the installation in one slot to a different slot at any time. Before you modify slot images or reconfigure a slot, clone it first. HPE recommends cloning because the cloned copy provides a backup. If you want to revert to the original configuration, you can use the clone.

The cloning process copies the software for the admin node, the leader node, and the compute nodes to the slot you specify. The ICE compute nodes do not participate in the cloning process because they are diskless.

### Procedure

1. Log into the admin node as the root user.
2. Enter the `clone-slot` command in the following format:

```
clone-slot --source source_slot_number --dest destination_slot_number
```

The variables are as follows:

Variable	Specification
<i>source_slot_number</i>	The slot number that contains the configuration you want to clone.
<i>destination_slot_number</i>	The slot number to receive the copy of the configuration.
	<b>NOTE:</b> The cloning process completely destroys all data in slot <i>destination_slot_number</i> .

The `clone-slot` command does the following:

- Synchronizes the data
- Configures the `grub` and `fstab` entries to make the cloned slot bootable.

If the *source\_slot\_number* slot is the mounted, or active, slot, the `clone-slot` command acts as follows:

- It shuts down the cluster database on the admin node before it starts the backup operation.
- It restarts the cluster database when the backup is complete.

The preceding sequence ensures that the cluster database does not change during the cloning operation. The sequence also ensures that there is no data loss.

For more information, enter the following command:

```
# clone-slot --help
```

For example, the following command clones the configuration in slot 1 to slot 2 and overwrites the contents of slot 2:

```
# clone-slot --source 1 --dest 2
```

## Customizing slot labels

After an installation, the slot label is `(none)`. You can use the `cadmin` command to label the slots on a multiple-boot cluster.

### Procedure

1. Log into the admin node as the root user.
2. Enter the following command to retrieve the current labels:

```
admin:~ # cadmin --show-root-labels
slot 1: SMC 3.5.0 / sles12sp3: (none)
slot 2: SMC 3.5.0 / sles12sp3: Backup for slot 1
slot 3: SMC 3.5.0 / sles11sp4: (none)
slot 4: SMC 3.5.0 / rhel7.4: (none)
slot 5: SMC 3.5.0 / rhel7.4: patch 11395
```

3. Enter the following command to specify the slot and the label:

```
cadmin --set-root-label --slot num --label "mylabel"
```

The variables are as follows:

Variable	Specification
<i>num</i>	Use an integer from 1 to 10, inclusive, to specify the slot you want to label.
<i>mylabel</i>	Enter the name you want to apply to the slot.

For example:

```
# cadmin --set-root-label --slot 1 --label "Installed 05/15/18"
# cadmin --show-root-labels
    slot 1: SMC 3.5.0 / sles12sp3: Installed 05/15/18
    slot 2: SMC 3.5.0 / sles12sp3: Backup for slot 1
    slot 3: SMC 3.5.0 / sles11sp4: (none)
    slot 4: SMC 3.5.0 / rhel7.4: (none)
    slot 5: SMC 3.5.0 / rhel7.4: patch 11395
```

## Modifying boot options

You can use the `cadmin` command to set extra kernel boot parameters for the following on a per-image basis:

- ICE compute nodes
- Flat compute nodes
- Leader nodes

For example, assume that you want to add `cgroup_disable=memory` to the kernel boot parameters. You want to add `cgroup_disable=memory` to any node that boots the `ice-sles12sp3` image. Enter the following command:

```
% cadmin --set-kernel-extra-params --image ice-sles12sp3 cgroup_disable=memory
```

To change the boot parameters, issue additional `cadmin` commands. The following additional arguments might be useful to you when you update boot parameters:

- `--show-kernel-extra-params`
- `--unset-kernel-extra-params`
- `--show-nfsroot-extra-params`
- `--set-nfsroot-extra-params`
- `--unset-nfsroot-extra-params`

For information about boot options, see the following:

[HPE Performance Cluster Manager Installation Guide](#)

## Powering on and powering off cluster systems and cluster system components

You can use the `cpower` command to power on or power off all or part of a cluster system. The command allows you to manage the power status of the entire cluster or selected components of the cluster.

The following topics describe how to power off and power on cluster systems:

- [Using the cpower command](#) on page 161
- [Power commands for the entire cluster](#) on page 164
- [Power commands for ICE compute nodes and flat compute nodes](#) on page 165
- [Managing rack leaders](#) on page 166
- [Managing ICE compute IRUs](#) on page 167
- [Managing ICE compute blade switches](#) on page 167

## Using the `cpower` command

The `cpower` command allows you to control power up actions, power down actions, and other actions. The command can also display the power status of system components. The following is the command format:

```
cpower [option] target_type action target_list
```

The parameters are as follows:

- The *option* parameter is optional. Specify one of the following for *option*:

**-h | --help**

Displays the help message. If you enter the `cpower` command without any arguments, the command displays command usage information.

**-i seconds | --interval=seconds**

Specifies how long the identifying LED of the target stays lit. Specify an integer number of seconds. If you specify 0, the LED turns off immediately.

Valid with the `identity` action.

**-p | --poll**

Forces an immediate check for the status of a target. See the description for the `status` action.

Valid with the `status` action.

Valid target types: `leader`, `node`, `system`

**-u | --no-unmatched**

Suppress output messages that report **unmatched targets**. These messages describe names in the target list that do not match any component with the specified target type.

**-v | --verbose**

Reports all details in the command output, including all errors.

**-w seconds | --wait=seconds**

Waits until the specified action on the target completes or until the time specified by *seconds* expires. The *seconds* parameter is required. The command reports its progress as it executes.

Valid with actions `on`, `reset`, and `reboot`.

- The *target\_type* parameter is required. Specify one of the following:

**switch-blade**

Applies the action to the blade switches specified by *target\_list*.

**iru**

Applies the action to the independent rack units (IRUs) specified by *target\_list*. For the `on` and `off` actions, the dependent the blade switches and ICE compute nodes are targeted also.

**leader**

Applies the action to the rack leader nodes specified by *target\_list*.

**node**

Applies the action to the flat compute nodes or ICE compute nodes specified by *target\_list*.

**system**

Applies the action to the entire cluster, excluding the admin node. Do not specify *target\_list* with this target type.

- The *action* parameter is required. Specify one of the following:

**cycle**

Power cycles the target by sending an IPMI `cycle` command.

Valid target types: `leader`, `node`

The `-wait` option is available for this action.

**halt**

Halts the target by issuing a `halt` command through `ssh`.

Valid target types: `leader`, `node`, `system`

If the target type is `system`, flat compute and ICE compute nodes halt first; then, the leaders halt.

**identify**

Turns on the identifying LED of the target for the period specified by the `-i seconds` option.

Valid target types: `leader`, `node`

**off**

Powers off the target by sending an IPMI power off command.

Valid target types: `switch-blade`, `iru`, `leader`, `node`, `system`

If the target type is `system`, compute and ICE compute nodes are powered off first. Then, the leaders are powered off.

If the target type is `iru`, the associated blade switches are also powered off.

**on**

Powers up the target by sending an IPMI power-on command.

Valid target types: `switch-blade`, `iru`, `leader`, `node`, `system`

If the target type is `system`, leaders and compute nodes are powered on first. Then, the ICE compute nodes are powered on.

If the target is an ICE compute node, the `on` action ensures that the associated leader and IRU are on. If the associated leader is off, this action powers it on and waits for its successful boot with a 10-minute timeout. Then, the `on` action powers on the associated IRU if needed and the ICE compute node in turn.

If the target type is `iru`, the associated blade switches are also powered on.

For rack leaders and blade switches, the `on` action only powers on the specified target.

The `-wait` option is available for this action.

#### **reboot**

Reboots the target even if already booted by sending a `reboot` command through ssh.

Valid target types: `leader`, `node`

The `-wait` option is available for this action.

#### **reset**

Performs a hard reset on the target by sending an IPMI `reset` command.

Valid target types: `leader`, `node`

The `-wait` option is available for this action.

#### **shutdown**

Shuts down and power off the target by sending a `shutdown -h now` command through ssh.

Waits for targets to shut down.

Valid target types: `node`, `leader`, `system`

#### **status**

Displays the power status of the target.

Valid target types: `iru`, `leader`, `node`, `system`

For a target type of `node`, parameter *target\_list* is required.

Possible status values for a target:

- `BOOTED` -- The power is on and the target is booted.
- `ON` -- The power is on and the target is not booted.
- `OFF` -- The power is off.
- `UNKNOWN` -- The status for an ICE compute node where its leader either is not running or the `clmgr-power` service is not connected.

By default, the command reads information reported by the heartbeat daemon, described in the following topic:

#### **Heartbeat daemon** on page 259

Most node status transitions are reported by the `cpower` command as they occur. However, sometimes a node moves from `BOOTED` to some unbooted state. In this case, the `cpower` command, by default, might take longer to reflect the changed state. If you use the `--poll` option, you force the command to make an immediate poll of the target and to return a more accurate status. The `--poll` option is only valid with target types `leader`, `node`, and `system`.

- A *target\_list* is required, except when *target\_type* is `system`.

For *target\_list*, specify a comma-separated list of hostnames, IRUs, or blade switches. To display possible targets, use the `cnodes` command or the `discover` command and the cluster definition file.

You can use pattern-matching expressions (wildcards) to specify targets. When you use wildcards, enclose the *target\_list* parameter in quotation marks. The `cpower` command supports **globbing** expressions. The most commonly used expressions are the following:

\*

Matches one or more characters.

Example: "r\*lead" for all rack leaders

?

Matches exactly one character.

Example: "r1i?n\*" for all nodes in rack 1 whose IRU number is a single character

[]

Matches any of the range of characters specified within brackets.

Example: "r1i2n[1-3]" for nodes 1, 2, and 3 in IRU 2 of rack 1

For information about the `discover` command and the cluster definition file, see the following:

#### **HPE Performance Cluster Manager Installation Guide**

## **Power commands for the entire cluster**

To manage the power status of the entire cluster (excluding the admin node), specify the target type of `system` and the desired action on the `cpower` command.

The following are example power management commands for entire clusters.

- The following command powers down the cluster:

```
# cpower system off
```

The compute nodes and ICE compute nodes are powered down first. Then, the rack leaders are powered down.

- The following command powers up the cluster:

```
# cpower system on
leader node rllead power ON
600 sec wait for leader rllead to boot
direct node service0 power ON
leader node rllead is BOOTTED
leader node rllead is BOOTTED
compute node r1i0n0 BOOTTED
compute node r1i0n3 BOOTTED
compute node r1i0n4 BOOTTED
...
compute node r1i2n2 BOOTTED
compute node r1i2n11 BOOTTED
compute node r1i2n4 BOOTTED
compute node r1i2n14 BOOTTED
compute node r1i2n15 BOOTTED
```

The rack leaders and compute nodes are powered on first, followed by the ICE compute nodes.

- The following command queries the power on/off status of the cluster:

```
# cpower system status
service0      BOOTTED
rllead       BOOTTED
r1i0n0      BOOTTED
r1i0n1      BOOTTED
r1i0n2      BOOTTED
...
```

```
r1i3n14      BOOTED
r1i3n15      BOOTED
r1i3n16      BOOTED
r1i3n17      BOOTED
```

## Power commands for ICE compute nodes and flat compute nodes

To manage compute nodes and ICE compute nodes with the `cpower` command, specify the following:

- A target type of `node`
- An action
- A target list

The following are example power management commands for ICE compute nodes and flat compute nodes.

- The following command powers on `service0`, which is a flat compute node:

```
# cpower node on service0
```

- The following command powers on all flat compute nodes:

```
# cpower node on "service*"
```

To manage the boot order of a group of flat compute nodes, not ICE compute nodes, see the following:

[\*\*Changing compute node configuration elements\*\*](#) on page 171

- The following command queries and displays the status of all flat compute nodes:

```
# cpower node status "service*"
```

- The following command powers down one flat compute node:

```
# cpower node off service0
```

- The following command reboots compute node 0 with a three-minute timeout:

```
# cpower node reboot service0 -w 180
```

- The following command powers on the ICE compute node at rack 1, IRU 3, slot 10:

```
# cpower node on r1i3n10
```

If the associated leader node is off, this action powers on components in the following order:

- The leader node itself. The command powers on the leader note and waits for its successful boot. There is a 10-minute timeout.
  - The associated IRU, if needed.
  - The ICE compute node itself.
- The following command powers on a group of ICE compute nodes in a rack:

```
# cpower node on "r1i0n[2-5]" -w 300
cmc node r1i0c power ON
```

```
compute node r1i0n3 already BOOTTED
compute node r1i0n5 power ON
compute node r1i0n4 power ON
compute node r1i0n2 power ON
compute node r1i0n5 is BOOTTED
compute node r1i0n2 is BOOTTED
300 second timeout exceeded waiting for boot of r1i0n4
```

The command powers on and attempts to boot the ICE compute nodes in slots 2, 3, 4, and 5 in IRU 0 of rack 1. Note the 5-minute wait time for booting.

- The following command queries the status of all ICE compute nodes in a rack:

```
# cpower node status "r1i*n*"
```

- The following command powers off the specified ICE compute node.

```
# cpower node off r1i3n10
```

The associated rack leader and IRUs are unaffected.

- The following command reboots the specified ICE compute node:

```
# cpower node reboot r1i3n10
```

- The following command turns on the ID LED of an ICE compute node for 60 seconds:

```
# cpower node identify r1i3n10 -i 60
```

## Managing rack leaders

Rack leader power management requires you to use the `cpower` command to specify a target type of leader, an action, and a target list.

The following are example power management commands for rack leaders.

- The following command powers on the leader for rack 1:

```
# cpower leader on r1lead
```

- The following command shuts down the specified rack leader:

```
# cpower leader shutdown r3lead
```

```
leader node r3lead has been issued a shutdown -h now command
leader node r3lead is DOWN
```

The associated ICE compute nodes and IRUs are unaffected.

- The following command queries and then displays the status of all rack leaders:

```
# cpower leader status "*"
```

```
r1lead      BOOTED
r2lead      BOOTED
r3lead      OFF
```

- The following command reboots a rack leader:

```
# cpower leader reboot r3lead -w 180
```

There is a three-minute timeout.

- The following command turns on the ID LEDs of all the rack leaders for 60 seconds:

```
# cpower leader identify "r*lead" -i 60
```

## Managing ICE compute IRUs

IRU power management requires you to use the `cpower` command with the target type of `iru`, an action, and a target list. Specify an IRU by its rack number and its IRU number. For example, `r1i1` specifies IRU 1 on rack 1.

Powering on an ICE compute node powers on its associated leader and IRU, but the converse is not true. Likewise, powering on/off an IRU powers on/off its associated blade switches and ICE compute blades.

The following are example power management commands for IRUs.

- The following command powers on IRU 0 in rack 1:

```
# cpower iru on r1i0
```

- The following command powers off IRU 1 in rack 3 plus the associated blade switches and ICE compute nodes:

```
# cpower iru off r3i1
```

- The following command powers off all IRUs, blade switches, and ICE compute nodes in rack3:

```
# cpower iru off "r3i*"
```

Notice the use of quotation marks (" ") with the wildcard to ensure that a matching filename is not targeted.

## Managing ICE compute blade switches

Like IRUs, you can manage blade switches selectively. You can turn them on and off and query their power status. For blade switches, use the `cpower` command with the following:

- A target type of `switch-blade`
- An action
- A target list

Specify a blade switch by its switch number and its associated rack and IRU. For example, `r1i0s0` specifies switch 0 associated with IRU 0 on rack 1.

The following are example power management commands for blade switches.

- The following command powers on blade switch 0 associated with IRU 0 in rack 1:  

```
# cpower switch-blade on r1i0s0
```
- The following command powers off blade switch 1 associated with IRU 1 in rack 3:  

```
# cpower switch-blade off r3i1s1
```
- The following command returns the status of all blade switches:  

```
# cpower switch-blade status "*"
r1i0s0      ON
r1i0s1      ON
r1i1s0      ON
r1i1s1      ON
r1i2s0      ON
r1i2s1      ON
r1i3s0      ON
r1i3s1      ON
```

## Power and energy management

Power management includes the following:

- Monitoring energy use
- Managing energy use at the system level and at the job level
- Managing the thermal health of the cluster

For power management information, see the following:

[\*\*HPE Performance Cluster Manager Power Management Guide\*\*](#)

## pdsh and pdcp commands

The `pdsh` command is the parallel shell utility. The `pdcp` command is the parallel copy/fetch utility. The cluster manager populates some `pdsh` group files for the various node types. On the admin node, the cluster manager populates the `leader` and `compute` group files with the list of online nodes in each of those groups.

On a leader node, the cluster manager populates the `ice-compute` group with the list of all the online ICE compute nodes in that rack.

On a flat compute node, the cluster manager populates the `compute` group with the list of all the online flat compute nodes in the whole system.

The following topics explain how to use the `pdsh` and `pdcp` commands:

- [\*\*pdsh command examples\*\*](#) on page 169
- [\*\*Creating custom pdsh group files\*\*](#) on page 169

For more information, see the `pdsh(1)` and `pdcp(1)` manpages.

## **pdsh command examples**

The following are pdsh command examples:

- Example 1. From the admin node, to run the `hostname` command on all the leader nodes, enter the following:

```
# pdsh -g leader hostname
```

- Example 2. To display the hostname of all ICE compute nodes in the cluster, enter the following:

```
admin_node # pdsh -g ice-compute hostname
```

The preceding command runs the `hostname` command on all the ICE compute nodes.

If the preceding command does not work, verify that the routing information protocol (RIP) is enabled on the management switch. The RIP protocol is enabled by default, but it is possible that the protocol has been disabled. To run `pdsh` commands on all ICE compute nodes from the admin node, the RIP protocol must be enabled on the management switches.

For example, to retrieve the status of the RIP protocol on `mgtsw1`, enter the following command:

```
# switchconfig rip -i -s mgtsw1
```

To set the RIP protocol on `mgtsw1`, enter the following command:

```
# switchconfig rip -e -v all -s mgtsw1
```

- Example 3. To run the `hostname` command on just `r1lead` and `r2lead`, enter the following:

```
# pdsh -w r1lead,r2lead hostname
```

## **Creating custom pdsh group files**

Depending on your cluster configuration and how you use it, you might find it useful to create custom group files. To create a custom group, do the following:

1. Create a text file containing all desired nodes, one per line.
2. Choose a filename that matches the desired group name.
3. Place the text file in directory `/etc/dsh/group`.

For example, assume that you have a Lustre file system. You might want the following groups:

- A group that contains only MDS nodes
- A group that contains only OSS nodes
- A group that contains all MDS and OSS nodes

If the MDS nodes are `mds1` and `mds2` and the OSS nodes are `oss[1-6]`, you could define the following groups:

- Define group `mds` in file `/etc/dsh/group/mds` as follows:

```
mds1  
mds2
```

- Define group `oss` in file `/etc/dsh/group/oss` as follows:

```
oss1  
oss2  
oss3  
oss4  
oss5  
oss6
```

- Define group `lustre` in file `/etc/dsh/group/lustre` as follows:

```
mds1  
mds2  
oss1  
oss2  
oss3  
oss4  
oss5  
oss6
```

## Using the `cadmin` command, the administrative interface

After you log into the admin node, you can use the `cadmin` command to administer the cluster.

The following topics include examples that show how to use the `cadmin` command:

- [Bringing a node online or setting a node offline](#) on page 170
- [Creating and displaying node-specific notes](#) on page 171
- [Changing compute node configuration elements](#) on page 171
- [Changing the admin node hostname and IP address on the house network](#) on page 172
- [Displaying network information](#) on page 173
- [Changing switch management network settings](#) on page 174
- [Changing console management settings](#) on page 174
- [Managing UDP multicast \(UDPcast\) provisioning](#) on page 175

For information about how to preserve custom configuration changes across `update-configs` calls, see the following:

[Preserving custom configuration changes](#) on page 309

For information about the `cadmin` command, see the `cadmin(8)` manpage.

### Bringing a node online or setting a node offline

The following examples show how to bring a node online or set a node offline:

- To set `r1i0n0` offline, enter the following command:  
`# cadmin --set-admin-status --node r1i0n0 offline`
- To set `r1i0n0` online, enter the following command:  
`# cadmin --set-admin-status --node r1i0n0 online`

When you set the node administrative status to offline, the cluster manager changes the node status in the following:

- The cluster manager database
- Configuration files that depend on the database

The cluster manager ignores the node for all subsequent actions targeting online nodes.

## Creating and displaying node-specific notes

You can use the `cadmin` command to create and display node-specific notes in the cluster manager database. For example, you might want to document why you placed a node offline. Use the following command formats to do so:

- `cadmin --set-node-notes --node nodes "text"`
- `cadmin --show-node-notes --node nodes`

Example:

```
# cadmin --set-node-notes --node n1 "Booting problems May 3. Put offline."
# cadmin --show-node-notes --node n1
Booting problems May 3. Put offline.
```

## Changing compute node configuration elements

The following examples show how to change the hostname and IP address of a compute node.

- To change the hostname of `service0` to `myservice`, enter the following command:  
`admin:~ # cadmin --set-hostname --node service0 myservice`
- To retrieve the IP addresses currently configured for `myservice`, enter the following command:  
`admin:~ # cadmin --show-ips --node myservice
IP Address Information for SMC node: service0`  

ifname	ip	Network
myservice-bmc	172.24.0.3	head-bmc
myservice	172.23.0.3	head
myservice-ib0	10.148.0.254	ib-0
myservice-ib1	10.149.0.67	ib-1
myhost	172.24.0.55	head-bmc
myhost2	172.24.0.56	head-bmc
myhost3	172.24.0.57	head-bmc

- To change the IP address on `myservice-ib0`, enter the following command:

```
admin:~ # cadmin --set-ip --node myservice --net head myservice=172.23.0.199
```

- To set the boot order for compute node `myservice`, enter the following command:

```
# cadmin --set-boot-order --node myservice priority
```

For `priority`, specify any positive integer number. The default is 1. This value is the boot order specification.

You can use the `cadmin` command to control the boot order (boot sequence) of a group of flat compute nodes. You can implement this kind of control to ensure that the server nodes boot before clients nodes. For example, you might want to use this feature with NFS, CIFS, or SMB servers.

When you boot a group of compute nodes with varying boot order values, the cluster manager first boots all nodes with boot order 1. Then the cluster manager boots those nodes with boot order 2, and so on.

Some power-down operations honor a specified boot order. These operations power down the compute nodes starting with those operations that have the largest boot order number. The power-down operations that respect boot order are `off`, `shutdown`, and `halt`.

The `reboot`, `reset`, and `cycle` operations do not respect boot order. These operations act on all target flat nodes simultaneously.

For information about power-on operations and power-down operations, see the following:

**Powering on and powering off cluster systems and cluster system components** on page 160

To make large-scale configuration changes, use the `discover` command and a cluster definition file. For more information, see the following:

**HPE Performance Cluster Manager Installation Guide**

## Changing the admin node hostname and IP address on the house network

The procedure in this topic explains the following:

- How to retrieve information about the admin node
- How to update the admin node hostname or IP address

The examples show how to change the address information for the admin node on the house network.

### Procedure

1. Log into the admin node as the root user.
2. Use the `cadmin` command to retrieve information about the current house network IP address.

For example:

```
admin:~ # cadmin --show-house-network-info
-----Network Information-----
broadcast      :          137.38.82.255
base_ip        : 137.38.82.0          # the IP of the house network
netmask        : 255.255.255.0
gateway        : 137.38.82.254
ip             : 137.38.82.166          # the IP address of the admin node
```

3. Use the `cadmin` command in the following format to assign a new IP address to the admin node:

```
cadmin --set-house-network ip_addr,netmask,gateway_info
```

The variables are as follows:

Variable	Specification
<i>ip_addr</i>	The new IP address that you want to assign to the admin node.
<i>netmask</i>	The network mask you want to assign to the new IP address.
<i>gateway_info</i>	Either the default gateway you want to assign to the new IP address or the keyword <code>no_gateway</code> .

You can also use the `cadmin --set-house-network` command to specify a new network mask or new gateway information for the admin node. In that case, specify the existing admin node IP address, the new network mask, and/or the new default gateway.

For example, to change the hostname associated with the admin node to be `newname`, enter the following command:

```
admin:~ # cadmin --set-hostname --node admin newname
```

4. Use the `service network restart` command to restart the network services.

When you use the `--set-house-network` parameter to the `cadmin` command to change any of the networking information, restart network services.

The following examples show how to use the `service network restart` command:

Example 1. On a RHEL 7 or SLES 12 admin node, enter the following:

```
admin:~ # cadmin --set-house-network 137.38.82.165,255.255.255.0,137.38.82.253
admin:~ # systemctl network restart
```

Example 2. On a RHEL 7 or SLES 12 admin node, enter the following:

```
admin:~ # cadmin --set-house-network 137.38.82.165,255.255.255.0,no_gateway
admin:~ # systemctl network restart
```

## Displaying network information

The following examples show how to use the `cadmin` command to display network information.

- To set and show the cluster subdomain, enter the following commands:

```
admin:~ # cadmin --set-subdomain mysubdomain.domain.mycompany.com
admin:~ # cadmin --show-subdomain
The cluster subdomain is: mysubdomain
```

- To retrieve the admin node house network domain, enter the following command:

```
admin:~ # cadmin --show-admin-domain
The admin node house network domain is: domain.mycompany.com
```

## Changing switch management network settings

The following examples show how to use the `cadmin` command to change the switch management network settings.

- To retrieve the current switch management value for a specified node, enter the following command:

```
admin:~ # cadmin --show-switch-mgmt-network --node admin
no
```

In this example, returned value is `no`. This value means that there is no switch management network. This configuration is a nondefault configuration.

- To enable the switch management network for a specified node that is connected to managed top-level switches, enter the following command:

```
admin:~ # cadmin --enable-switch-mgmt-network --node admin
```

- To disable the switch management network for a specified node that is connected to managed top-level switches, enter the following command:

```
admin:~ # cadmin --disable-switch-mgmt-network --node admin
```

## Changing console management settings

If you have hundreds of flat compute nodes connected to the system, console logging and the number of active IPMI processes can affect performance.

To avoid excessive console logging and `ipmitool` processing, use the `cadmin` command to suppress console logging and reduce the number of active IPMI processes.

The following `cadmin` command parameters control console logging and the number of active IPMI processes:

- Console logging. Console logging is enabled by default.

The following parameters affect console logging:

- `--show-conserver-logging`
- `--set-conserver-logging`. You can set this value on a global basis or on a per-node basis. This setting affects ICE compute nodes tied to leaders.

- Console on demand. Console on demand is disabled by default.

This feature allows IPMI to connect to the BMC to access the console when there is an active console session through the `console` command. For this feature to work, console logging must be enabled.

The following parameters affect the console on-demand setting:

- `--show-conserver-on-demand`
- `--set-conserver-on-demand`

## Managing UDP multicast (UDPcast) provisioning

The cluster manager supports UDP multicast provisioning, which allows you to quickly install hundreds of compute nodes at once. UDPcast allows many nodes to join a multicast stream of the content being transported. With all the nodes sharing a single stream, the network is protected from being saturated by disjoint installations. Regardless of cluster type or node type, UDPcast is the default transport method for provisioning.

The following topics explain UDPcast provisioning:

- [\*\*UDPcast overview\*\*](#) on page 175
- [\*\*UDPcast configuration tuning\*\*](#) on page 176

For more information about various transport methods, see the following:

### [\*\*HPE Performance Cluster Manager Installation Guide\*\*](#)

## UDPcast overview

UDPcast is the basic tool used for multicast installation. It has two primary commands:

- `udp-sender`. Sends a single image stream to one or more receivers.
- `udp-receiver`. Issued by the recipients to listen to the stream.

The following is additional information about UDPcast:

- Flamethrower

Flamethrower is a wrapper program. The cluster manager uses Flamethrower to manage UDPcast content when installing systems and pushing images.

It maps `udp-sender` commands to content to be transported. It starts a `udp-sender` on a unique port for each component to be transported. When `udp-sender` terminates (due to a transfer being complete), Flamethrower starts a new one.

The content managed by Flamethrower includes the Flamethrower directory itself, the system imager boot environment, and any available images. For each image, there are two components: the image itself and the overrides associated with the image.

On a system with three images, there are typically 10 different pieces of content to manage, each with a dedicated `udp-sender` process running on a unique port.

On the admin node, `udp-sender` is run in tar-pipe mode, which means the image is run through tar through a pipe. Separate tar files for each image do not need to be maintained. What is being transported is always the current image.

- Flamethrower directory

All of the content managed by Flamethrower is listed in the Flamethrower directory. The directory contains a module file for each piece of content that is to be sourced by Bash.

When a node is interested in multicast content, it first uses `udp-receiver` to transfer the Flamethrower directory. Once the node has the directory, it has the list of components to transport and the port numbers to use. It then uses `udp-receiver` to transfer the desired content.

- Management Ethernet

The management Ethernet switches must be configured to properly handle multicast traffic. Switches that are supported and configured by the cluster manager are likely to be configured correctly

automatically. Switches that are not configured by the cluster manager must be configured to transport multicast traffic.

The multicast IP addresses are adjustable for the RDV address (the address used for nodes to find each other). The data transport IP addresses are not configurable. The admin node uses 239.0.0.1 by default for RDV, which often requires special switch configuration to work properly. The leader nodes serve the ICE compute nodes. The leader nodes use 224.0.0.1 for RDV by default.

For more information about these IP addresses and configuration adjustments, see the following:

#### **UDPcast configuration tuning** on page 176

- Node memory used for flat compute and leader nodes

Flat compute (service) nodes and leader nodes installed using UDPcast must have enough system memory to hold the image. The image is stored in to a `tmpfs` file system on the node during installation to make the transport more efficient. With hundreds of nodes listening to a stream, writing the data directly to disk would slow down the transfer for all nodes. For this reason, the data is saved to `tmpfs` first and then expanded onto the system disk. If you have nodes with little memory, UDPcast installation could fail for this reason.

- Node memory used for ICE compute nodes in `tmpfs` mode

The UDP receiver is used in tar-pipe mode. That is, the files are expanded from a pipe directly to the `tmpfs` file system. The `tmpfs` file system is used as the root file system.

## **UDPcast configuration tuning**

This topic describes settings you can fine-tune to optimize UDPcast performance. The goal is to get most nodes to listen to a stream at the same time. Various settings affect the wait time for neighbors to join. It is acceptable for nodes to join different streams. The UDP receiver waits for the current stream to complete and joins when a new stream starts. In this case, some nodes can grab the first stream and other nodes can join the second. You can tune the following attributes:

- `flamethrower-directory-portbase`

The `flamethrower_directory_portbase` attribute is the port number for the Flamethrower directory itself. This directory is important because all nodes need access to the Flamethrower directory to find the appropriate port number for pertinent content. This port number is provided as a kernel parameter for flat compute (service) and leader nodes when using the UDPcast transport as well as ICE compute nodes when in `tmpfs` mode. The default is 9000.

---

**NOTE:** A technical support representative can help you using the `cattr` command to adjust the value if necessary.

---

- `udpcast-min-receivers`

This attribute defines the minimum number of receivers that must be present before a UDP sender can start a stream. The admin node uses this value when it serves flat compute and leader nodes. The leader nodes that serve ICE compute nodes use this global value in `tmpfs` mode. You can use the `cadmin` command to change this value.

For more information, see the following:

- The `udpcast-max-wait` attribute. A description of this attribute appears later in this topic.
- The `udp-sender` manpage.
- The `cadmin` manpage. See the `--set-udpcast-min-receivers` and `--show-udpcast-min-receivers` parameters.
- `udpcast-min-wait`

The `udpcast-min-wait` attribute defines the minimum time that the UDP sender waits before starting a given stream. The UDP sender waits the minimum time for `udpcast-min-receivers` receivers (described earlier) to join the stream. The admin node uses this value when it serves flat compute and leader nodes. The leader nodes that serve ICE compute nodes use this global value in `tmpfs` mode.

For more information, see the following:

- The `udpcast-min-receivers` and `udpcast-max-wait` attributes. Descriptions of these attributes appear later in this topic.
  - The `udp-sender` manpage.
  - The `cadmin` manpage. See the `--set-udpcast-min-wait` and `--show-udpcast-min-wait` parameters.
  - `udpcast-max-wait`
- The `udpcast-max-wait` attribute defines the maximum time a UDP sender waits before starting a stream. If the minimum number of receivers have not joined by this time, the stream starts anyway. The admin node uses this value when it serves flat compute and leader nodes. The leader nodes that serve ICE compute nodes use this global value in `tmpfs` mode.
- For more information, see the following:
- The `udp-sender` manpage.
  - The `cadmin` manpage. See the `--set-udpcast-max-wait` and `--show-udpcast-max-wait` parameters.
  - `udpcast-max-bitrate`
- The `udpcast-max-bitrate` attribute defines the stream bit rate that a UDP sender attempts to achieve. If the bit rate is too fast, the result is an excessive number of retransmits and retries. The default is 900m. The admin node uses this value when it serves flat compute and leader nodes. The leader nodes that serve ICE compute nodes use this global value in `tmpfs` mode.
- For more information, see the following:
- The `udp-sender` manpage.
  - The `cadmin` manpage. See the `--set-udpcast-max-bitrate` and `--show-udpcast-max-bitrate` parameters.
  - `udpcast-mcast-rdv-addr`

The `udpcast-mcast-rdv-addr` attribute is an IP address. Senders and receivers use this IP address to find each other (rendezvous).

This setting affects switch configuration. If the cluster includes switches that were not configured by HPE tools, ensure the following:

- Multicast traffic must be properly routed inside the switches.
- Multicast traffic must be properly routed between the spine switches and the leaf switches.

The default RDV addresses are as follows:

- 239.0.0.1. The admin node, flat compute nodes, and leader nodes use this address when pushing images for the first time. This address is used because 224.0.0.1 does not cross switch VLANs.
- 224.0.0.1. Leader nodes that serve ICE compute nodes in `tmpfs` boot mode use this address, which is the default. The default is suitable in this case because VLAN crossing is not necessary.

---

**NOTE:** If you adjust the `udpcast-mcast-rdv-addr` value, you might need to adjust the `udpcast-rexmit-hello-interval` attribute.

The `udpcast-mcast-rdv-addr` value takes effect on the leader nodes after the `cimage` command pushes (or repushes) files from the admin node. The image push process reconfigures Flamethrower and the node boot files on leader nodes.

---

The `udpcast-mcast-rdv-addr` value resides in the network boot files of nodes that are being booted or installed with UDPcast. The `udpcast-mcast-rdv-addr` value on the nodes must match the value on the server. To adjust this value, use the `cadmin` command.

For more information, see the following:

- The `udp-sender` manpage.
- The `cadmin` manpage. See the `--set-udpcast-mcast-rdv-addr` and `--show-udpcast-mcast-rdv-addr` parameters.
- `udpcast-rexmit-hello-interval`

The `udpcast-rexmit-hello-interval` attribute defines how often a UDP sender process sends a hello packet. This value is especially important when the RDV address is not 224.0.0.1. Remember that the admin node, for example, defaults to 239.0.0.1 for UDP sender processes.

When a UDP receiver process starts for an RDV address other than 224.0.0.1, the operating system sends an IGMP packet. The Ethernet switch detects this packet. The Ethernet switch then updates its tables with this information. This action allows the multicast packets to properly route through the switch. A problem can arise if the UDP receiver sends its connection packet before the switch updates the switch routing. In this case, the UDP receiver waits forever for a UDPcast stream.

When you set a `udpcast-rexmit-hello-interval` value, the UDP sender sends a hello packet at regular intervals and UDP receivers respond to it. In this way, if the UDP receiver missed the initial packet, the UDP receiver sends a fresh request after seeing the hello packet from the UDP sender.

By default, for admin node UDP senders, this value is 5000 (5 seconds). By default, on leader nodes, this value is 0 (disabled). On leader nodes, this value typically does not need to be set. The RDV address is 224.0.0.1, and there are no VLANs being crossed. If you change the RDV address used by leader nodes, also adjust the `udpcast-rexmit-hello-interval` value. To adjust this value, use the `cadmin` command.

For more information, see the following:

- The `udp-sender` manpage.
- The `cadmin` manpage. See the `--set-udpcast-rexmit-hello-interval` and `--show-udpcast-rexmit-hello-interval` parameters.

---

**NOTE:** You can adjust UDPcast settings using the `cadmin` or `cattr` command. After doing so, push the new images to the leaders. This action ensures the following:

- Flamethrower on the leader nodes that serve ICE compute nodes is set up, and the needed UDP sender processes are running on the designated ports.
- The ICE compute node `tmpfs` network boot files have the appropriate configuration details.

---

## Console management

Clusters use the open-source console management package called `conserver` to perform the following functions:

- Manage the console devices of all managed nodes in a cluster. The `conserver` package allows all consoles to be accessed from the admin node.
- Console logging. These logs can be found at `/var/log/consoles` on the admin node and leader nodes. An `autofs` configuration file lets you access leader-node-managed console logs from the admin node. This file is as follows:

```
admin # cd /net/r1lead/var/log/consoles/
```

A `conserver` daemon runs on the admin node and the leader nodes. The admin node manages leader node and compute node consoles. The leader nodes manage blade consoles. The `conserver` daemon uses `ipmitool` to connect to the consoles. Users connect to the daemon to access them. Multiple users can connect, but nonprimary users are read-only.

The `/etc/conserver.cf` file is the configuration file for the `conserver` daemon.

For both the admin node and the leader nodes, the following script on the admin node generates the `/etc/conserver.cf` file:

```
/opt/sgi/sbin/generate-conserver-files
```

This script is called from the `discover-rack` command as part of rack discovery or rediscovery. The script generates `conserver.cf` for the racks and for the admin node.

---

**NOTE:** The `conserver` package replaces `cconsole` for access to all consoles (blades, leader nodes, managed compute nodes).

---

For more information about `conserver`, see the following manpages:

- `console(1)`, a console server client program.
- `conserver(8)`, the console server daemon.
- `conserver.cf(5)`, the console configuration file for `conserver`.
- `conserver.passwd(5)`, user access information for `conserver`.

For information about `conserver`, see the following:

<http://www.conserver.com/>

To use the `conserver` console manager, perform the following steps:

### Procedure

1. Enter the following command to see the list of available consoles:

```
admin:~ #console -x
  service0          on /dev/pts/2           at Local
  r2lead            on /dev/pts/1           at Local
  r1lead            on /dev/pts/0           at Local
  r1i0n8            on /dev/pts/0           at Local
  r1i0n0            on /dev/pts/1           at Local
```

2. Connect to the service console and enter the appropriate login when prompted.

For example:

```
admin:~ # console service0
```

3. Connect to the leader node console and enter the appropriate login when prompted.

For example:

```
admin:~ # console r1lead
```

4. Enter the following to trigger system request commands `sysrq` (after you connect to a console):

```
Ctrl-e c l 1 8          # set log level to 8
Ctrl-e c l 1 <sysrq cmd>      # send sysrq command
```

5. Enter the following to display a list of `conserver` escape keys:

```
Ctrl-e c ?
```

## Synchronizing system time

Network time protocol (NTP) is the primary mechanism that keeps the cluster nodes synchronized.

The following topics describe how this mechanism operates on the cluster components:

- [Admin node NTP](#) on page 180
- [Leader node NTP](#) on page 181
- [BMC setup with NTP](#) on page 181
- [Compute node NTP](#) on page 181
- [ICE compute node NTP](#) on page 181
- [NTP workarounds](#) on page 181

## Admin node NTP

During the system configuration process, the `configure-cluster` command guides you through the process of setting up NTP on the admin node. Typically, the admin node NTP client points to the house network time server.

The NTP server provides NTP service to system components. Each node uses the server when it boots. The admin node sends NTP broadcasts to some networks to keep the nodes in sync after they boot.

## Leader node NTP

The NTP client on the leader node gets time from the admin node when it is booted. The leader node NTP client stays in sync by connecting to the admin node for time.

The NTP server on the leader node provides NTP service to ICE compute nodes. The ICE compute nodes sync their time when they boot. The leader node sends NTP broadcasts to some networks to keep the ICE compute nodes in sync after they boot.

## BMC setup with NTP

The BMC controllers on managed compute nodes, ICE compute nodes, and leader nodes are also kept in sync with NTP. You might need the latest BMC firmware for the BMCs to sync with NTP properly. The NTP server information for the BMCs is provided by special options stored in the DHCP server configuration file.

## Compute node NTP

When it boots, the NTP client on a managed compute node sets its time from the admin node. Managed compute nodes listen to NTP multicast transmissions from the admin node to stay in sync. These nodes do not provide any NTP service.

## ICE compute node NTP

When it boots, the NTP Client on an ICE compute node sets its time from the leader node. It listens to NTP multicast transmissions from the leader node to stay in sync.

## NTP workarounds

During the initial installation and configuration of a hierarchical system, the NTP services might not be available. In this situation, NTP cannot serve time to system components.

A standard NTP server, running for the first time, takes quite some time before it offers service. In this situation, the leader nodes and compute nodes might not get time from the admin node as they come online. ICE compute nodes might also fail to get time from the leader node when they first come up. After the `ntp` servers have a chance to create their drift files, `ntp` servers offer time with far less delay on subsequent reboots.

The following workarounds are in place for situations when NTP cannot serve the time:

- The admin node and leader nodes have the `time` service enabled (`xinetd`).
- All system node types have the `netdate` command.
- A special startup script is on leader, compute, and ICE compute nodes. This script runs before the NTP startup script.

This script attempts to get the time using the `ntpdate` command. The `ntpdate` command might fail because the NTP server it is using is not ready yet to offer time service. In this case, it uses the `netdate` command to get the clock close.

The `ntp` startup script starts the NTP service as normal. Because the clock is known to be close, NTP fixes the time when the NTP servers start offering time service.

# Booting leaders or flat compute nodes from a local disk

By default, flat compute nodes (including leaders) boot over the network using the GRUB 2 bootloader and miniroot from the admin node.

The cluster manager allows you to boot a flat compute node from a local disk. For example, you can use this feature in the following situations:

- The node has local images with kernels that are not registered on the admin node.
- You want to boot the node independently from the admin node.

Unless you have a compelling reason to do otherwise, HPE recommends that you use the default boot mode for flat compute nodes. There are multiple reasons for this recommendation. For one, regardless of the boot-from-disk mode, the cluster manager does not provide any kernel parameter management capability. You must boot as you would with a standalone system. Secondly, the boot-from-local-disk feature is not supported on MD RAIDs. The boot-from-local-disk feature is supported for the following disk configurations:

- Physical disks
- Hardware-defined RAIDs
- BIOS SW RAIDs where BIOS can find the boot sector storing the GRUB 2 bootloader

The cluster manager supports disjoint boot and admin-assisted boot for booting from a local disk. The following topics describe these modes:

- [\*\*Disjoint boot mode\*\*](#) on page 182
- [\*\*Admin-assisted boot mode\*\*](#) on page 182

For a general description of the default boot process for flat compute nodes, see the following:

[\*\*Booting a flat compute node or a leader node on an installed cluster\*\*](#) on page 290

## Disjoint boot mode

In a disjoint boot, the node boots without retrieving the bootloader or the miniroot from the admin node. This boot could be useful if the network or the admin node is down. By default, the node uses the on-disk GRUB 2 bootloader and boots the most recently installed slot. In an environment with multiple root slots, a menu lets you choose a different slot from the console.

To select disjoint boot mode, adjust the boot order in BIOS to select booting from a disk. If you reinstall the node, change the BIOS boot order back to select booting from the network. On some platforms, you can use the `ipmitool` command `chassis bootdev pxe`. On UEFI platforms, you can use the `efibootmgr` command.

## Admin-assisted boot mode

In admin-assisted boot mode, the cluster manager sends the GRUB 2 bootloader to the node but does not send or use the miniroot. The bootloader chain loads to the appropriate boot partition, and then it instructs the node to boot from disk.

To select admin-assisted boot mode, use one of the following methods:

- Use the following `cadmin` command specification:  
`cadmin --enable-disk-bootloader --node node`
- Specify the following parameter on the `discover` command or in the `discover` configuration file:  
`disk_bootloader=yes`

To disable admin-assisted boot mode, use one of the following methods:

- Use the following `cadmin` command:  
`cadmin --disable-disk-bootloader --node node`
- Specify the following parameter on the `discover` command or in the `discover` configuration file:  
`disk_bootloader=no`

To display the value of the current disk bootloader, use the following command:

```
cadmin --show-disk-bootloader --node node
```

---

**NOTE:** If a node has been marked for installation, the cluster manager supersedes this boot mode with its normal over-the-network boot operation. On noninstallation boots, however, the cluster manager honors the boot mode specification.

---

## Changing the size of `/tmp` on ICE compute nodes

The following procedure explains how to change the size of `/tmp` on ICE compute nodes.

### Procedure

1. From the admin node, use the `cd` command to change to the following directory:

```
/opt/sgi/share/per-host-customization/global
```

2. Open the `sgi-fstab.sh` file.
3. Change the `size=` parameter for the `/tmp` mount in both locations that it appears.
4. Push the image out to the racks to pick up the change.

For example:

```
# cimage --push-rack --customizations-only sgi-fstab.sh image_name "r*"
```

For more information about using the `cimage` command, see the following:

[Using the `cimage` command to manage ICE compute node images](#) on page 225

## Switching ICE compute nodes to a `tmpfs` root

You can use a `/tmpfs` root with an ICE compute node that has 4GB of memory or more. A standard `/tmpfs` mount has access to half the system memory. The standard ICE compute node image is just over 1 GB in size.

The following commands are useful if you want to change ICE compute nodes to use a `tmpfs` root file system:

- To view the current root setting of an ICE compute node, enter the following:

```
admin:~ # cimage --show-nodes r1i0n0
r1i0n0: ice-sles12sp3 2.6.27.19-5-smp tmpfs
```

- To switch ICE compute nodes to a `tmpfs` root, use the optional `--tmpfs` parameter to the `cimage --set` command.

For example:

```
admin:~ # cimage --set --tmpfs ice-sles12sp3 2.6.27.19-5-smp r1i0n0
```

- To set an ICE compute node back to an NFS root, use the `--nfs` parameter to the `cimage --set` command.

For example:

```
admin:~ # cimage --set --nfs ice-sles12sp3 2.6.27.19-5-smp r1i0n0
```

## Switching flat compute nodes to a `tmpfs` root

You can boot flat compute nodes with `tmpfs` root file system. The benefit from configuring this capability is that the `tmpfs` file system does not incur root file system access-to-disk latency.

The drawback to this capability relates to memory. The memory that hosts the root file system in `tmpfs` mode cannot be freed for applications that allocate memory. Compute nodes that boot with the disk root file system make more memory available to applications.

The following procedures explain how to configure this capability:

- [Configuring flat compute nodes to boot a `tmpfs` root file system on RHEL 7 nodes](#) on page 184
- [Configuring flat compute nodes to boot a `tmpfs` root file system on SLES 12 nodes](#) on page 186

## Configuring flat compute nodes to boot a `tmpfs` root file system on RHEL 7 nodes

The following procedure explains how to configure flat compute nodes to use a `tmpfs` root file system. This procedure applies to nodes that run RHEL 7.

### Procedure

1. Use the `pdsh` command to retrieve information about the file systems that the cluster manager used to boot the nodes.

For example:

```
[root@cladmin ~]# pdsh -g compute df -h | dshbak -c
-----
n[01,03-06]
-----
```

```

Filesystem      Size  Used  Avail  Use%  Mounted on
/dev/sda33     110G  2.0G  102G   2%   /
devtmpfs        7.7G   0    7.7G   0%   /dev
tmpfs          7.7G  4.0K  7.7G   1%   /dev/shm
tmpfs          7.7G  9.0M  7.7G   1%   /run
tmpfs          7.7G   0    7.7G   0%   /sys/fs/cgroup
/dev/sda13     283M 110M  158M  41%  /boot
tmpfs          1.6G   0    1.6G   0%   /run/user/0
-----
n02
-----
Filesystem      Size  Used  Avail  Use%  Mounted on
/dev/sdb33     110G  1.9G  102G   2%   /
devtmpfs        7.7G   0    7.7G   0%   /dev
tmpfs          7.7G  4.0K  7.7G   1%   /dev/shm
tmpfs          7.7G  9.0M  7.7G   1%   /run
tmpfs          7.7G   0    7.7G   0%   /sys/fs/cgroup
/dev/sdb13     283M 110M  158M  41%  /boot
tmpfs          1.6G   0    1.6G   0%   /run/user/0

```

**2. Use the `cinstallman` command to specify that selected nodes boot with the `tmpfs` file system.**

For example:

```

[root@cladmin ~]# cinstallman --set-rootfs tmpfs --node n*
Setting up network boot for service0
Setting up network boot for service1
Setting up network boot for service2
Setting up network boot for service3
Setting up network boot for service4
Setting up network boot for service5
cladmin: update-configs: updating SMC configuration files
cladmin: update-configs: -> cminfo
cladmin: update-configs: 0.35s
cladmin: update-configs: Configuration files generated. 0.35s
cladmin: update-configs: -> updating admin node
Number of files transferred: 0
cladmin: update-configs: Configuration pushed to admin. 1.66s
cladmin: update-configs: Configuration files pushed. 1.66s
cladmin: update-configs: done.
cladmin: update-configs: TOTAL = 2.02s

```

**3. Use the `cpower` command to boot the nodes.**

For example:

```

[root@cladmin ~]# cpower node reboot n*
direct node n05 has been issued a reboot command
direct node n04 has been issued a reboot command
direct node n06 has been issued a reboot command
direct node n01 has been issued a reboot command
direct node n03 has been issued a reboot command
direct node n02 has been issued a reboot command

```

**4. Use the `pdsh` command to verify the boot and verify the system condition.**

For example:

```
[root@cladmin ~]# pdsh -g compute df -h | dshbak -c
-----
n[01,03-06]
-----
Filesystem      Size  Used  Avail Use% Mounted on
tmpfs           7.7G  1.9G  5.9G  25% /
devtmpfs        7.7G   0    7.7G  0% /dev
tmpfs           7.7G  4.0K  7.7G  1% /dev/shm
tmpfs           7.7G  9.0M  7.7G  1% /run
tmpfs           7.7G   0    7.7G  0% /sys/fs/cgroup
tmpfs           1.6G   0    1.6G  0% /run/user/0
-----
n02
-----
Filesystem      Size  Used  Avail Use% Mounted on
tmpfs           7.7G  1.9G  5.9G  24% /
devtmpfs        7.7G   0    7.7G  0% /dev
tmpfs           7.7G  4.0K  7.7G  1% /dev/shm
tmpfs           7.7G  9.0M  7.7G  1% /run
tmpfs           7.7G   0    7.7G  0% /sys/fs/cgroup
tmpfs           1.6G   0    1.6G  0% /run/user/0
```

## Configuring flat compute nodes to boot a `tmpfs` root file system on SLES 12 nodes

The following procedure explains how to configure flat compute nodes to use a `tmpfs` root file system. This procedure applies to nodes that run SLES 12.

### Procedure

1. Use the `cnodes` command to list the compute nodes.

For example:

```
cladmin:~ # cnodes -a
n01
n02
n03
n04
n05
n06
```

2. Use the `cinstallman` command to specify the `tmpfs` file system.

For example, the following command sets node n04 to boot with the `tmpfs` file system:

```
cladmin:~ # cinstallman --set-rootfs tmpfs --node n04
Setting up network boot for service3
cladmin: update-configs: updating SMC configuration files
cladmin: update-configs: -> cminfo
cladmin: update-configs: 0.20s
cladmin: update-configs: Configuration files generated. 0.20s
cladmin: update-configs: -> updating admin node
Number of files transferred: 1
```

```
c1admin: update-configs: Configuration pushed to admin. 1.62s
c1admin: update-configs: Configuration files pushed. 1.62s
c1admin: update-configs: done.
c1admin: update-configs: TOTAL = 1.82s
```

3. Use the following command to check the status of each node:

```
c1admin:~ # cpower system status
n01      BOOTED
n02      BOOTED
n03      BOOTED
n04      BOOTED
n05      BOOTED
n06      BOOTED
```

4. Use the `cpower` command to reboot the node.

For example:

```
c1admin:~ # cpower node reboot n04
direct node n04 has been issued a reboot command
```

5. (Optional) Use Linux commands to verify the boot and verify the system condition.

For example:

```
c1admin:~ # ssh n04
n04:~ # df -h
Filesystem      Size  Used Avail Use% Mounted on
tmpfs           7.8G  2.1G  5.7G  27% /
devtmpfs        7.7G    0   7.7G   0% /dev
tmpfs           7.8G    0   7.8G   0% /dev/shm
tmpfs           7.8G   11M   7.7G   1% /run
tmpfs           7.8G    0   7.8G   0% /sys/fs/cgroup
n04:~ # exit
logout
Connection to n04 closed.
c1admin:~ # ssh n03 df -h
Filesystem      Size  Used Avail Use% Mounted on
/dev/sda32       110G  2.2G  102G   3% /
devtmpfs        7.7G  8.0K   7.7G   1% /dev
tmpfs           7.8G    0   7.8G   0% /dev/shm
tmpfs           7.8G   19M   7.7G   1% /run
tmpfs           7.8G    0   7.8G   0% /sys/fs/cgroup
/dev/sda12      283M   47M   221M  18% /boot
```

## Configuring local storage space for swap and scratch disk space

You can configure a hierarchical cluster to support local storage space on ICE compute nodes. The nodes are also known as **blades**. Solid-state drive (SSD) devices and 2.5" disks are available for this purpose.

HPE supports a set of parameters that you can use to configure partitions on your system. You can define the size and status for both swap and scratch partitions.

You can set the partition values on a global basis or on an individual basis. If you set a value on a global basis, the value applies to all ICE compute nodes. You can also set the value to apply to only one node.

By default, the disks are partitioned only if blank. Swap is off. Scratch is set to occupy the whole disk space. Scratch is mounted at `/tmp/scratch`.

You can use the `cattr` command to retrieve the status of a setting, to enable a setting, or to disable a setting. If you do not set any parameters, the system uses the defaults.

The cluster manager `/etc/init.d/set-swap-scratch` script configures the swap and scratch space based on the settings you specify with the `cattr` command.

The following list explains the local storage space settings:

#### **`blade_disk_allow_partitioning`**

Determines whether you can repartition and reformat the local storage disk. Specify `on` or `off`.

Default is `on`.

To protect user data, the cluster manager prevents you from repartitioning a disk that is already partitioned. In this case, you need a blank disk to use for the `swap` and `scratch` partitions.

#### **`blade_disk_raid_level`**

Specifies whether you can enable RAID0 (striping) or RAID1 (mirroring) when you have two disks for swap and scratch. The values are as follows:

##### **`off`**

Does not enable RAID. Default.

`0`

Enables RAID0 (striping) for the swap and scratch partitions.

`1`

Enables RAID1 (mirroring) for the swap and scratch partitions.

#### **`blade_disk_reformat_scratch_at_boot`**

Specifies whether you are allowed to format the scratch partition every time the ICE compute node boots. The values are as follows:

##### **`off`**

Prevents formatting of the scratch partition at boot. Default.

`0`

Enables formatting of the scratch partition every time the ICE compute node boots.

#### **`blade_disk_reformat_swap_at_boot`**

Specifies whether you are allowed to format the swap partition every time the ICE compute node boots. The values are as follows:

##### **`off`**

Prevents formatting of the swap partition at boot. Default.

`0`

Enables formatting of the swap partition every time the ICE compute node boots.

#### **`blade_disk_scratch_mount_point`**

Specifies the mount point for the scratch partition. Default is `/tmp/scratch`.

You can mount the disk to any mount point. If it does not exist, the cluster manager creates the mount point directory. The cluster manager needs permission to create the mount point at the mount point you specify. On the ICE compute nodes, the root mount point (`/`) is not writable. If you want to mount to `/scratch`, make sure to create that folder as part of the ICE compute node image.

**`blade_disk_scratch_size`**

Specifies the scratch size, in megabytes. Specify either an integer number of megabytes or one of the special values, as follows:

**-0**

Uses all free space for scratch when partitioning. Default.

**0**

Does not create a scratch partition on the local storage disk. Prevents the cluster manager from creating a scratch partition.

**1, 2, ...**

Specifies an integer number of megabytes for the scratch partition.

**`blade_disk_scratch_status`**

Determines whether the cluster manager creates a scratch partition on the local storage disk. Specify `on` or `off`. Default is `off`, which means that the cluster manager does not create a scratch partition.

The cluster manager assigns the label `SGI_SCRATCH` when it partitions the disk. It mounts the scratch on the partition labeled `SGI_SCRATCH`.

**`blade_disk_swap_size`**

Specifies the swap size, in megabytes. Specify one of the following values:

**-0**

Uses all free space when partitioning.

**0**

Does not create a swap partition on the local storage disk. Prevents the cluster manager from creating a swap partition. Default.

**1, 2, ...**

Specifies an integer number of megabytes for the swap partition.

**`blade_disk_swap_status`**

Determines whether the cluster manager creates a swap partition on the local storage disk. Specify `on` or `off`. Default is `off`, which means that the cluster manager does not create a swap partition.

The cluster manager assigns the label `SGI_SWAP` when it partitions the disk. It enables the swap only if an `SGI_SWAP` label exists.

The following topics show the `cattr` commands you can use to configure the swap and scratch disk space:

- [Retrieving the status of a local storage space setting](#) on page 189
- [Enabling, disabling, or respecifying a local storage space setting](#) on page 190

## **Retrieving the status of a local storage space setting**

The following procedure explains how to display the status of a local storage space setting.

## Procedure

1. Log into the admin node as the root user.
2. Enter the `cattr get` command, in the following format, to retrieve the current setting:

```
cattr get setting [-N node_id] --default default
```

The variables are as follows:

Variable	Specification
<i>setting</i>	One of the local storage space settings. For the list of settings, see the following: <a href="#">Configuring local storage space for swap and scratch disk space</a> on page 187
<i>node_id</i>	The system ID for one ICE compute node. Specify this argument only if you want to set one of the local storage space settings for an individual ICE compute node.
<i>default</i>	The default value for this setting.

Example 1. The following command returns `on`, which indicates that the setting is enabled and applies to all ICE compute nodes:

```
# cattr get blade_disk_allow_partitioning --default on
on
```

Example 2. Assume that you set the `blade_disk_scratch_size` to 2 megabytes. To retrieve the current scratch size, enter the following command:

```
# cattr get blade_disk_scratch_size --default -0
2
```

## Enabling, disabling, or respecifying a local storage space setting

The following procedure explains how to modify a local storage space setting.

## Procedure

1. Log into the admin node as the root user.
2. Enter the `cattr set` command, in the following format, to enable, disable, or specify a value for a local storage space setting:

```
cattr set [-N node_id] setting value
```

The variables are as follows:

---

Variable	Specification
<i>node_id</i>	The system ID for one ICE compute node. Specify this argument only if you want to set one of the local storage space settings for an individual ICE compute node.
<i>setting</i>	One of the local storage space settings.
<i>value</i>	<code>on</code> , <code>off</code> , an integer value that represents megabytes, or a mount point. For information about possible values, see the individual setting information in the following topic: <b><a href="#">Configuring local storage space for swap and scratch disk space</a></b> on page 187

---

Example 1. The following command turns on the `blade_disk_allow_partitioning` setting for all ICE compute nodes:

```
# cattr set blade_disk_allow_partitioning on
```

Example 2. The following command turns on `blade_disk_allow_partitioning` for ICE compute node `r1i0n0`:

```
# cattr set -N r1i0n0 blade_disk_allow_partitioning on
```

Example 3. The following command sets the scratch partition mount point for the local disk associated with ICE compute node `r1i0n0` to `/tmp/scratch22`:

```
# cattr set -N r1i0n0 blade_disk_mount_point /tmp/scratch22
```

3. Use the `cimage` command to push the changes out to the desired nodes:

```
# cimage --push-rack image_name racks
```

## Using the `cattr` command to modify system attributes

You can use the `cattr` command to assign attributes to cluster nodes. You can assign attributes either on a global basis, to the entire system, or on an individual node basis.

The `cattr` command can retrieve attribute settings, set attributes, remove attributes, and perform other functions. Enter the following to retrieve a `cattr` command help statement and the list of attributes you can manipulate:

```
# cattr -h
```

---

**NOTE:** When possible, use the `cadmin` command, rather than the `cattr` command, to modify system attributes. When you use the `cadmin` command to modify an attribute, the `cadmin` command regenerates the configuration and eliminates the need for you to issue an `update-configs` command. To make custom configuration changes that you want preserved across `update-configs` calls, see the following:

**[Preserving custom configuration changes](#)** on page 309

---

For information about how to modify local storage space attributes, see the following:

**[Configuring local storage space for swap and scratch disk space](#)** on page 187

# About disk quotas

Within the compute image for an ICE leader node, the cluster manager sets default per-directory disk **quotas**, which can also be called **project quotas**. The quota mechanism prevents a disk from filling up and inhibiting a node from booting.

Soft quotas and hard quotas apply to any entity that writes to disk. For example, quotas pertain to a user writing to disk actively or a user job that writes to disk.

Quotas prevent an ICE compute node from accidentally filling the disk space of the associated leader node over the network file system (NFS). Quotas apply when a ICE compute node is booted with NFS root directories, not `tmpfs` directories.

The cluster manager sets default quota settings in each software image, rather than in each node. You can adjust these quota settings at your site. The soft quotas and hard quotas are as follows:

- A soft quota is an initial limit. After an ICE compute node exceeds a soft quota, the ICE compute node can continue to use resources up until it reaches the upper hard limit.
- A hard quota is a firm limit.

The default quotas are as follows:

- Soft quota = 2048 minutes
- Hard quota = 2148 minutes
- Quota timer = 1 day

The ICE compute node might fail to boot properly in the following cases:

- If a hard quota is exceeded
- If a soft quota is exceeded past the time set in the timer

A hierarchical cluster prevents additional writes to a disk when either of the following events occur:

- A disk reaches its hard limit.
- A disk reaches its soft limit and the timer has expired.

The following topics provide more information about quotas:

- [\*\*Retrieving quota information\*\*](#) on page 192
- [\*\*Setting quotas\*\*](#) on page 194
- [\*\*Viewing the ICE compute node read/write quotas\*\*](#) on page 195

## Retrieving quota information

The following procedure explains how to retrieve quota values for a specific image.

## Procedure

1. Log into the admin node as the root user.
2. Enter the following command to retrieve a list of the images on the system:

```
# cinstallman --show-images
Image Name          BT  VCS  Compat_Distro
ice-rhel7.4          0   0    rhel7
            3.10.0-514.el7.x86_64
ice-rhel7.4-kdump      0   1    rhel7
            3.10.0-514.el7.x86_64
ice-sles12sp3-mofed      0   1    sles12
            3.0.101-80-default
lead-rhel7.4          0   1    rhel7
            3.10.0-514.el7.x86_64
lead-rhel7.4-ha         0   1    rhel7
            3.10.0-514.el7.x86_64
rhel7.4                0   1    rhel7
            3.10.0-514.el7.x86_64
sles12sp3-mofed         0   1    sles12
            3.0.101-80-default
```

The example output shows the following ICE compute node images:

- ice-rhel7.4
- ice-rhel7.4-kdump
- ice-sles12sp3-mofed

3. Enter one of the following commands to retrieve information about one of the quotas or the quota timer:

```
cadmin --show-soft-quota --image image_name
cadmin --show-hard-quota --image image_name
cadmin --show-quota-timer --image image_name
```

For *image\_name*, enter one of the names from the `Image Name` column in the previous step.

For example:

```
# cadmin --show-soft-quota --image ice-rhel7.4
2048m
# cadmin --show-hard-quota --image ice-rhel7.4
2148m
# cadmin --show-quota-timer --image ice-rhel7.4
1d
```

The `cadmin` command output displays the quotas using the format of the underlying tool, which is the XFS file system project quota infrastructure. For information about the format, see the `xfs_quota(8)` manpage.

4. (Optional) Set site-specific quotas.

Proceed to the following:

[Setting quotas](#) on page 194

## Setting quotas

The following procedure explains how to change a quota or the quota timer.

### Procedure

1. Log into the admin node as the root user.
2. Verify the current value for the quota setting you want to change.

For information about how to verify quota settings, see the following:

[Retrieving quota information](#) on page 192

3. Modify the quota setting.
  - To set a site-specific *value*, use one of the following commands:

```
cadmin --set-soft-quota --image image_name value
cadmin --set-hard-quota --image image_name value
cadmin --set-quota-timer --image image_name value
```

The variables are as follows:

Variable	Specification
<i>image_name</i>	<p>One of the names in the <code>Image Name</code> column of output from the <code>cinstallman --show-images</code> command.</p> <p>For information about the <code>cinstallman --show-images</code> command, see the following:</p> <p><a href="#">Retrieving quota information</a> on page 192</p>
<i>value</i>	<p>An integer value followed by a unit specification.</p> <p>For <code>--set-soft-quota</code> or <code>--set-hard-quota</code> operations, specify the following:</p> <ul style="list-style-type: none"><li>◦ <code>k</code> for kilobytes</li><li>◦ <code>m</code> for megabytes</li><li>◦ <code>g</code> for gigabytes</li><li>◦ <code>t</code> for terabytes</li></ul> <p>For the <code>--set-quota-timer</code> operation, specify the following:</p> <ul style="list-style-type: none"><li>◦ <code>m</code> for minutes</li><li>◦ <code>d</code> for days</li><li>◦ <code>h</code> for hours</li><li>◦ <code>w</code> for weeks</li></ul>

The following examples specify site-specific values for the quotas associated with the `ice-sles11sp3` compute image:

```
# cadmin --set-soft-quota --image ice-sles11sp4 4200m
# cadmin --set-hard-quota --image ice-sles11sp4 4196m
# cadmin --set-quota-timer --image ice-sles11sp4 3d
```

- To reset a site-specific value back to the factory default value, use one of the following commands:

```
cadmin --unset-soft-quota --image image_name
cadmin --unset-hard-quota --image image_name
cadmin --unset-quota-timer --image image_name
```

The following examples reset site-specific values back to the factory default values:

```
# cadmin --unset-soft-quota --image ice-sles11sp4
# cadmin --unset-hard-quota --image ice-sles11sp4
# cadmin --unset-quota-timer --image ice-sles11sp4
```

#### 4. Push out the changes to the ICE compute nodes.

For information about how to push changes, see the following:

[Pushing images from the admin node to the targeted nodes](#) on page 217

## Viewing the ICE compute node read/write quotas

You can retrieve the read quota and the write quota for each ICE compute node.

The following procedure explains how to retrieve current usage.

### Procedure

1. Log into the admin node as the root user.
2. Use the `ssh` command to log into one of the leader nodes.

To retrieve a list of leader nodes, enter `cnodes --leader`.

The following example shows how to retrieve a list of leader nodes and how to log into one of them:

```
# cnodes --leader
r1lead
r2lead
# ssh r1lead
```

3. Enter the following command to retrieve a list of projects:

```
# less /etc/projects
1:/var/lib/sgi/per-host/ice-rhel7.4/1/i0n0
2:/var/lib/sgi/per-host/ice-rhel7.4/1/i0n1
3:/var/lib/sgi/per-host/ice-rhel7.4/1/i0n2
4:/var/lib/sgi/per-host/ice-rhel7.4/1/i0n3
5:/var/lib/sgi/per-host/ice-rhel7.4/1/i0n4
6:/var/lib/sgi/per-host/ice-rhel7.4/1/i0n5
7:/var/lib/sgi/per-host/ice-rhel7.4/1/i0n6
8:/var/lib/sgi/per-host/ice-rhel7.4/1/i0n7
9:/var/lib/sgi/per-host/ice-rhel7.4/1/i0n8
10:/var/lib/sgi/per-host/ice-rhel7.4q/1/i0n9
.
```

The project numbers are the leftmost integers in the output. Enter q to exit the less command.

4. Use the `xfs_quota` command, in the following format, to retrieve the current usage values:

```
xfs_quota -x -c 'quota -ph project_num'
```

For `project_num`, specify one of the project numbers you retrieved in the preceding step.

For example:

```
r1lead:~ # xfs_quota -x -c 'quota -ph 1'
Disk quotas for Project #1 (1)
Filesystem  Blocks  Quota  Limit Warn/Time    Mounted on
/dev/disk/by-label/sgiroot
       64.6M      0     1G  00 [-----]  /
```

## Creating custom partitions

By default, the cluster manager provides partition layouts. These layouts allow one or more instances of the cluster manager to be installed on the same root drive or drives. Each instance of the cluster manager is housed in a slot. A **slot** is a disjoint subset of partitions. The cluster manager documentation refers to the default partitioning scheme as **slot partitioning**. The cluster manager supports 1 to 10 slots. By default, there are 2 slots.

If the default partition scheme does not work for your cluster, you can create a custom partitioning scheme.

The following topics describe custom partitioning:

- [\*\*Custom partitioning notes, constraints, and cautions\*\*](#) on page 196
- [\*\*Configuring custom partitioning for flat compute and leader nodes\*\*](#) on page 197
- [\*\*Configuring custom partitioning for admin nodes\*\*](#) on page 199
- [\*\*Managing custom partitions\*\*](#) on page 199

For more information about the default partition layout and the role of slots, see the following:

[\*\*HPE Performance Cluster Manager Installation Guide\*\*](#)

## Custom partitioning notes, constraints, and cautions

Before you implement custom partitioning, consider the following items:

- Reserving disk space for partitions

If you want scratch space reserved on a disk for your partitions, use the disk reservation feature rather than custom partitioning. Use custom partitioning when core operating system partitions are in play. The disk reservation feature and custom partitioning are not compatible.

For information about the disk reservation feature, see the following:

[\*\*Configuring scratch disk space on system disks\*\*](#) on page 140

- RAID configurations

Custom partitioning does not affect RAID configurations: MD RAID, BIOS SW RAID, or single disks (including HW RAIDs). Custom partitioning is activated after any RAIDs are created or assembled. Custom partitioning does not define the RAIDs.

- Applicable cluster nodes

The cluster manager supports custom partitioning on admin, leader, and flat compute nodes. However, the manner by which you configure the custom partitioning varies.

The cluster manager does not support custom partitioning on ICE compute nodes.

- High Availability (HA) configurations

Custom partitioning is not supported on physical admin nodes in HA configurations.

- Slots

The slot partitioning scheme of the admin node determines the initial slot partitioning of all other nodes. Custom partitioning does not support slot partitioning (install slots). Hence, custom partitioning of the admin node precludes the use of slots on any nodes in the cluster.

The use of slots, especially on the admin node, allows you to more smoothly handle cluster manager and Linux distribution updates. Slots provide convenient fallback locations when upgrading. Even if you desire custom partitions on other node types, consider your future upgrade plans before deciding to use custom partitions on the admin node itself.

- Data loss

If a node is configured for custom partitioning and you reinstall the node, the following occurs:

- Its partitions are erased.
- All data on the root drive is lost.

Similarly, assume the following:

- A node is configured with custom partitioning.
- You configure it for default slot partitioning.

In this case, data loss occurs on the root drive when the node is reinstalled.

## Configuring custom partitioning for flat compute and leader nodes

The following procedure explains how to configure custom partitioning for flat compute and leader nodes.

1. Describe the custom partition layout in a specially formatted configuration file.

You can choose any filename as long as it begins with an alphabetic character and has a file extension of .cfg.

Write the file to the /opt/clmgr/image/scripts/pre-install directory.

Within the file, create a table. In this table, each row defines a partition. The columns, separated by the character |, contain the partition specifications. In order, the columns contain the following partition specifications:

- Partition number
- Mount point
- Size
- Filesystem type  
XFS, ext3, or ext4.

- Filesystem label
- EFI-only option

A file system that can only be created/mounted on a UEFI platform.

- Mount/filesystem options
- A `mkfs` command

The command specification includes substitutions for file system label and the partition device.

For an example configuration file, see the following:

```
/opt/clmgr/image/scripts/pre-install/custom-partitions.cfg
```

The example file contains comments that describe requirements. For example, the file explains how to order the partitions and explains the constraints on various specifications. The following is a sample partition layout:

```
#part| mount_point | sgdisk_size | fs | fs_label | efi_only | mount_opts | mkfs_command
1 |           | 50M | ext3 | sgidata | no | | mke2fs -F -L %fs_label -j %dev
2 | swap | 2048M | swap | sgiswap | no | | mkswap -f -L %fs_label %dev
3 |           | 50M | vfat | ADMINEFI | yes | defaults | mkdosfs -I -n %fs_label %dev
4 | / | 16384M | ext3 | sgirroot | no | defaults | mke2fs -F -L %fs_label -j %dev
5 | /boot | 400M | ext3 | sgiboot | no | defaults | mke2fs -F -L %fs_label -j %dev
6 | /boot/efi | 150M | vfat | SGIEFI | yes | defaults | mkdosfs -I -n %fs_label %dev
7 | /var | 32768M | ext3 | myvar | no | defaults | mke2fs -F -L %fs_label -j %dev
8 | /scratch | FILL | xfs | scratch | no | defaults | mkfs.xfs -f -L %fs_label %dev
```

## 2. Set the `custom-partitions` system attribute to the name of this configuration file.

The following two methods are available for setting this attribute:

- Using the `cadmin` command

Use one of the following command entries:

```
# cadmin --set-custom-partitions --node target_nodes filename
```

or

```
# cadmin --set-custom-partitions --image image_name filename
```

For `filename`, specify the filename of the configuration file you created.

When the node reboots, the repartitioning takes place. For the second entry, any node that is assigned this image receives the custom partitioning scheme at its next reboot. If there is a conflict in the partitioning specification between the node and image, the node specification takes precedence.

- Using the `discover` command

The `discover` command configures or reconfigures a node into the cluster. You can use the `discover` command to set the `custom_partitions` attribute. You can specify the `custom_partitions=filename` parameter in the `discover` configuration file or you can use the `--custom_partitions=filename` option on a `discover` command-line entry.

For more information about the `discover` command, see the following:

[discover command](#) on page 154

## Configuring custom partitioning for admin nodes

As with leader and flat compute nodes, you must describe your custom partition layout in a specially formatted configuration file. However, with admin nodes, you must do the configuring at cluster installation time. To get a template of this configuration file to use during installation, access the file named `custom_partition_example` on the installation media. For help and support, contact your customer support representative.

## Managing custom partitions

The following are some management actions you might need to take when you have nodes with custom partitions:

- Display information about an image or node regarding custom partitioning.

Use the `cadmin` command in the following format:

```
cadmin --show-custom-partitions --image image_name | --node target_nodes
```

The command returns the name of the custom partition configuration file or `Disabled`.

For images, `Disabled` means that the cluster manager uses slot partitioning when the image is installed on a node. Slot partitioning is the default.

For nodes, `Disabled` indicates one of the following:

- The node currently has slot partitioning.
- The node is going to be repartitioned with slot partitioning the next time the node boots.

- Clear the `custom-partitions` attribute for an image or selected nodes.

Use the `cadmin` command in the following format:

```
cadmin --unset-custom-partitions --image image_name | --node target_nodes
```

The command disables custom partitioning for the nodes or for the image effective for the next reboot. A node specification takes precedence over the specification associated with an image.

## Backing up and restoring the system database

The cluster manager database is critical to the operation of your cluster. The database includes relevant data for the managed objects in a cluster. Make sure you back up the database on a regular basis.

Managed objects on a cluster include the following:

- The cluster itself

The whole cluster is a managed object. A hierarchical cluster is modeled as a meta-cluster. This meta-cluster contains the racks each modeled as a subcluster.

- Nodes

Admin node, leader nodes, compute nodes, ICE compute nodes (blades), and chassis management controllers (CMCs) are modeled as nodes.

- Networks

The preconfigured and potentially customized IP networks.

- NICs

The network interfaces for Ethernet and InfiniBand adapters.

- The node images installed on each particular node.

HPE recommends that you keep three backups of your system database at any given time. Implement a rotating backup procedure following the son-father-grandfather principle.

The following procedures explain how to back up and restore the system database:

- [\*\*Backing up the cluster database\*\*](#) on page 200
- [\*\*Restoring the cluster database\*\*](#) on page 201

## Backing up the cluster database

### Procedure

1. Log into the admin node as the root user.

2. Enter the following commands to stop the cluster manager and back up the cluster database:

```
# systemctl stop cmu  
# sqlite3 /opt/clmgr/database/db/cmu.sqlite3 ".backup file"
```

For *file*, specify a name for the backup file.

For example:

```
# sqlite3 /opt/clmgr/database/db/cmu.sqlite3 ".backup cmu.backup.sqlite3"
```

3. Write the backup file to another computer system at your site for safekeeping.

The cluster database is the internal database that hosts information about each cluster component. A copy of the original cluster database can be valuable when performing a disaster recovery. Make sure to take additional, periodic database backups in the future as you modify your system.

## Restoring the cluster database

### Procedure

1. Log into the admin node as the root user.
2. Enter the following commands to restore the cluster database and start the cluster manager:

```
# cp file /opt/clmgr/database/db/cmu.sqlite3  
# systemctl start cmu
```

For *file*, specify the name of the backup file.

For example, the following lines show how to restore the database. Answer **y** when prompted to affirm the database overwrite.

```
# cp -i cmu.backup.sqlite3 /opt/clmgr/database/db/cmu.sqlite3  
cp: overwrite '/opt/clmgr/database/db/cmu.sqlite3'? y  
# systemctl start cmu
```

## Enabling EDNS

Extension mechanisms for DNS (EDNS) can cause excessive logging activity when not working properly. The cluster manager limits EDNS logging. This section describes how to delete this code and allow EDNS to work unrestricted and log messages.

To enable EDNS on your cluster, perform the following steps:

### Procedure

1. Open the `/opt/clmgr/lib/Tempo/Named.pm` file with your favorite editing tool.
2. (Optional) Remove the limit on the `edns_udp_size` parameter.  
Comment out or remove the following line:  
`$limit_edns_udp_size = "edns-udp-size 512;";`
3. Remove the following lines so that EDNS logging is no longer disabled:

```
logging {  
category lame-servers {null; };  
category edns-disabled { null; }; };
```

# Managing software images

The following table shows the operating systems that the cluster manager supports on each node type:

Admin node	Leader node	ICE compute node	Flat compute node
SLES 12 SP3	SLES 12 SP3	SLES 12 SP3	SLES 12 SP3
RHEL 7.4	RHEL 7.4	RHEL 7.4	RHEL 7.4
CentOS 7.4	CentOS 7.4	CentOS 7.4	CentOS 7.4
		RHEL 6.9	RHEL 6.9
		SLES 11 SP4	SLES 11 SP4
		CentOS 6.9	CentOS 6.9

This chapter includes the following topics:

- [About cluster images](#) on page 202
- [About installation repositories](#) on page 205
- [Adding software to the cluster manager repository database](#) on page 208
- [Selecting the repositories to be used in the installation](#) on page 209
- [Creating images to host new software](#) on page 211
- [Installing new software into new images](#) on page 213
- [Associating nondefault images with targeted nodes](#) on page 215
- [Pushing images from the admin node to the targeted nodes](#) on page 217
- [Miscellaneous image management tasks](#) on page 218

## About cluster images

The following topics explain the images that reside on cluster nodes and the image management commands:

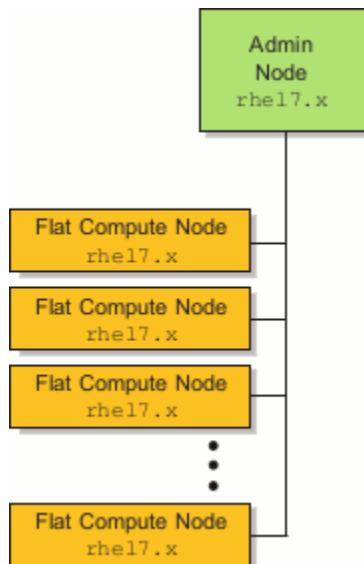
- [Node types and default image names](#) on page 202
- [Image management commands](#) on page 204

## Node types and default image names

Clusters include different types of nodes, and there is a unique software image for each individual node type. When you install additional software on your cluster, you might need to modify the software on some of the nodes.

To get a profile of the node types on your cluster, use the `cnodes` command. The command can display the profile of all nodes (`cnodes --all`) or you can specify the display nodes of a specified type. Use the `--help` option to see the various specifications.

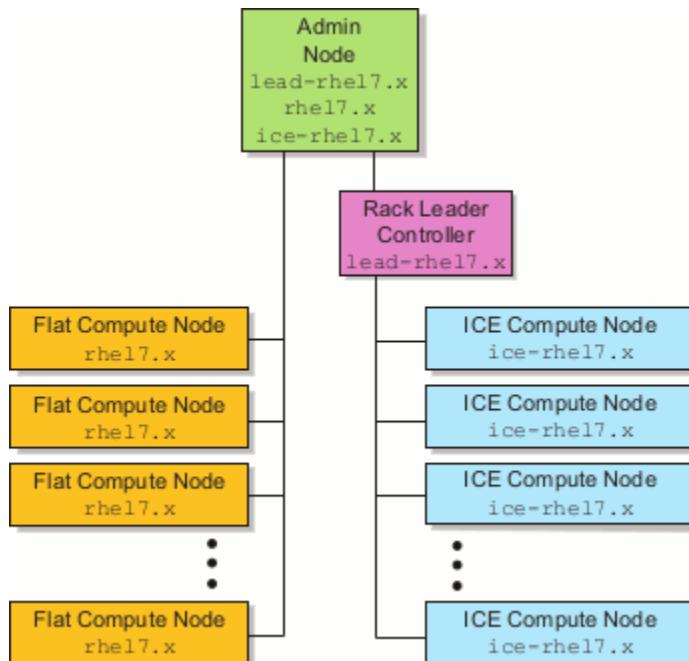
The following figure shows a simple flat cluster and the software images that reside on each node. The admin node hosts an image called `rhel7.x`, which is the default image for the compute nodes in the cluster.



**Figure 50: Flat cluster - node software images noted**

The following figure shows a simple hierarchical cluster and the software images that reside on each node. The admin node hosts the following images, which are the default images for the other nodes in the cluster:

- `lead-rhel7.x` is the default image for the leader node.
  - `rhel7.x` is the default image for the flat compute nodes.
  - `ice-rhel7.x` is the default image for the ICE compute nodes.



**Figure 51: Hierarchical cluster - node software images noted**

The figures show that all cluster nodes subordinate to the admin node have the same operating system. This depiction is for convenience only and is not a requirement.

## Image management commands

The following list shows the primary image management commands:

### **cadmin**

Displays and manages repository group assignments. This command plays a minor role in image management.

### **cimage**

Used with hierarchical cluster images only. Performs the following functions:

- Assigns images to nodes.
- Pushes images to racks.
- Lists available images.
- Lists images currently on nodes.

### **cinstallman**

Used with hierarchical cluster images and operating system distribution images. Performs the following functions:

- Creates an image from scratch.
- Recreates an image. Any nodes associated with the image before you run the command are associated with the image after the command runs.
- Uses existing images that might have been created by some other means.
- Deletes images.
- Shows available images.
- Updates or manages images. Uses the `yume` command.
- Formally tracks revisions to images.
- Assigns images to nodes.
- Defines rebooting behavior. Sets whether a node can image itself or whether a node boots from its disk the next time the node reboots.

### **crepo**

Manages repositories. Performs the following functions:

- Adds, deletes, and displays repositories.
- Selects and clears repositories for RPM list generation.
- Creates, deletes, and displays logical groups of repositories.

To retrieve a list of parameters for each command, do one of the following:

- Retrieve the manpage for the command
- Enter the command name and `--help` on the command line

## About installation repositories

The `crepo` command manages the cluster manager software and the Linux distribution software. You can also use the `crepo` command to manage custom repositories you create yourself or to add media.

Often, the repositories for the cluster manager and for the Linux distribution reside in the following directory on the admin node:

`/opt/clmgr/repos`

On other clusters, the repositories could reside on a remote server.

The following topics explain concepts related to installation repositories:

- [\*\*Repository metadata\*\*](#) on page 205
- [\*\*Remote repositories\*\*](#) on page 206
- [\*\*General repository management parameters\*\*](#) on page 206
- [\*\*RPM lists\*\*](#) on page 207

## Repository metadata

The cluster manager associates the following with each repository:

- A name

The repository name is a metadata item that you supply to the `crepo` command. Except for custom repositories, the `crepo` command extracts the repository name from the media when adding a repository. The `crepo --show` command displays the repository name and the repository location in the following format:

`name : location`

The following topic contains an example:

[\*\*About installation repositories\*\*](#) on page 205

- A directory
- Selection status
- Suggested package lists

The repository information is determined from the media itself when you add the following types of software:

- HPE
- Linux distributors (SLES, RHEL, or CentOS)
- Any other YaST-compatible software

For customer-supplied repositories, supply information to the `crepo` command when adding the repository.

The `crepo` command uses items like the selection status and the suggested package lists to build the RPM lists required for software installations.

## Remote repositories

The cluster manager supports the use of both local repositories and remote repositories. You can access the repositories by using `http` and `https`.

To use remote repositories for HPE distribution media or Linux distribution media, the repositories must contain the complete, expanded media including any dot files (`.filename`). If the remote media is HPE distribution media or Linux distribution media, the cluster manager processes the default RPM lists the same way it processes locally hosted media. You can use remote repositories on cluster nodes that are routed to the servers upon which the remote repositories reside. If necessary, establish the correct routing.

For remote Linux distribution repositories, if the distribution spans one source (for example, the distribution includes both DVD1 and DVD2), take care when expanding DVD2. If you overwrite DVD1 files, the overwrite breaks distribution detection.

The following is an example of a correct copy:

```
# cp -a /dvd-mnt-rhel7-dvd1 /web-export/rhel7
# cp -a /dvd-mnt-rhel7-dvd2/* /web-export/rhel7
```

## General repository management parameters

You can use the `crepo` command to manage your software repositories. The following table summarizes various management actions.

Repository Action	Description
Adding	<p>Use the <code>--add</code> option.</p> <p>For information, see the following:</p> <p><a href="#">Adding software to the cluster manager repository database</a> on page 208</p>
Selecting	<p>For information, see the following:</p> <p><a href="#">Selecting the repositories to be used in the installation</a> on page 209</p>
Displaying	<p>Use the <code>--show</code> option to display all available repositories.</p> <p>Use the <code>--show-distro</code> option to display repositories only for Linux distributions.</p>
Grouping	<p>For information, see the following:</p> <p><a href="#">Selection using a group assignment</a> on page 210</p>
Deleting	<p>Use the <code>--del repo_name</code> option to delete a repository. This parameter also deletes the associated <code>/opt/clmgr/tftpboot</code> directory, if present.</p>
Refreshing metadata	<p>Use the <code>--refresh repo_name</code> option to refresh metadata.</p> <p>You cannot refresh HPE distribution repositories or Linux distribution repositories. Instead, put any updated or additional RPMs in a custom repository.</p>

## RPM lists

The `crepo` command constructs default RPM lists based on the repository metadata. These lists include suggested package lists and selection status. The `cinstallman` command can use RPM lists when creating images. The RPM lists are generated only if a single distribution is selected. You can find the lists in `/opt/clmgr/image/rpmlists`. The lists match the form `generated-* rpmlist`. The `crepo` command can display its actions when it updates or removes generated RPM lists.

For example:

```
# crepo --select SUSE-Linux-Enterprise-Server-12-SP3
Updating: /opt/clmgr/image/rpmlists/generated-ice-sles12sp3.rpmlist
Updating: /opt/clmgr/image/rpmlists/generated-sles12sp3.rpmlist
```

When generating RPM lists, the `crepo` command combines a list of distribution RPMs with suggested RPMs from every other selected repository. The following directory contains the generated RPM lists:

`/opt/clmgr/image/rpmlists`

To override an RPM list, you can take the generated lists and adjust them to suit your needs. To pass your updated or customized RPM list to the `cinstallman` command, use the `--rpmlist` parameter. To override suggested RPM lists, as opposed to using your own custom RPM list, you can create an override RPM list associated with a media type. To override the suggested RPM lists, create an override RPM list in the following directory:

`/opt/clmgr/image/rpmlists/override/`

For example, to change the default RPM list for the SMC media, create the following file:

`/opt/clmgr/image/rpmlists/override/cm-1.0-sles12sp3`

For more information about `rpmlist` customization information, see the following:

[Creating ICE compute node images](#) on page 229

For information about generating RPM lists for repository groups, see the following:

[Selection using a group assignment](#) on page 210

## Adding and updating software images

After installation and configuration, you might need to modify images or add new images. The following are some image management tasks that you might need to perform on a production system:

- Adding site-specific software.
- Adding or updating a software component in an image. For example, kernel or configuration files.
- For flat compute nodes, installing packages on the running nodes themselves.

For information about how to create images from Linux distributions, see the following:

[HPE Performance Cluster Manager Installation Guide](#)

The task of creating or modifying new software images is a multistep process. Complete the procedures in the following topics:

### Procedure

1. [Adding software to the cluster manager repository database](#) on page 208
2. [Selecting the repositories to be used in the installation](#) on page 209
3. [Creating images to host new software](#) on page 211

4. [Installing new software into new images](#) on page 213
5. [Associating nondefault images with targeted nodes](#) on page 215
6. [Pushing images from the admin node to the targeted nodes](#) on page 217

## Adding software to the cluster manager repository database

To install new software, first use the `crepo` command to add the software to the repository. Log into the admin node as the root user, and use the `crepo` command in the following format:

```
crepo --add location [--custom 'repository_name']
```

The following subsections describe the command options for various repositories:

- [Standard repository](#) on page 208
- [Custom repository](#) on page 209
- [Multiple media sources](#) on page 209
- [Nested repositories](#) on page 209

### Standard repository

To build a repository for an OS distribution or HPE software media, use the `crepo` command in the following format:

```
crepo --add location
```

For *location*, specify one of the following:

- If the new software resides on a remote web server, specify a URL.  
Make sure that the host name in the URL is a fully qualified domain name (FQDN). If you do not specify an FQDN, the `crepo` command might not properly resolve the host location.  
The cluster manager adds the new repository in the remote web location.
- If the new software resides on a server that is NFS-mounted to the admin node, specify the full path to the ISO file on that server. Use the following format for *location*:

```
host:/path_to_ISO_file
```

The cluster manager adds the new repositories to the following directory:

```
/opt/clmgr/repos
```

- If the new software resides on the admin node, specify one of the following:
  - The path to the ISO file on the admin node
  - The path to mounted media

The cluster manager adds the new repositories to the following directory:

```
/opt/clmgr/repos
```

## Custom repository

When `--add` is used with the `--custom` option, the value *location* must point to an existing directory (with existing RPMs). Repository metadata is created inside the existing directory. Make sure that *location* resides under the following directory:

```
/opt/clmgr/repos
```

The *location* can be a remote repository.

For more information about custom repositories, see the following:

[\*\*Using a custom repository for site packages\*\*](#) on page 219

## Multiple media sources

For Linux distribution media that has more than one DVD, you can use a `crepo --add` call on the first media and a separate call for the second media. The `crepo` command combines the two DVDs into a single repository.

## Nested repositories

Some Linux distributions have subdirectories with additional separate repositories in them. The `crepo` command searches for these additional repositories and includes the subdirectory repositories when the media is selected. The result is that certain distribution media may have more than one URL associated with the media.

The following procedure explains how to set up nested repositories on a remote server.

### Procedure

1. Use the `crepo` command to register the remote Linux distro URL.

Example using a remote repository for RHEL 7.X media:

```
# crepo --add http://updateserver.example.com//rhel/7.x/x86_64/os/
```

2. Add a separate custom repository that points to each separate subdirectory you want to include.

The following example uses the required `ScalableFileSystem` component:

```
# crepo --add http://updateserver.example.com//rhel/7.4/x86_64/os/ \
ScalableFileSystem --custom rhel7.4-scalabe-fs
```

## Selecting the repositories to be used in the installation

Over time, for any given software installation or software update, it is unlikely that you need all available repositories. The `crepo` command lets you select the pertinent repositories to be used later by default with the `cinstallman` command.

The `cinstallman` command lets you override the default selection. As noted earlier, the `crepo` command uses the selection status to determine whether RPM list generation is required for the repository. When you select media repositories, the `crepo` command generates RPM lists per node type with the following name:

```
/etc/clmgr/image/rpmlists/generated--*.rpmlist.
```

The following output from a `crepo --show` example shows the available repositories:

```
# crepo --show
* SUSE-Linux-Enterprise-Software-Development-Kit-12-SP3 : /opt/clmgr/repos/other/sle-sdk12sp3
* SUSE-Linux-Enterprise-Server-12-SP3 : /opt/clmgr/repos/distro/sles12sp3
* sles12sp3-debuginfo-updates : /opt/clmgr/repos/sles12sp3-debuginfo-updates
```

```
* sles12sp3-sdk-updates : /opt/clmgr/repos/sles12sp3-sdk-updates
* HPE-MPI-1.2-sles12sp3 : /opt/clmgr/repos/cm/HPE-MPI-1.2-sles12sp3
* sles12sp3-updates : /opt/clmgr/repos/sles12sp3-updates
```

There are three fields in each line of the output: selection status, repository name, and the path to the repository. An asterisk for the selection status indicates that the repository is selected.

The following topics explain repository selection methods:

- [Explicit selection using the crepo command](#) on page 210
- [Selection using a group assignment](#) on page 210
- [Selection using the cinstallman command](#) on page 211

## Explicit selection using the `crepo` command

Use the `crepo` command to explicitly select repositories in the following cases:

- There are relatively few repositories involved
- The updates are likely to be a one-time or infrequent event

On the admin node, use the `crepo` command in the following format to select a repository:

```
crepo --select repository_name
```

For example, to select the repository SUSE-Linux-Enterprise-Server-12-SP3, enter the following:

```
# crepo --select SUSE-Linux-Enterprise-Server-12-SP3
```

The `crepo` command accepts globbing characters (wildcards). With long repository names, you can use wildcards, as the following shows:

```
# crepo --select SUSE-*-SP3
```

You can specify only one repository with the command. To select multiple repositories, use multiple commands.

To clear a repository, use the following format:

```
crepo --unselect repository_name
```

## Selection using a group assignment

Your site might have a compute environment that requires many repositories or sufficiently interesting sets of repositories. In this case, you might want to define repository groups to be used with the `cinstallman` command.

To create a repository group, use the `crepo` command in the following format on the admin node:

```
crepo --add-group group_name list
```

For *list*, specify zero or more repository names, space-delimited. Subsequent `crepo --add-group` commands that specify a group name add members to the group.

When you create a repository group that includes a supported Linux distribution and the necessary components, the `crepo` command generates RPM lists for the default images of the various node types. For example, for group `xyz` that includes the Linux distribution RHEL 7.4, it creates the following:

```
/opt/clmgr/rpmlists/generated-group-xyz-rhel7.4.rpmlist
/opt/clmgr/rpmlists/generated-group-xyz-ice-rhel7.4.rpmlist
/opt/clmgr/rpmlists/generated-group-xyz-lead-rhel7.4.rpmlist
```

Related group actions:

- The following command displays groups or group members of specified groups:

```
crepo --show-group [group_names]
```

If no group names are specified, the cluster manager displays all groups names.

- The following command deletes group members:

```
crepo --del-group group_name [list]
```

For *list*, specify zero or more repository names, space-delimited. If you do not specify any repository names, the system deletes *group\_name*.

If a repository group member is deleted from the `crepo` repository, the group membership remains intact. Hence, you can recreate a repository using the same name and have its group membership already established.

- You can use the `cadmin` command to assign repository groups to images for use by the `cinstallman` command. Use the following `cadmin` options:

- `--set-repo-group`
- `--unset-repo-group`
- `--show-repo-group`

## Selection using the `cinstallman` command

The `cinstallman` command lets you override default selections with either its `--repos` option or its `--repo-group` parameter. The following list shows the order of precedence for repository selection:

1. The `--repos` or the `--repo-group` specification (not both) on the `cinstallman`
2. If targeting an image, the repository group (if any) associated with the image
3. The default selections set by explicit `crepo` selections

## Creating images to host new software

Generally, you want to house software updates in new images. You might want to do so for scalability, change tracking, and reusability. However, there are instances where you might choose to do otherwise. For examples, see the following:

- [Changing the services on the ICE compute nodes](#) on page 224
- [Installing new software into new images](#) on page 213

The following topics explain ways to create images in the cluster manager:

- [Cloning an existing cinstallman image](#) on page 212
- [Using the cluster manager version control system \(VCS\)](#) on page 212
- [Capturing an image from a running compute node](#) on page 212

## Cloning an existing `cinstallman` image

Use one of the following methods to display the list of available images to clone:

- List the contents of directory `/opt/clmgr/image/images`
- Use the `cinstallman -show-images` command

To clone an existing image on the admin node, use the `cinstallman` command in the following format:

```
cinstallman --create-image --clone --source original --image new_image
```

The cluster manager adds image `new_image` to the list of available images in `/opt/clmgr/image/images`.

## Using the cluster manager version control system (VCS)

Using VCS to create an image is similar to the cloning method. The difference is that when you use VCS, you do not need to employ a second name for the changed image. VCS does the implicit cloning and lets you change the clone. The changed image retains the same name as the original, but VCS assigns different version numbers to the images. Depending on your image management needs, you might find the VCS scheme sufficient.

In this scheme, you move to the image environment on the admin node, `/opt/clmgr/image/images`, and update the original image.

For more information about VCS, see the following:

[Using the version control system](#) on page 231

## Capturing an image from a running compute node

If you want to capture the operating environment from a running flat compute node, you can capture this environment in an image. In this case, make sure that the operating system repositories currently selected match that of the image being captured from the running node. To select a running node, first compare its on-disk image to its image on the admin node.

To capture the image, use the `cinstallman` command in the following format:

```
cinstallman --create-image --image new_image --from-node target_node [options]
```

The variables are as follows:

---

Variable	Specification
<i>new_image</i>	A name for the new image.  If the image name <i>new_image</i> exists, the cluster manager overwrites or refreshes the existing image.  Otherwise, the cluster manager adds image <i>new_image</i> to the list of available images in /opt/clmgr/image/images.
<i>target_node</i>	The name of the node upon which the image is running. If needed, use the <code>cinstallman --show-nodes</code> command to identify <i>target_node</i> .
<i>options</i>	Specify zero or more options. The <code>cinstallman</code> command passes any specified <i>options</i> arguments to the <code>rsync</code> command.  Examples of such arguments are as follows: <ul style="list-style-type: none"> <li>• <code>--exclude</code></li> <li>• <code>--exclude-file</code></li> <li>• <code>--one-file-system</code></li> </ul> For more information, see the <code>rsync</code> manpage.

---

For related information, see the following:

- [Selecting the repositories to be used in the installation](#) on page 209
- [Comparing the image on a running node with images on the admin node](#) on page 223

## Installing new software into new images

When installing software to cluster images, there are various cases to consider.

Some scenarios involve installing the software update directly on a running flat compute or leader node. Depending on your site requirements, the direct installation onto a running node might be preferred because it avoids the overhead associated with reimaging nodes.

The following topics describe various scenarios:

- [Installing packages from repositories into an image](#) on page 213
- [Installing miscellaneous RPMs into an image](#) on page 214
- [Installing packages from repositories onto running compute nodes or leader nodes](#) on page 215
- [Installing miscellaneous RPMs onto running compute nodes or leader nodes](#) on page 215

## Installing packages from repositories into an image

After you select a repository, you can add one or more packages from the repository to an existing image. Use the `cinstallman` command in one of the following formats:

- On RHEL platforms, use the following command:  
`cinstallman --yum-image [--duk] --image image install packages`
- On SLES platforms, use the following command:  
`cinstallman --zypper-image [--duk] --image image install packages`

The variables are as follows:

---

Variable	Specification
--duk	This parameter is conditional.  Use this parameter if you want to update an image, but you do not want to update the kernel. When you specify <code>--duk</code> , you might save some processing time. By default, the cluster manager updates the kernel in an image when you add a package to an image.
<i>image</i>	The image into which you want to install the packages.
<i>packages</i>	One or more packages from the selected repositories.

---

For example:

```
# cinstallman --yum-image --duk --image ice-rhel7.4 install \
additional_new_package
```

## Installing miscellaneous RPMs into an image

The following procedure explains how to install a miscellaneous set of RPMs into an existing image.

### Procedure

1. Log into the admin node as the root user, and enter the following command to switch to the images directory:

```
# cd /opt/clmgr/image/images/
```

2. Copy the RPMs you want to add to the desired image.

For example, if your RPMs are in `/tmp/newrpm.rpm` and you want to update the `rhel7.4-new` image, enter the following:

```
# cp /tmp/newrpm.rpm rhel7.4-new/tmp
```

3. Change to the image environment.

For example:

```
# chroot rhel7.4-new
```

4. Install the RPMs.

For example:

```
# rpm -Uvh /tmp/newrpm.rpm
```

5. Enter the following command to exit the `chroot` environment:

```
# exit
```

## Installing packages from repositories onto running compute nodes or leader nodes

Rather than installing a software package in an existing image on the admin node, you can install the package directly on a running flat compute or leader node. Later, you might want to capture the image from the running node and store the image on the admin node for further use.

To install packages from repositories already selected, use the `cinstallman` command in one of the following formats:

- On RHEL platforms, use the following format:

```
cinstallman --yum-node --node flat_leaders install packages
```

- On SLES platforms, use the following format:

```
cinstallman --zypper-node --node flat_leaders install packages
```

The variables are as follows:

Variable	Specification
<i>flat_leaders</i>	The compute nodes to receive the packages.
<i>packages</i>	One or more of the packages from the selected repositories.

For example, the following command installs package `zlib-devel` on SLES compute node `service0`:

```
# cinstallman --zypper-node --node service0 install zlib-devel
```

For related information, see the following:

[Capturing an image from a running compute node](#) on page 212

## Installing miscellaneous RPMs onto running compute nodes or leader nodes

You can install any set of RPMs onto a running flat compute or leader node. Later, you can capture and store the image. In this case, use standard Linux tools to manually install the RPMs.

The following general procedure explains how to install the RPMs:

### Procedure

1. Log into the compute node.
2. Copy the RPMs you need to the following directory:  
`/tmp/your_rpm_directory`
3. Use the following command to install the RPMs:  
`rpm -Uvh /tmp/your_rpm_directory`

## Associating nondefault images with targeted nodes

Unless otherwise instructed, the cluster manager uses the default images for cluster nodes at boot time. The default images are determined by node type.

The following topics explain how to change the default image-node association for one or more nodes:

- [\*\*Associating a nondefault image with flat compute nodes and leader nodes\*\*](#) on page 216
- [\*\*Associating a nondefault image with ICE compute nodes\*\*](#) on page 216

For related information, see the following:

[\*\*Node types and default image names\*\*](#) on page 202

## Associating a nondefault image with flat compute nodes and leader nodes

To associate a nondefault image with one or more flat compute or leader nodes, use the `cinstallman` command in the following format:

```
cinstallman --assign-image --node nodes --image new_image [--kernel version]
```

The variables are as follows:

Variable	Specification
<i>nodes</i>	The name of one or more cluster nodes.
<i>new_image</i>	The image to be associated with the <i>nodes</i> .
<i>version</i>	The kernel to be used.

Example 1. The following example command assigns image `sles12sp2-new` to all flat compute nodes:

```
# cinstallman --assign-image --node "service*" --image sles12sp3-new
```

Example 2. The following example assigns image `lead-rhel7.4-nis` with kernel `2.6.32-504.el6.x86_64` to all leader nodes:

```
# cinstallman --assign-image --node "r*lead" --image lead-rhel7.4-nis \
--kernel 2.6.32-504.el6.x86_64
```

Example 3. The following `cinstallman` command returns the available kernels. As the output shows, the associated kernel for an image follows the kernel name.

```
# cinstallman --show-images
Image Name          BT VCS Compat_Distro
ice-rhel7           1  1   rhel7
                  2.6.32-504.el6.x86_64
lead-sles12sp3      0  1   sles12
                  3.0.76-0.11-default
lead-rhel7.4        0  1   rhel7
                  2.6.32-504.el6.x86_64
sles12sp3           0  1   sles12
                  3.0.76-0.11-default
ice-sles12sp3       1  1   sles12
                  3.0.76-0.11-default
```

## Associating a nondefault image with ICE compute nodes

To associate a nondefault image with one or more ICE compute nodes, use the `cimage` command in the following format:

```
cimage --set new_image [kernel_version] nodes
```

The variables are as follows:

---

Variable	Specification
<i>new_image</i>	The image to be associated with the nodes.
<i>kernel_version</i>	The kernel to be used. <i>n</i>
<i>nodes</i>	One or more cluster nodes.

---

The following command assigns image `ice-sles11-new` to all ICE compute nodes:

```
# cimage --set ice-sles11-new "r*i*n*"
```

To change the default kernel association, use the `cimage --show-images` command to see the available image-kernel combinations.

## Pushing images from the admin node to the targeted nodes

You can push nondefault images onto nodes after you complete the following tasks:

- You added updates to the images
- You associated the updated images with the appropriate nodes

The act of pushing images to nodes is also called **provisioning**. The procedure for pushing out the images varies with node type.

The following topics explain how to push images to nodes:

- [Pushing images to flat compute nodes and leader nodes](#) on page 217
- [Pushing images to ICE compute nodes](#) on page 218

---

**NOTE:** For reasons related to node type, security, and performance, the pushing procedures use the default transport method for provisioning. For information about other transport methods for other situations, see the following:

[HPE Performance Cluster Manager Installation Guide](#)

## Pushing images to flat compute nodes and leader nodes

The following procedure describes how to push images from the admin node to flat compute nodes or leader nodes.

### Procedure

1. Use the `cinstallman` command in the following format to schedule the push for the next boot:

```
cinstallman --next-boot image --node target_nodes
```

For example, the following command pushes changes to all flat compute nodes on the next reboot:

```
# cinstallman --next-boot image --node "service*"
```

2. Reboot the node.

Use the `cpower` command in the following format:

```
# cpower node reboot target_nodes
```

## Pushing images to ICE compute nodes

The following procedure describes how to push ICE compute nodes images from the admin node to targeted ICE compute nodes.

### Procedure

1. Enter the command in the following format to stop the targeted ICE compute nodes:

```
cpower node halt target_nodes
```

For `target_nodes`, specify the node names. To specify all ICE compute nodes, specify "`r*i*n*`". To specify only selected nodes, specify `rxixnx`. Use integers, integer ranges, or wildcards for the `x` characters.

For example, the following command stops all ICE compute nodes:

```
# cpower node halt "r*i*n*"
```

2. Use the `cimage` command in the following format to push the new image to the affected rack leaders:

```
cimage --push-rack image_name target_racks
```

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

For example, the following command pushes changes to all the ICE compute nodes:

```
# cimage --push-rack ice-rhel7.4-new "r*"
```

3. Enter the command in the following format to power up the targeted ICE compute nodes:

```
cpower node on target_nodes
```

For example, enter the following command to power up all ICE compute nodes:

```
# cpower node on "r*i*n*"
```

The leader nodes send the new images for the ICE compute nodes when the ICE compute nodes reboot.

## Miscellaneous image management tasks

The following topics explain how to complete various image management tasks:

- [Using a custom repository for site packages](#) on page 219
- [Creating images in an environment with multiple operating systems](#) on page 221
- [Comparing the image on a running node with images on the admin node](#) on page 223
- [Performing ICE compute node per-host customization](#) on page 224
- [Changing the services on the ICE compute nodes](#) on page 224
- [Using the cimage command to manage ICE compute node images](#) on page 225
- [Using cinstallman to install packages into software images](#) on page 227
- [Creating ICE compute node images](#) on page 229

## Using a custom repository for site packages

You can maintain software packages specific to your site and have them available to the `crepo` command. HPE recommends that you put site-specific packages in a separate location. Do not store site-specific packages in the same location as cluster manager packages or operating system packages.

The following procedure describes how to create a custom repository.

### Procedure

1. Use the `crepo` command in the following format to create a site-specific repository and add the contents of the new directory to the repository:

```
crepo --add location_site-specific_packages --custom 'repository_name'
```

For *location\_site-specific\_packages*, specify the location using a URL, an NFS path (*host:/path*), or a path local to the admin node. The destination must be populated already with the required RPMs.

For *repository\_name*, create a name for the new site-specific repository.

This command also creates the required repository metadata.

For example:

```
# crepo --add /opt/clmgr/repos/site-local/site-rpms --custom 'Site-RPMs'
```

2. Use `crepo` command in the following format to make the new site-specific repository available to the `cinstallman` command:

```
crepo --select repository_name
```

For example:

```
# crepo --select Site-RPMs
```

3. (Optional) Add the new RPM base names to an existing RPM list.

This step makes your site-specific RPMs available by default when you create node images in the future.

The substeps are as follows:

- Use the `cp` command to copy an existing generated RPM list.
- Open the new RPM list file with a text editor.
- Add each new RPM to the file on individual lines.
- Save and close the file.

For example, assume that you want to add the following site-specific RPMs to the RPM list called `generated-rhel7.4.rpmlist`:

```
kernel-debug-debuginfo-2.6.32-431.el6.x86_64.rpm  
kernel-debuginfo-2.6.32-431.el6.x86_64.rpm  
kernel-debuginfo-common-x86_64-2.6.32-431.el6.x86_64.rpm
```

Complete the following steps:

- Enter the following commands:

```
# cp /opt/clmgr/image/rpmlists/generated-rhel7.4.rpmlist \
/opt/clmgr/image/rpmlists/site-rhel7.4.rpmlist
# vi /opt/clmgr/image/rpmlists/site-rhel7.4.rpmlists
```
  - Use the `vi` editor to add the following lines to file `site-rhel7.4.rpmlists`:

```
kernel-debug-debuginfo
kernel-debuginfo
kernel-debuginfo-common
```
  - Save and close the file.
4. Use the `cinstallman` command to install the new packages into an image, onto a node, or into a new image that contains these packages.
- To install the new packages into an existing ICE compute node image, use one of the following formats:
    - RHEL:

```
cinstallman --yum-image --image image install package package ...
```
    - SLES:

```
cinstallman --zypper-image --image image install package package ...
```
- The variables are as follows:
- 
- | Variable       | Specification   |
|----------------|---|
| <i>image</i>   | The name of image that you want to install into the packages. |
| <i>package</i> | One or more of the packages you wrote to the repository.      |
- 
- For example:
- ```
# cinstallman --yum-image --image ice-rhel7.4 install \
kernel-debuginfo kernel-debug-debuginfo kernel-debuginfo-common
```
- If necessary, enter the `cimage --show-images` command to retrieve a list of existing images.

- To install the new packages onto a running flat compute node, use one of the following formats:
  - RHEL:

```
cinstallman --yum-node --node node_ID package package ...
```

The following is a RHEL example:

```
# cinstallman --yum-node --node service0 install \
kernel-debuginfo kernel-debug-debuginfo kernel-debuginfo-common
```

- SLES:

```
cinstallman --zypper-node --node node_ID package package ...
```

For *node\_ID*, specify the node ID of the flat compute node.

- To create an image that includes the packages, use the following format:

```
cinstallman --create-image --image new_image --rpmlist path
```

The variables are as follows:

| Variable         | Specification                   |
|------------------|---------------------------------|
| <i>new_image</i> | A name for the new image.       |
| <i>path</i>      | The full path to the new image. |

For example:

```
# cinstallman --create-image --image my-image \
--rpmlist /opt/clmgr/image/rpmlists/site-rhel7.4.rpmlists
```

## 5. (Conditional) Push the changes to the desired nodes.

For information, see the following:

[Pushing images from the admin node to the targeted nodes](#) on page 217

## Creating images in an environment with multiple operating systems

The cluster manager supports the ability for cluster nodes to have different operating systems. For such clusters, images management is similar to image management in a homogeneous cluster. After initial installation, the default images on the admin node reflect the OS running on the admin node. For instance, if you are running SLES 12 on the admin node, the default images for the cluster nodes in the image directory are the following:

- lead-sles12
- sles12
- ice-sles12

To use images for another OS in the cluster, create the default images for that OS.

The following procedure explains how to create CentOS cluster node images while running SLES on the admin node. The procedure uses SLES 12 for the admin node OS. The procedure builds CentOS 7.4 cluster node images.

## Procedure

1. Enter the following command to see the repositories that are currently selected:

```
admin:~ # crepo --show
* HPE-MPI-1.2-sles12sp3 : /opt/clmgr/repos/cm/HPE-MPI-1.2-sles12sp3
* sles12sp3-updates : /opt/clmgr/repos/updates/sles12sp3-updates
* Cluster-Manager-1.0.0-sles12sp3 : /opt/clmgr/repos/cm/Cluster-Manager-1.0.0-sles12sp3
* SUSE-Linux-Enterprise-Server-12-SP3 : /opt/clmgr/repos/distro/sles12-SP3
* SUSE-Linux-Enterprise-High-Availability-Extension-12 : /opt/clmgr/repos/other/sle-ha12
```

2. Unselect the SLES 12 repositories, as follows:

```
crepo --unsel Clu*sles12
crepo --unsel sles12-updates
crepo --unsel SUSE*
crepo --unsel other_sles12_repo_if_any
crepo --unsel other_sles12_repo_if_any
.
.
.
```

3. Confirm that the SLES12 repositories are no longer selected, as follows:

```
admin:~ # crepo --show
HPE-MPI-1.2-sles12sp3 : /opt/clmgr/repos/cm/HPE-MPI-1.2-sles12sp3
Cluster-Manager-1.0.0-sles12sp3 : /opt/clmgr/repos/cm/Cluster-Manager-1.0.0-sles12sp3
sles12sp3-updates : /opt/clmgr/repos/updates/sles12sp3-updates
SUSE-Linux-Enterprise-Server-12-SP3 : /opt/clmgr/repos/distro/sles12sp3
SUSE-Linux-Enterprise-High-Availability-Extension-12 : /opt/clmgr/repos/other/sle-ha12
```

4. Add the CentOS (and related RHEL) repositories, as follows:

```
crepo --add /mirrors/centos/7.4/x86_64/CentOS-7.4-x86_64-bin-DVD1.iso
crepo --add /mirrors/sgi/dist.enqr/latest-rhel6-x86_64/ISO/smc-3.5.0-cd1-media-rhel6-x86_64.iso
crepo --add other_rhel7.4_centos7.4_iso_if_any
crepo --add other_rhel7.4_centos7.4_iso_if_any
.
.
```

5. Confirm the additions, as follows:

```
admin:~ # crepo --show
Cluster-Manager-1.0.0-rhel6 : /opt/clmgr/repos/cm/Cluster-Manager-1.0.0-rhel6
CentOS-7.4 : /opt/clmgr/repos/distro/centos7.4
Cluster-Manager-1.0.0-sles12sp3 : /opt/clmgr/repos/cm/Cluster-Manager-1.0.0-sles12sp3
SUSE-Linux-Enterprise-High-Availability-Extension-12 : /opt/clmgr/repos/other/sle-ha12
SUSE-Linux-Enterprise-Server-12-SP3 : /opt/clmgr/repos/distro/sles12sp3
HPE-MPI-1.2-sles12sp3 : /opt/clmgr/repos/cm/HPE-MPI-1.2-sles12sp3
```

6. Use the following commands to select the CentOS and related RHEL repositories:

```
crepo --sel Clu*rhel6
crepo --sel CentOS-7.4
crepo --sel other_rhel7.4_centos7.4_repo_if_any
crepo --sel other_rhel7.4_centos7.4_repo_if_any
.
.
```

7. Confirm the selections, as follows:

```
admin:~ # crepo --show
* Cluster-Manager-1.0.0-rhel6 : /opt/clmgr/repos/cm/Cluster-Manager-1.0.0-rhel6
* Cluster-Manager-1.0.0-sles12sp3 : /opt/clmgr/repos/cm/Cluster-Manager-1.0.0-sles12sp3
```

```

sles12sp3-updates : /opt/clmgr/repos/updates/sles12sp3-updates
* CentOS-7.4 : /opt/clmgr/repos/distro/centos7.4
SUSE-Linux-Enterprise-High-Availability-Extension-12 : /opt/clmgr/repos/other/sle-ha12
HPE-MPI-1.2-sles12sp3 : /opt/clmgr/repos/cm/HPE-MPI-1.2-sles12sp3
SUSE-Linux-Enterprise-Server-12-SP3 : /opt/clmgr/repos/distro/sles12sp3

```

The selection process generates the necessary RPMs to build the new default images. The following RPM lists are placed in directory /opt/clmgr/image/rpmlists:

```

generated-centos7.4.rpmlist
generated-ice-centos7.4.rpmlist
generated-lead-centos7.4.rpmlist

```

8. Create the default images for CentOS using the appropriate RPM lists as follows:

```

# cinstallman --create-image --image centos7.4 \
--rpmlist /opt/clmgr/image/rpmlists/generated-centos7.4.rpmlist
# cinstallman --create-image --image ice-centos7.4 \
--rpmlist /opt/clmgr/image/rpmlists/generated-ice-centos7.4.rpmlist
# cinstallman --create-image --image lead-centos7.4 \
--rpmlist /opt/clmgr/image/rpmlists/generated-lead-centos7.4.rpmlist

```

The cluster manager adds these new images to the admin node image directory. At this point, you can selectively use these new CentOS images for your cluster nodes as you would any other image.

## Comparing the image on a running node with images on the admin node

You might want to compare the image on a running node with the image on the admin node. The following are some situations in which this comparison can be helpful:

- You suspect that the image on the running node is somehow corrupted or has undesirable content.
- You want to confirm that you have added required packages to the running node.
- You are looking for a node that hosts an image that you want to use to reimagine other nodes.
- Over a time, you are unsure about untracked changes to the image on the running node.

The cluster manager can display RPM and file differences in the following file:

/var/log/cinstallman-idiff.log

Additionally, you can customize the comparison by supplying a **whitelist file**. A whitelist file consists of files and directories to include in the comparison.

---

**NOTE:** This feature applies only to flat compute nodes and leader nodes. However, you can use any image from the image directory on the admin node for the comparison. The node type of the image can be different from the node type of the running node.

---

To perform the comparison, use the `cinstallman` command in the following format:

`cinstallman --idiff --node running_node --image base_image [--rev n] [--file whitelist_file]`

The variables are as follows:

| Variable            | Specification                                                         |
|---------------------|-----------------------------------------------------------------------|
| <i>running_node</i> | The name of the running node for the comparison.                      |
| <i>base_image</i>   | The name of the image to be used from the admin node image directory. |

*Table Continued*

| Variable              | Specification                                                                                                                                                                                                                                                                     |
|-----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>n</i>              | <p>The image revision (an integer) to be used for the comparison. By default, the command uses the highest revision level of the image.</p> <p>To display existing revisions, use the following command:</p> <pre>cinstallman --history --image <i>image_name</i></pre>           |
| <i>whitelist_file</i> | <p>The name of the whitelist file to be used for comparisons. By default, the command uses the following file:</p> <pre>/etc/opt/sgi/idiff-selection-criteria.conf</pre> <p>To build a custom whitelist file, copy the default file to a second file and modify its contents.</p> |

Example 1. The following command compares the image on running node `service0` to the default image for compute nodes for SLES 12 SP3:

```
# cinstallman --idiff --node service0 --image sles12sp3
```

Example 2. The following command runs the same comparison as example 1, but it uses custom whitelist file `/tmp/whitelist1`:

```
# cinstallman --idiff --node service0 --image sles12sp3 \
--file /tmp/whitelist1
```

## Performing ICE compute node per-host customization

You can add per-host ICE compute node customization to the compute node images. To accomplish the customization, add scripts to one of the following directories on the admin node:

- `/opt/sgi/share/per-host-customization/global/`
- `/opt/sgi/share/per-host-customization/mynewimage/`

Scripts in the global directory apply to all ICE compute nodes images. Scripts under the image name apply only to the image in question. The scripts are cycled through once per host when being installed on the leader nodes.

To run all or a subset of the customization scripts upon demand, using the `cimage` command with the `--customizations-only` parameter.

For more information about the customization scripts, see the following file:

```
/opt/sgi/share/per-host-customization/README
```

The following file contains an example global script:

```
/opt/sgi/share/per-host-customization/global/sgi-fstab.sh
```

---

**NOTE:** When creating custom images for ICE compute nodes, make sure to clone the original cluster manager images. You can fall back to the original images if necessary.

---

For more information about the `cimage` command, see the `cimage` manpage.

## Changing the services on the ICE compute nodes

By default, most services are disabled in ICE compute node images. This default condition improves HPE Message Passing Interface (MPI) job performance on ICE compute nodes.

The services file is a configuration framework file. For more information about adjusting configuration framework files, see the following:

#### **About the hierarchical cluster system configuration framework** on page 306

The following procedure explains how to obtain information about the services that run on ICE compute nodes. The procedure also explains how to change the list of active services.

#### **Procedure**

1. Log into the admin node as the root user.
2. Enter the following command to change to the directory where the compute images reside:  

```
# cd /opt/clmgr/image/images
```
3. Enter the `ls` command, and examine the available compute images.

For example:

```
# ls -F
ACHTUNG           DO_NOT_TOUCH_THESE_DIRECTORIES    sles12sp3
ice-sles12sp3    lead-sles12sp3
CUIDADO          README
```

4. Enter the `cd` command in the following format to change to the directory that hosts the compute image services file:

```
cd image_name/etc/opt/sgi/conf.d
```

For *image\_name*, specify one of the compute image names.

For example:

```
# cd ice-sles11sp4/etc/opt/sgi/conf.d
```

5. Use a text editor or text viewer to display `80-compute-distro-services`, the services file.

6. Peruse the file and decide if you want to change the settings for any of the services.

If you want to change any services, complete the rest of this procedure.

7. Copy the original services file to a `.local` version.

For example:

```
# cp 80-compute-distro-services 80-compute-distro-services.local
```

Always edit a copy, not the original.

8. Use a text editor to make the desired changes in the `.local` version of the services file.

9. Enter the following command to propagate the new services file to the ICE compute nodes:

```
# cimage --push-rack image_name "r*"
```

After this command finishes, the cluster manager runs the `.local` version of the services file instead of the original file.

## **Using the `cimage` command to manage ICE compute node images**

The `cimage` command allows you to list, modify, and set software images on the ICE compute nodes in your system.

For a help statement, enter the following command:

```
admin:~ # cimage --help
```

The following examples walk you through some typical `cimage` command operations.

Example 1. The following command lists the available images and their associated kernels:

```
# cimage --show-images
image: ice-sles12sp3
    kernel: 4.4.59-69-default
    kernel: 4.4.59-81
```

Example 2. Assume that you entered the following command:

```
# cimage --show-nodes r1
r1i0n0: ice-sles12sp3 4.4.21-69-default nfs
r1i0n8: ice-sles12sp3 4.4.21-69-default nfs
```

The command lists the following:

- The compute nodes in rack 1
- The image and kernel they are set to boot
- The root file system type (NFS or tmpfs)

Example 3. Assume that you entered the following commands:

```
# cimage --set ice-sles12sp3 4.4.21-81 r1i0n0
# cimage --show-nodes r1
r1i0n0: ice-sles12sp3 4.4.21.81
r1i0n1: ice-sles12sp3 4.4.21.69-default-smp
r1i0n2: ice-sles12sp3 4.4.21.69-default-smp
.
.
.
```

The commands set the `r1i0n0` compute node to boot the `4.4.21-81` kernel from the `ice-sles12sp3` image. The commands also display new node information.

Example 4. The following command sets all nodes in all racks to boot the `4.4.21.81` kernel from the `ice-sles12sp3` image:

```
# cimage --set ice-sles12sp3 4.4.21.81 "r*i*n*"
```

Example 5. The following command sets two ranges of nodes to boot the `4.4.21.81` kernel:

```
# cimage --set ice-sles12sp3 4.4.21.81 "r1i[0-2]n[5-6]" "r1i[2-3]n[0-4]"
```

Example 6. The following commands clone the `ice-sles12sp3` image to a new image and modify the new image:

```
# cinstallman --create-image --clone --source ice-sles11 --image mynewimage
Cloning ice-sles12sp3 to mynewimage ... done
# cp *.rpm /opt/clmgr/image/images/mynewimage/tmp
# chroot /opt/clmgr/image/images/mynewimage/ bash
# rpm -Uvh /tmp/*.rpm
# exit
```

# cloning  
# adds the image and its  
# kernels to the database  
# copy needed RPMS to  
# a temp directory  
# enter the directory  
# install the RPMs  
# exit the chroot

Example 7. If you change the kernels in an image, refresh the kernel database entries for your image.

If you do not change the kernels in the cloned image created, you do not have to refresh the kernel database entries.

The following command refreshes the kernel database entries:

```
# cimage --update-db mynewimage
```

Example 8. The following command pushes new software images to compute blades in a rack or rack set:

```
# cimage --push-rack mynewimage "r*"
rllead: install-image: mynewimage
rllead: install-image: mynewimage done.
```

Example 9. The following command lists the images in the database and the kernels they contain:

```
# cimage --show-images
```

```
image: ice-sles12sp3
      kernel: 4.4.21.81-carlsbad
      kernel: 4.4.21.81-smp

image: mynewimage
      kernel: 4.4.21.81-carlsbad
      kernel: 4.4.21.81-smp
```

Example 10. The following command specifies that a set of compute nodes to boot an image:

```
# cimage --set mynewimage 4.4.21.81-smp "r1i3n*"
```

Reboot the compute nodes to run the new images.

Example 11. The following command completely removes an image you no longer use. The cluster manager removes the image from the admin node and from all compute nodes in all racks:

```
# cimage --del-image mynewimage
rllead: delete-image: mynewimage
rllead: delete-image: mynewimage done.
```

Example 12. You can use the `cimage` command with its `--push-rack` option to specify that the command push out only customization scripts to nodes. When you use the `--customizations-only` option, the command does not include images in the push. In these cases, make sure that you pushed a new image to the nodes before you push the customization scripts. You can specify a list of scripts to run, or you can specify the keyword `all` to direct the cluster manager to run all scripts.

The command format is as follows:

```
cimage --push-rack --customizations-only [script, script, ...]
```

Or

```
cimage --push-rack --customizations-only all
```

For more information about customization, see the `cimage` manpage or see the following:

[\*\*Performing ICE compute node per-host customization\*\*](#) on page 224

## Using `cinstallman` to install packages into software images

A repository is where the following reside:

- HPE Performance Software
- The Linux distribution media
- Any other media or custom repositories you have added

The `cinstallman` command provides a repository list to commands such as `yume`.

---

**NOTE:** Always work with copies of software images.

The `cinstallman` command can update packages within `systemimager` images. You can also use `cinstallman` to install a single package within an image.

However, `cinstallman` works only with the configured repositories. If you install your own RPM, that package must be part of an existing repository. Use the `crepo` command to create a custom repository into which you can group custom packages.

The `cinstallman` command uses the `yum` (RHEL) and `zypper` (SLES) Linux distribution commands to manage package metadata. These commands maintain a cache of the package metadata. If you change the repositories, the caches for the nodes or images you manage can become outdated. When you install packages in images or on nodes, the `cinstallman` command helps you to account for such cache changes with the following options:

- RHEL:

```
--yum-image clean all  
--yum-node clean all
```

- SLES:

```
--zypper-image clean --all  
--zypper-node clean --all
```

The following examples show install the `zlib-devel` package in the flat compute node image. The next time you image or install a compute node, it includes the new package. The commands are as follows:

- RHEL:

```
# cinstallman --yum-image --image my-image install zlib-devel
```

- SLES:

```
# cinstallman --zypper-image --image my-image install zlib-devel
```

---

**NOTE:** Unlike the `install` action, the `cinstallman` command `--update-node` and `--update-image` parameters do not require you to account for the possible cache changes.

You can perform a similar operation for ICE compute node images. Remember the following:

- If you update an ICE compute node image on the admin node, use the `cimage` command to push the changes.

For more information about the `cimage` command, see the following:

**Using the `cimage` command to manage ICE compute node images** on page 225

- If you update a compute node image on the admin node, reimagine or reinstall the compute node. Use the `cinstallman` command to direct the compute node to reimagine itself with a specified image the next time it boots. The parameters you need are as follows:

- `cinstallman --assign-image`
- `cinstallman --next-boot`

## **Creating images for ICE compute nodes and flat compute nodes**

You can use the `cinstallman` command to create ICE compute node and flat compute node images. The command generates a root directory for images automatically. Fresh installations of the cluster manager software create these images during the `configure-cluster` installation step.

RPM lists that control the packages that get installed in the images and are listed in the following directory:

```
/opt/clmgr/image/rpmlists
```

For example, one file could be `/opt/clmgr/image/rpmlists/ice-sles12sp3.rpmlist`.

Do not edit the default lists. The default files are recreated by the `crepo` command when repositories are added or removed. Therefore, use only the default RPM lists as a model for your own.

The following topics explain how to create images:

- [\*\*Creating ICE compute node images\*\*](#) on page 229
- [\*\*Creating flat compute node images\*\*](#) on page 230

For related information, see the following:

[\*\*RPM lists\*\*](#) on page 207

### **Creating ICE compute node images**

The following procedure uses a SLES example and explains how to create an ICE compute node image using the `cinstallman` command.

#### **Procedure**

1. Make a copy of the ICE compute node image RPM list and work on the copy.

For example:

```
# cp /opt/clmgr/image/rpmlists/ice-sles12sp3.rpmlist \
/opt/clmgr/image/rpmlists/my-compute-node.rpmlist
```

2. Add or remove packages from the RPM list.

Keep in mind that needed dependencies are pulled in automatically.

3. Run the `cinstallman` command to create the root.

For example:

```
# cinstallman --create-image --image my-compute-node-image --rpmlist \
/opt/clmgr/image/rpmlists/my-compute-node.rpmlist
```

This example uses the name `my-compute-node-image` as the name.

Output is logged to `/var/log/cinstallman` on the admin node.

The `cinstallman` command makes the new image available to the `cimage` command.

For information about how to use the `cimage` command to push this new image to leader nodes, see the following:

[\*\*Using the cimage command to manage ICE compute node images\*\*](#) on page 225

## Creating flat compute node images

The following procedure uses a SLES example and explains how to create a flat compute node image using the `cinstallman` command.

### Procedure

1. Make a copy of the example flat compute node image RPM list and work on the copy, as follows:

```
# cp /opt/clmgr/image/rpmlists/sles12sp3.rpmlist \
/opt/clmgr/image/rpmlists/my-service-node.rpmlist
```

2. Add or remove any packages from the RPM list.

Keep in mind that needed dependencies are pulled in automatically.

3. Use the `cinstallman` command with the `--create-image` option to create a root directory for the image, as follows:

```
# cinstallman --create-image --image my-service-node-image --rpmlist \
/opt/clmgr/image/rpmlists/my-service-node.rpmlist
```

This example uses `my-service-node-image` as the name of the image.

Output is logged to `/var/log/cinstallman` on the admin node.

After the `cinstallman` command finishes, the image is ready to be used with flat compute nodes. For example:

- You can supply this image as an optional image name on the `discover` command.
- You can use the `cinstallman --assign-image` command to assign this image to an existing flat compute node.
- You can use the `cinstallman --next-boot` command to assign this image to a flat compute node the next time the node boots.

# Using the version control system

Node-specific software resides on the admin node. When you install and configure the cluster software, the installer pushes the node-specific software to each node in the cluster. Over time, you might need to modify the software images. For example, you might need to add a workload manager or file system software.

If you modify the images frequently, your image repository eventually contains several different versions and becomes difficult to manage. As an alternative to managing these images manually, you can use VCS. The cluster manager version control system (VCS) archives, tracks, and manages the various versions of an image. The cluster manager includes an implementation of VCS that uses the `cinstallman` command.

---

**NOTE:** Before you add or modify software, HPE recommends that you back up the original, default software images.

Do not modify the files within the version control system repository. If you edit files in the VCS repository, the integrity of VCS becomes compromised.

---

The following topics explain how to use VCS to manage system images:

- [\*\*VCS terminology\*\*](#) on page 231
- [\*\*Creating images\*\*](#) on page 232
- [\*\*Managing clones\*\*](#) on page 232
- [\*\*Committing the working copy\*\*](#) on page 232
- [\*\*Reverting the working copy to a specified revision\*\*](#) on page 232
- [\*\*Reviewing revision history\*\*](#) on page 233
- [\*\*Reviewing changes between revisions and the working copy\*\*](#) on page 233
- [\*\*Amending a commit message\*\*](#) on page 233
- [\*\*Removing revisions\*\*](#) on page 234
- [\*\*VCS examples\*\*](#) on page 234

## VCS terminology

The following terminology pertains to the image files:

- The **working copy** of an image is the copy that is stored on the admin node in the following directory:  
`/opt/clmgr/image/images/image_name`

The *image\_name* directory contains additional subdirectories and files. The system image includes all the subdirectories and files. The format for the *image\_name* directory name is one of the following:

- `os_name`. For example, `rhel7.4`. This name is the name of the image that can reside on the flat compute nodes in the cluster.
- `lead-os_name`. For example, `lead-rhel7.4`. This name is the name of the image that can reside on the rack leader controllers in the cluster.
- `ice-os_name`. For example, `ice-rhel7.4`. This name is the name of the image that can reside on the ICE compute nodes in the cluster.

When you install cluster software, the installer pushes the working copy image from the admin node to the appropriate nodes in the cluster.

- A **committed copy** of an image is a copy that resides in the VCS repository. It is best to check in, or **commit**, copies of images as you modify them to ensure that modifications are not lost.

## Creating images

When you create an image, it resides in the following directory:

```
/opt/clmgr/image/images/image_name
```

In most cases, after the new image is created, the cluster manager sends a copy to the VCS repository and sets the revision number to 1. This set of events is true, for example, after you run the `configure-cluster` command during installation and configuration. However, if you capture an image from a node, the cluster manager does not automatically check in that image.

For information about how to create images, see the following:

[Creating images to host new software](#) on page 211

## Managing clones

With VCS, cloning works like creating an image. When cloning, you can use the `cinstallman` command parameter called `--rev` to specify that you want a revision of the image to be the source for the clone.

For information about creating images, see the following:

[Creating images](#) on page 232

## Committing the working copy

After you change the working copy of an image, use the `--commit` parameter of the `cinstallman` command to commit your changes to VCS. The working copy of an image resides in the following directory:

```
/opt/clmgr/image/images/image_name
```

The commit requires you to enter a log message. You can specify the log message with the `--msg` option or you can let the command read in the message from the terminal.

## Reverting the working copy to a specified revision

To revert the working copy of an image to a specified revision, use the `--revert` parameter of the `cinstallman` command. This parameter accomplishes the following:

- The parameter removes the working copy of the image.
- The parameter replaces the working copy with a copy of the revision you specify from the VCS repository.

## Reviewing revision history

Each time you commit, the cluster manager adds a log message that notes the associated change. Use the following command to list the revision history of an image or of a range of images:

```
cinstallman --history
```

## Reviewing changes between revisions and the working copy

To display what has changed in an image, use the `--changed` parameter of the `cinstallman` command. When you do not specify a revision, the cluster manager compares the working copy to the highest version checked in to VCS. The working copy resides in the following directory:

```
/opt/clmgr/image/images/image_name
```

Preceding each file in the list of changed files is an 11-character summary of the differences. For an explanation of the 11-character summary, see the description of the `itemize-changes` option on the `rsync(1)` manpage.

The following information describes various ways to compare revisions:

- If you specify a single revision, the cluster manager compares the working copy to the specified revision.
- If you specify a revision range, the cluster manager displays changes between the two revisions.
- To display changes in brief `diff` mode, use the `--cmp-tool` option.
- To display file changes in `diff` format, use one of the following parameters of the `cinstallman` command:
  - `--file`. The `--file` parameter targets a specific file.
  - `--diff-tool`. The `--diff-tool` parameter lets you specify a specific `diff` tool. Make sure that the tool you specify processes arguments the same way that `diff` command processes arguments.

## Amending a commit message

To adjust the commit message of a committed change, use the following command:

```
cinstallman --commit-msg
```

If you do not specify the `--commit-msg` parameter, the command reads the message from the terminal.

# Removing revisions

The `--del-revisions` parameter to the `cinstallman` command deletes all stored revisions but leaves the working copy. This parameter can be useful if you want to free space used by revisions or want to start over with the revision history.

The following two commands free all space used in the revision history and then commit a new first revision:

```
# cinstallman --del-revisions --image myimage
# cinstallman --commit --image myimage --msg "Initial commit"
```

The working copy remains intact and the two revisions effectively collapse into one.

## VCS examples

The VCS examples assume that you logged in as the root user. Long output lines are wrapped for inclusion in this documentation.

The following are the example topics:

- [Adding a revision and querying changes](#) on page 234
- [Reverting to a previous revision](#) on page 235
- [Cloning an image](#) on page 236
- [Deleting all revisions permanently](#) on page 237

## Adding a revision and querying changes

The following example shows how to add the file `test_file` to the compute node image `sles12sp3`.

### Procedure

1. Enter the following command to view the status of image `sles12sp3` in the VCS repository:

```
icicle:~ # cinstallman --history --image sles12sp3
Revision history for image sles12sp3, revisions 1 through 1
-----
Revision: 1, Commit Time: Mon 18 May 2018 11:40:24 AM CDT
=====
Image created using cinstallman.
```

All images you create using the `cinstallman` command are added automatically to VCS as revision 1.

2. Enter the following command to add `test_file` to the working copy of the image:

```
icicle:~ # echo "test file" > \
/opt/clmgr/image/images/sles12sp3/tmp/test_file
```

3. Enter the following command to show the differences between the working copy and revision 1.

```
icicle:~ # cinstallman --changed --image sles12sp3
icicle: cinstallman: Comparing revision 1 and working copy for image sles12sp3...
cmd: rsync -avHix --dry-run --delete /opt/clmgr/image/images//sles12sp3/ \
/opt/clmgr/image/vcs/sles12sp3/1/
sending incremental file list
.d...t..... tmp/
```

```
>f++++++ tmp/test_file  
sent 4961524 bytes received 16926 bytes 1991380.00 bytes/sec total size is 4015834620  
speedup is 806.64 (DRY RUN)
```

The preceding output shows one difference and the addition of file `test_file`.

- Enter the following command to commit the new image that contains file `test_file`:

```
icicle:~ # cinstallman --commit --image sles12sp3 \  
--msg "Added test_file to /tmp"  
icicle: cinstallman: vcs: Using rsync to commit image sles12sp3...  
cmd: rsync -aqHx --link-dest=/opt/clmgr/image/vcs/sles12sp3/1\  
/opt/clmgr/image/images/sles12sp3/ /opt/clmgr/image/vcs/sles12sp3/2/  
icicle: cinstallman: image sles12sp3 committed to vcs, rev: 2
```

- Enter the following command to verify that there are no differences between the working copy and the committed copy:

```
icicle:~ # cinstallman --changed --image sles12sp3  
icicle: cinstallman: Comparing revision 2 and working copy for image sles12sp3...  
cmd: rsync -avHix --dry-run --delete /opt/clmgr/image/images//sles12sp3/ \  
/opt/clmgr/image/vcs/sles12sp3/2/  
sending incremental file list  
  
sent 4961516 bytes received 16918 bytes 1991373.60 bytes/sec total size is 4015834611  
speedup is 806.65 (DRY RUN)
```

- Enter the following command to retrieve the revision history of the image:

```
icicle:~ # cinstallman --history --image sles12sp3  
Revision history for image sles12sp3, revisions 1 through 2  
-----  
Revision: 1, Commit Time: Mon 18 May 2018 11:40:24 AM CDT  
=====  
Image created using cinstallman.  
  
Revision: 2, Commit Time: Wed 20 May 2018 08:17:08 AM CDT  
=====  
Added test_file to /tmp  
  
Done
```

- Enter the following command to display the list of all files changed between revision 1 and revision 2:

```
icicle:~ # cinstallman --changed --image sles12sp3 --rev 1..2  
icicle: cinstallman: Comparing revisions 1 and 2 for image sles12sp3...  
cmd: rsync -avHix --dry-run --delete /opt/clmgr/image/vcs/sles12sp3/2/ \  
/opt/clmgr/image/vcs/sles12sp3/1/  
sending incremental file list  
>f.st..... etc/opt/sgi/vcs-log-entry  
.d..t..... tmp/  
>f++++++ tmp/test_file  
  
sent 5135898 bytes received 16931 bytes 3435219.33 bytes/sec total size is 4015834611  
speedup is 779.35 (DRY RUN)
```

Notice the presence of the `vcs-log-entry` file, which is always modified upon commits.

## Reverting to a previous revision

Assume that you have revised and checked in an image. Later, you decide that you want to revert to a previous image. Use the `cinstallman` command to perform the revert.

## Procedure

1. Enter the following command to declare that you want `sles12sp3` image version 1 software image to be the working copy:

```
icicle:~ # cinstallman --revert --image sles12sp3 --rev 1
icicle: cinstallman: Removing image work dir: /opt/clmgr/image/images/sles12sp3
icicle: cinstallman: vcs: Syncing revision 1 in to place...
cmd: rsync -aqHx /opt/clmgr/image/vcs/sles12sp3/1/ /opt/clmgr/image/images/sles12sp3/
icicle: cinstallman: Working copy of sles12sp3 now at revision 1
```

2. Enter the following command to retrieve the revision history:

```
icicle:~ # cinstallman --history --image sles12sp3
Revision history for image sles12sp3, revisions 1 through 2
-----
Revision: 1, Commit Time: Mon 18 May 2018 11:40:24 AM CDT
=====
Image created using cinstallman.

Revision: 2, Commit Time: Wed 20 May 2018 08:17:08 AM CDT
=====
Added test_file to /tmp

Done
```

As the preceding output shows, reverting the working copy does not affect the revision history.

3. To make the working copy correspond to the highest revision (the normal order of things), you can use the following command to commit the current working image (same content as revision 1):

```
icicle:~ # cinstallman --commit --image sles12sp3 \
--msg "Copy of image minus test_file in /tmp."
icicle: cinstallman: vcs: Using rsync to commit image sles12sp3...
cmd: rsync -aqHx --link-dest=/opt/clmgr/image/vcs/sles12sp3/2\
 /opt/clmgr/image/images/sles12sp3/ /opt/clmgr/image/vcs/sles12sp3/3/
icicle: cinstallman: image sles12sp3 committed to vcs, rev: 3
```

4. Enter the following command to retrieve the revision history:

```
icicle:~ # cinstallman --history --image sles12sp3
Revision history for image sles12sp3, revisions 1 through 3
-----
Revision: 1, Commit Time: Mon 18 May 2018 11:40:24 AM CDT
=====
Image created using cinstallman.

Revision: 2, Commit Time: Wed 20 May 2018 08:17:08 AM CDT
=====
Added test_file to /tmp

Revision: 3, Commit Time: Wed 20 May 2018 08:25:39 AM CDT
=====
Copy of image minus test_file in /tmp.

Done
```

## Cloning an image

The following example shows how to clone an image based on one of the previous revision images.

## Procedure

1. Enter the following command to clone an image based on revision 2, which includes `test_file`:

```
icicle:~ # cinstallman --create-image --clone --source sles12sp3 --rev 2 \
--image sles12sp3+test_file
About to use mksimage --Copy to clone the image...
icicle: cinstallman: vcs: Syncing revision 2 of sles12sp3 to new sles12sp3+test_file ...
cmd: rsync -aqHx /opt/clmgr/image/vcs/sles12sp3/2/ \
/opt/clmgr/image/images/sles12sp3+test_file/
icicle: cinstallman: Working copy of sles12sp3+test_file now at revision 2 of image sles12sp3
Ran sgi-mkautoinstallscript
.
.
```

Notice that the `--rev 2` argument in the preceding command directs the `cinstallman` command to create the clone from revision 2.

2. Enter the following command to verify the images that exist on the admin node after the cloning operation:

```
icicle:~ # cinstallman --show-images
Image Name                                BT VCS Compat_Distro
sles12sp3                                    0  1    sles12
        4.4.59-92.17-default
sles12sp3+test_file                         0  1    sles12
        4.4.59-92.17-default
lead-sles12sp3                            0  1    sles12
        4.4.59-92.17-default
ice-sles12sp3                             0  1    sles12
        4.4.59-92.17-default
```

3. Enter the following command to retrieve the revision history:

```
icicle:~ # cinstallman --history --image sles12sp3+test_file
Revision history for image sles12sp3+test_file, revisions 1 through 1
-----
Revision: 1, Commit Time: Wed 20 May 2018 08:29:07 AM CDT
=====
Image clone of sles12sp3 by cinstallman
Done
```

## Deleting all revisions permanently

The following procedure explains how to use the `cinstallman` command to permanently delete all revisions.

## Procedure

1. Enter the following command to delete all revisions:

```
icicle:~ # cinstallman --del-revisions --image sles12sp3
Removing all revisions of sles12sp3, leaving the working copy...
```

2. Enter the following command to retrieve the revision history and verify the deletion:

```
icicle:~ # cinstallman --history --image sles12sp3
icicle: cinstallman: There are no checked in revisions of this image.
Image history failed. See above for error messages
```

3. (Optional) Enter the following command to commit the current image to the version control system:

```
icicle:~ # cinstallman --commit --image sles12sp3 --msg "A new beginning :)"  
icicle: cinstallman: vcs: First revision, rsync the image work dir to the first revision...  
cmd: rsync -aqHx /opt/clmgr/image/images/sles12sp3/ /opt/clmgr/image/vcs/sles12sp3/1/
```

4. (Optional) Enter the following command to verify the commit:

```
icicle:~ # cinstallman --history --image sles12sp3  
Revision history for image sles12sp3, revisions 1 through 1  
-----  
Revision: 1, Commit Time: Wed 20 May 2018 08:32:55 AM CDT  
=====  
A new beginning :)
```

Done

# Fabric management

Your HPE cluster implements one of the following network topologies:

- Hypercube
- Enhanced Hypercube
- All-to-All
- Fat Tree

The connective software that supports the cluster topology is a fabric network. The fabric network includes internal InfiniBand switches or Omni-Path switches. These switches are located in the individual rack units (IRUs). Each system is configured with one or two separate *fabrics* or **subnets**. The documentation typically refers to these subnets as `ib0` and `ib1`.

On compute nodes, there might be several interfaces called `ib0`, `ib1`, and so on. Each interface might be connected to the same subnet.

The fabric network is one of the following:

- Intel Omni-Path Fabric Network

This chapter introduces the Intel Omni-Path fabric network and describes commands you can use to manage the Omni-Path fabric.

For more information about the Omni-Path fabric network, see the documentation from Intel Corporation at the following link:

<http://www.intel.com/support/network/omni-adptr/sb/CS-035857.htm>

- InfiniBand Fabric Network

The InfiniBand technology facilitates fast communication between the compute nodes within a rack and the compute nodes in separate racks. The InfiniBand network uses Open Fabrics Enterprise Distribution (OFED) software to monitor and control the InfiniBand fabric. For information about OFED, see <http://www.openfabrics.org>.

The cluster system uses a distributed memory scheme. Parallel processes in an application pass messages, and each process has its own dedicated processor and address space. By default, the HPE Message Passing Interface (MPI) software uses only the `ib0` subnet. Typically, storage uses the `ib1` subnet. Other InfiniBand configurations are possible and can lead to better performance with specific workloads. For example, you can configure the MPI from HPE library, the Message Passing Toolkit (MPT), to use one or two InfiniBand subnets to optimize application performance.

For information about the HPE Message Passing Interface (MPI) and MPT software, see the following:

**[HPE Message Passing Interface \(MPI\) User Guide](#)**

The following topics explain how to manage network fabrics:

- **[Omni-Path fabric management](#)** on page 240
- **[InfiniBand fabric management](#)** on page 242
- **[Utilities and diagnostics for Omni-Path fabrics and InfiniBand fabrics](#)** on page 249

# Omni-Path fabric management

The following topics explain how to manage the Omni-Path fabric management software:

- [Starting and stopping the Omni-Path fabric managers](#) on page 240
- [Managing Omni-Path fabric software](#) on page 240

## Starting and stopping the Omni-Path fabric managers

- The following command starts the Omni-Path fabric managers:

```
# systemctl start opafm
```

- Each fabric manager instance is a separate process. There is no interdependency among fabric manager instances. To start or stop an individual instance, enter one of the following commands:

```
◦ /usr/lib/opa-fm/bin/opafmctrl start -i fabric_manager_instance
```

```
◦ /usr/lib/opa-fm/bin/opafmctrl stop -i fabric_manager_instance
```

For *fabric\_manager\_instance*, specify 0 for the first plane or 1 for the second plane.

- The following command stops all the running Omni-Path fabric managers:

```
# systemctl stop opafm
```

## Managing Omni-Path fabric software

The following examples show the commands that you can use to manage Omni-Path fabric software. These commands assume that the Omni-Path fabric software is installed and running as expected. If these commands fail, install the Omni-Path software on the servers.

Example 1. The following command displays the Omni-Path software version:

```
[root@service0 ~]# opaconfig -v  
10.6.0.0.134
```

Example 2. The ibstatus command checks the state of a node and tests to see if an HFI is installed on the admin node. You can run this command from any node.

The following command shows that the link is up and that there is a single host fabric interface (HFI) adapter on the node named service0:

```
[root@service0 ~]# ibstatus  
Infiniband device 'hfil_0' port 1 status:  
    default gid: fe80:0000:0000:0000:0011:7501:0167:1ecd  
    base lid: 0xa  
    sm lid: 0x1  
    state: 4: ACTIVE  
    phys state: 5: LinkUp  
    rate: 100 Gb/sec (4X EDR)  
    link_layer: InfiniBand
```

The following command shows two adapters installed on the same leader node:

```
r1lead:~ # ibstatus  
Infiniband device 'hfil_0' port 1 status:
```

```

default gid:      fe80:0000:0000:0000:0011:7501:0163:ce82
base lid:        0xa
sm lid:         0xa
state:          4: ACTIVE
phys state:     5: LinkUp
rate:           100 Gb/sec (4X EDR)
link_layer:     InfiniBand

Infiniband device 'hf1_1' port 1 status:
default gid:      fe80:0000:0000:0001:0011:7501:0163:cea3
base lid:        0x11a
sm lid:         0x11a
state:          4: ACTIVE
phys state:     5: LinkUp
rate:           100 Gb/sec (4X EDR)
link_layer:     InfiniBand

```

**Example 3.** You can check the state of the fabric manager running on the leader or admin node. If there are two separate fabrics, then SM 0 and SM 1 display as running. The command is as follows:

```

rllead:~ # /usr/lib/opa-fm/bin/opafmctrl status
Checking IFS Fabric Manager
Checking SM 0: fm0_sm: Running
Checking FE 0: fm0_fe: Disabled
Checking SM 1: fm1_sm: Running
Checking FE 1: fm1_fe: Disabled
Checking SM 2: fm2_sm: Disabled
Checking FE 2: fm2_fe: Disabled
Checking SM 3: fm3_sm: Disabled
Checking FE 3: fm3_fe: Disabled

```

**Example 4.** You can use `systemctl` commands, such as the following, to check the state of the fabric:

```

# systemctl status opafm
opafm.service - OPA Fabric Manager
   Loaded: loaded (/usr/lib/systemd/system/opafm.service; disabled; vendor preset: disabled)
   Active: active (running) since Thu 2018-04-26 14:52:29 CDT; 7min ago
     Process: 5186 ExecStart=/usr/lib/opa-fm/bin/opafmd -D (code=exited, status=0/SUCCESS)
    Main PID: 5190 (opafmd)
      Tasks: 17 (limit: 512)
     CGroup: /system.slice/opafm.service
             └─5190 /usr/lib/opa-fm/bin/opafmd -D
                   ├─5191 /usr/lib/opa-fm/runtime/sm -e sm_0

```

As the following command shows, for two fabrics, the `systemctl` command hints at two subnet managers running:

```

# systemctl status opafm
opafm.service - OPA Fabric Manager
   Loaded: loaded (/usr/lib/systemd/system/opafm.service; enabled; vendor preset: disabled)
   Active: active (running) since Fri 2018-04-20 11:29:17 CDT; 6 days ago
    Main PID: 1747 (opafmd)
       CGroup: /system.slice/opafm.service
                 ├─1747 /usr/lib/opa-fm/bin/opafmd -D
                 ├─1788 /usr/lib/opa-fm/runtime/sm -e sm_0
                 └─1789 /usr/lib/opa-fm/runtime/sm -e sm_1

```

**Example 5.** The following command checks the count of the switches on the fabric:

```

rllead:~ # opareport -o lids -q -Q -F nodetype:SW
LID Summary
283 LID(s) in Fabric:
          LID(Range) NodeGUID          Port Type Name

```

|        |                    |                      |
|--------|--------------------|----------------------|
| 0x0002 | 0x00117501026776dd | 0 SW SGI Switch Node |
| 0x0003 | 0x00117501026775f3 | 0 SW SGI Switch Node |
| 0x0004 | 0x0011750102677602 | 0 SW SGI Switch Node |
| 0x0005 | 0x0011750102677702 | 0 SW SGI Switch Node |
| 0x0006 | 0x00117501026776d9 | 0 SW SGI Switch Node |
| 0x0007 | 0x001175010267770a | 0 SW SGI Switch Node |
| 0x0008 | 0x00117501026776fc | 0 SW SGI Switch Node |
| 0x0009 | 0x00117501026776f1 | 0 SW SGI Switch Node |
| 0x000b | 0x00117501026776e0 | 0 SW SGI Switch Node |
| 0x000c | 0x00117501026776f6 | 0 SW SGI Switch Node |
| 0x000d | 0x00117501026776d4 | 0 SW SGI Switch Node |
| 0x000e | 0x0011750102677723 | 0 SW SGI Switch Node |
| 0x000f | 0x00117501026776eb | 0 SW SGI Switch Node |
| 0x0010 | 0x00117501026776f4 | 0 SW SGI Switch Node |
| 0x0011 | 0x00117501026776f0 | 0 SW SGI Switch Node |
| 0x0012 | 0x00117501026776f3 | 0 SW SGI Switch Node |

16 Reported LID(s)

The following command checks the count of HFIs on the fabric:

```
r1lead:~ # opareport -o lids -q -Q -F nodetype:FI
LID Summary
283 LID(s) in Fabric:
  LID(Range)  NodeGUID          Port Type Name
0x0001       0x001175010167006e  1 FI service1 hfi1_0
0x000a       0x001175010163ce82  1 FI r1lead hfi1_0
0x0013       0x001175010179e6d5  1 FI r1i3n0 hfi1_1
0x0014       0x001175010179010d  1 FI r1i3n1 hfi1_1
0x0015       0x001175010179e244  1 FI r1i3n2 hfi1_1
0x0016       0x001175010179e246  1 FI r1i3n3 hfi1_1
0x0017       0x001175010179e02f  1 FI r1i3n4 hfi1_1
0x0018       0x001175010179e24c  1 FI r1i3n5 hfi1_1
...
0x0119       0x001175810179e01b  1 FI r1i4n32 hfi1_0
0x011a       0x001175810179e242  1 FI r1i4n33 hfi1_0
0x011b       0x001175810179e27c  1 FI r1i4n34 hfi1_0
0x011c       0x001175810179d92c  1 FI r1i4n35 hfi1_0
267 Reported LID(s)
```

The values reported from the HFI and switch report must equal the number of devices on the fabric. In this case, 16+267=283.

## InfiniBand fabric management

The following topics explain how to manage the InfiniBand fabric management software:

- [InfiniBand fabric overview \(hierarchical clusters\)](#) on page 242
- [Using the InfiniBand management tool GUI](#) on page 243
- [Fabric management commands](#) on page 245
- [Automatic InfiniBand fabric management](#) on page 248

## InfiniBand fabric overview (hierarchical clusters)

HPE supports the OFED OpenSM software package and the `sgifmcli` tool for InfiniBand fabric management.

The InfiniBand fabric connects the compute nodes, leader nodes, and the ICE compute nodes. It does not connect to the admin node or the chassis management control (CMC) blades. Hierarchical systems usually have two separate InfiniBand fabrics, which are referred to as `ib0` and `ib1`.

Each InfiniBand fabric (also sometimes called an InfiniBand subnet) has its own subnet manager, which runs on a leader node. For a system with two or more racks, the subnet manager for each fabric is usually configured to run on different leader nodes. In a single rack system, both subnet managers run on the single leader node. Each subnet manager might also be paired with a standby subnet manager. If the primary subnet manager fails, the standby subnet manager takes over.

Leader nodes do not always have InfiniBand fabric host channel adapters (HCAs). In some cases, one or two leader nodes have HCAs to run the OFED subnet manager. In other cases, subnet management is done on separate fabric management nodes, so no leader nodes have InfiniBand HCAs.

Leader nodes associate a subnet manager instance with a particular port on the leader node. Usually, the following mapping exists:

- `ib0` is mapped to port 1 of the InfiniBand host channel adapter (HCA) on the subnet manager node
- `ib1` is mapped to port 2 of the HCA on the subnet manager node.

The subnet manager for `ib0` and `ib1` is configured using the corresponding `/etc/ofa/opensm-ib[01].conf` file.

---

**NOTE:** After a system reboots, the `opensm` daemons start running automatically.

For information about how to configure the InfiniBand fabric using the GUI, see the following:

**HPE Performance Cluster Manager Installation Guide**

For information about `sgifmcli`, see the following:

**Fabric management commands** on page 245

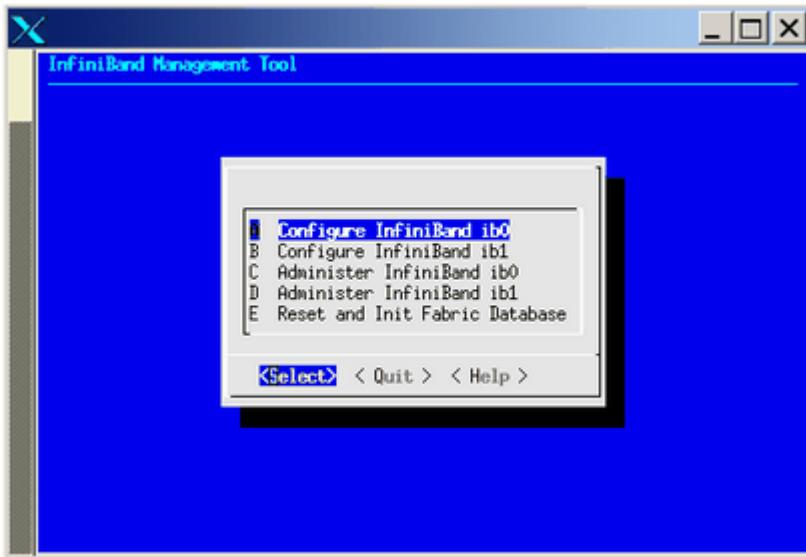
## Using the InfiniBand management tool GUI

You can use the InfiniBand management GUI tool to configure, administer, or verify the InfiniBand fabric on the cluster.

To start the tool, log into the admin node and enter the following command:

```
admin:~ # tempo-configure-fabric
```

The following figure shows the **InfiniBand Management Tool** GUI:



**Figure 52: InfiniBand Management Tool screen**

To highlight and select the action you want, use the following:

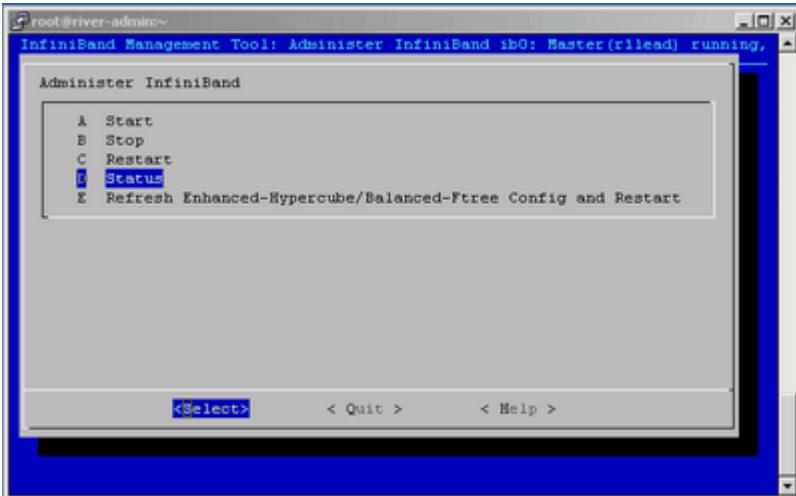
- The mouse
- The keyboard arrow keys
- The **Return** key
- The **Tab**

After you highlight menu choice, the following actions are possible:

- **Select** selects an action and displays to a submenu.
- **Quit** returns to the previous screen.
- **Help** displays online help for each of the GUI actions.

If the `tempo-configure-fabric` command fails in a configuration or administrative operation, use the `sgifmcli` command to debug the problem.

After configuring and bringing up the InfiniBand network, select the **Administer InfiniBand ib0** option or the **Administer InfiniBand ib1** option. You can use this screen to start, stop, restart, or refresh a fabric. You can verify the status through the **Status** option. The following is an example of the GUI after you select **Administer InfiniBand ib0** or **Administer InfiniBand ib1**:



**Figure 53: Administer InfiniBand Status option**

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

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

To return to the configure-cluster GUI, press the **Enter** key.

The **Refresh Enhanced Hypercube Config and Restart** option applies only to the Enhanced Hypercube topology. You are required to refresh the fabric configuration when you either add, remove, or move one or more ICE compute nodes. The refresh action updates the guid routing order file that balances InfiniBand traffic for the Enhanced Hypercube topology. In addition, this action also automatically restarts the master subnet manager and the optional standby subnet manager for the specified fabric. Ideally, perform a refresh action for a fabric when there are no jobs running in the system. Restarting the subnet manager can have an adverse impact on the running jobs in the system.

For information about `sgifmcli`, see the following:

[\*\*Fabric management commands\*\*](#) on page 245

## Fabric management commands

HPE recommends that you use the InfiniBand Management tool GUI for most fabric management operations. To start the GUI, enter the following command at the admin node system prompt:

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

For more information about the GUI, see the following:

[\*\*Using the InfiniBand management tool GUI\*\*](#) on page 243

The following topics explain how to use the fabric management commands:

- [\*\*sgifmcli fabric management command\*\*](#) on page 246
- [\*\*sgifmdb fabric management database command\*\*](#) on page 247

## **sgifmcli fabric management command**

For advanced fabric management, use the `sgifmcli` command. For example, use `sgifmcli` for the following actions:

- Initializing and configuring external InfiniBand switches

To configure an external InfiniBand switch, clusterwide InfiniBand connectivity is not required. The only requirements are as follows:

- The supplied switch host name is resolvable
- A working networking connection to the external InfiniBand switch exists

- Verifying the integrity of the InfiniBand fabric. This activity requires that the fabric is configured properly.

See the `sgifmcli(8)` manpage for the following information:

- Command syntax and examples.
- A list of switches that HPE supports on clusters.
- Information about adding external InfiniBand switches to your cluster fabric.
- Information about fabric verification operation.

The following are additional command examples.

Example 1. The syntax to start a subnet manager master is as follows:

```
sgifmcli --start --id identifier
```

For example, to start the `master_ib0` subnet manager master, enter the following:

```
# sgifmcli --start --id master_ib0
```

At this point, a master for the fabric `ib0` is running on the `r1lead`. The fabric `ib0` is available for compute jobs. If a standby is defined, the command launches both the standby and the master.

Example 2. The syntax to stop a subnet manager master is as follows:

```
sgifmcli --stop --id identifier
```

The following command stops the `master_ib0` subnet manager master running on host `r1lead`:

```
# sgifmcli --stop --id master_ib0
```

If a standby is defined, the standby also stops.

Example 3. The command syntax that checks the status of a subnet manager master is as follows:

```
sgifmcli --status --id identifier
```

The following command displays the status of the `master_ib0` subnet manager master:

```
# sgifmcli --status --id master_ib0
Master SM
Host = rlead
Guid = 0x0002c902002838f5
Fabric = ib0
Topology = hypercube
Routing Engine = dor
OpenSM = running
```

The command reports the status of the master subnet manager master\_ib0 running on host rllead. If a standby is defined, the command reports the status of both the standby and the master.

Example 4. The syntax to remove a subnet manager master is as follows:

```
sgifmcli --remove --id identifier
```

To remove the master\_ib0 subnet manager master, first stop it and then perform the **--remove** option, as follows:

```
# sgifmcli --stop --id master_ib0  
  
# sgifmcli --remove --id master_ib0
```

The subnet manager master is removed from the entity list. If a standby is defined, the command removes both the standby and the master.

Example 5. To remove the standby\_ib0 subnet manager standby, first stop its master. Then, use the **--remove** option, as follows:

```
# sgifmcli --stop --id master_ib0  
# sgifmcli --remove --id standby_ib0
```

The subnet manager standby is removed from the entity list. If a standby has been defined, the command removes both the standby and the master.

Example 6. To find the ID of the master subnet manager in the database, enter the following:

```
# sgifmcli --dump --id ib0 | grep MASTER
```

Example 7. To print the fabric configuration, enter the following:

```
# sgifmcli --showconfig
```

```
-----  
NAME = ib1  
TYPE = ibfabric  
MASTER =  
STANDBY =  
SWITCH_LIST =  
-----  
NAME = ib0  
TYPE = ibfabric  
MASTER =  
STANDBY =  
SWITCH_LIST =
```

Example 8. To list the switches related to a particular fabric, enter the following command:

```
# sgifmcli --switchlist --id fabric
```

## **sgifmdb fabric management database command**

The fabric component maintains a database of managed objects. The database version is automatically set during cluster install. You do not need to set it. Most likely, this database will change over time. To manage multiple database versions, use the **sgifmdb** command, which reports the managed objects database version.

For information about the **sgifmdb** command, enter the following from the admin node:

```
admin:~ # sgifmdb -h  
SGI Fabric Component DB tool  
Usage: db_version [--get|-g] [--dump|-d] [-v|--version] [-r|--reset] [--help|-h]
```

```
-g, --get      Read DB version object from DB
-d, --dump    Dump the DB
-v, --version Print version
-r, --reset   Reset the database and start clean
-h, --help    Show this text
```

## Automatic InfiniBand fabric management

By default, each subnet manager performs a light sweep of the fabric it is managing every 10 seconds. The time is set on the admin node in the `sweep_interval` variable in the following file:

```
/opt/sgi/var/sgifmcli/opensm-ibx.conf.templ
```

If a subnet manager detects a change in the fabric during a light sweep, it performs a **heavy** sweep. Examples of changes are updates such as the addition or deletion of a node. The heavy sweep changes the fabric configuration to reflect the current state of the system.

There is one `opensm` instance for each fabric. Each instance associates itself with a particular globally unique identifier (GUID) for a port on the node upon which `opensm` runs. This association is configured with the `guid` entry in the corresponding `opensm-ibx.conf` file.

---

**NOTE:** If your cluster has more than 256 nodes, increase the `sweep_interval` variable to 90 seconds or more.

To reset the sweep interval from the GUI, do the following:

- Edit the `sweep_interval` variable in the `/opt/sgi/var/sgifmcli/opensm-ib0.conf.templ` file.
- To launch the GUI, run the `tempo-configure-fabric` command.
- Do a **Commit** operation in the GUI.

Alternatively, use the `sgifmcli` command `--arglist` parameter. This parameter sets various subnet manager configuration parameters including the sweep interval.

For example:

```
# sgifmcli --set --id master-ib0 --arglist sweep-interval=90
```

---

For information about fabric sweeping, see the `opensm(8)` manpage on the leader node.

## Network topology

Cluster systems with a hypercube topology use the dimension order routing (DOR) algorithm. The dimension order routing algorithm is based on the min-hop algorithm. The algorithm uses the shortest paths. The goal of using the shortest path is in contrast to other goals. A different goal might spread out traffic across different paths with the same shortest distance. When choosing the shortest path, it chooses among the available shortest paths based on an ordering of dimensions.

Cluster systems with a fat-tree topology use UPDN as the default routing algorithm. Unicast routing algorithm (UPDN) is also based on the minimum hops to each node, but it is constrained to ranking rules.

There are two `opensm` daemons, one for each fabric, `opensmd-ib0` and `opensmd-ib1`. The `init.d` scripts control the `opensm` daemons. Each `init.d` script has a separate configuration file for each fabric, `opensm-ib0` and `opensm-ib1`.

The `sminfo` command shows the GUID of the subnet manager master.

For more information on routing variables, see the `opensm(8)` manpage.

# Utilities and diagnostics for Omni-Path fabrics and InfiniBand fabrics

The diagnostics package on your cluster contains tools and diagnostic software for the Open Fabrics Enterprise Distribution (OFED) software. On hierarchical clusters, these tools reside on the leader nodes in the `/usr/sbin` directory. In addition, the `opensm(8)` manpage describes options that control logging and debugging.

To run the commands in the following topics, log into a node that is attached to the fabrics. The following topics include information about fabric diagnostic software:

- [\*\*ibstat and ibstatus commands \(Omni-Path and InfiniBand\)\*\*](#) on page 249
- [\*\*perfquery command \(InfiniBand\)\*\*](#) on page 250
- [\*\*ibnetdiscover command \(InfiniBand\)\*\*](#) on page 250
- [\*\*ibdiagnet command \(InfiniBand\)\*\*](#) on page 251
- [\*\*Logging and debugging options\*\*](#) on page 252

## ibstat and ibstatus commands (Omni-Path and InfiniBand)

The `ibstat` command displays the status of the host channel adapters (HCAs) in your Omni-Path fabric or InfiniBand fabric. The status includes the HCAs on the leader nodes.

Example 1. The following shows `ibstat` output.

```
rllead:/usr/bin # ibstat
CA 'mlx4_0'
    CA type: MT4099
    Number of ports: 2
    Firmware version: 2.40.5030
    Hardware version: 0
    Node GUID: 0xe41d2d03006f51e0
    System image GUID: 0xe41d2d03006f51e3
    Port 1:
        State: Active
        Physical state: LinkUp
        Rate: 56
        Base lid: 1
        LMC: 0
        SM lid: 1
        Capability mask: 0x0251486a
        Port GUID: 0xe41d2d03006f51e1
        Link layer: InfiniBand
    Port 2:
        State: Active
        Physical state: LinkUp
        Rate: 56
        Base lid: 1
        LMC: 0
        SM lid: 1
        Capability mask: 0x0251486a
        Port GUID: 0xe41d2d03006f51e2
        Link layer: InfiniBand
```

Example 2. The following `ibstatus` command shows the link rate. The `ibstatus` command is more terse than the `ibstat` command.

```
r1lead:/usr/bin # ibstatus
Infiniband device 'mlx5_0' port 1 status:
    default gid:      fec0:0000:0000:0000:e41d:2d03:006f:51e1
    base lid:        0x1
    sm lid:         0x1
    state:          4: ACTIVE
    phys state:     5: LinkUp
    rate:           56 Gb/sec (4X FDR)
    link_layer:     InfiniBand

Infiniband device 'mlx5_0' port 2 status:
    default gid:      fec0:0000:0000:0001:e41d:2d03:006f:51e2
    base lid:        0x1
    sm lid:         0x1
    state:          4: ACTIVE
    phys state:     5: LinkUp
    rate:           56 Gb/sec (4X FDR)
    link_layer:     InfiniBand
```

## **perfquery command (InfiniBand)**

The `perfquery` command finds errors on one or more host channel adapters (HCAs) and errors on switch ports. You can also use `perfquery` to reset HCA and switch port counters.

For example output, enter one or more of the following commands on a node that is attached to the InfiniBand fabric. Typically, a leader node or a flat compute node is attached to the InfiniBand fabric. Example commands are as follows:

- To display command options, enter the following:  
`# perfquery --help`
- To display a list of counters, enter the following:  
`# perfquery`

## **ibnetdiscover command (InfiniBand)**

The `ibnetdiscover` command configures the InfiniBand fabric.

For example output, enter one or more of the following commands on a node that is attached to the InfiniBand fabric. Typically, a leader node or a flat compute node is attached to the InfiniBand fabric. The commands are as follows:

- To display command options, enter the following:  
`# ibnetdiscover --help`
- To display status information, enter the following:  
`# ibnetdiscover`

## **opareport command (Omni-Path)**

The `opareport` command provides Omni-Path fabric analysis and reports. Run this command on a node that is connected to the Intel Omni-Path Fabric with the Intel Omni-Path Fabric Suite FastFabric tool set installed. Typically, this node is a leader node or a flat compute node.

## **ibdiagnet command (InfiniBand)**

The `ibdiagnet` command scans the fabric and extracts information about connectivity and devices.

For example output, type one or more of the following commands on a node that is attached to the InfiniBand fabric. Typically, a leader node or a flat compute node is attached to the InfiniBand fabric. The commands are as follows:

- To display command options, enter the following:

```
# ibdiagnet --help
```

- To display a summary report, enter the following:

```
# ibdiagnet
```

The following example shows how to use `ibdiagnet` to load the fabric for testing.

```
r1lead:/opt/sgi/sbin # ibdiagnet -c 5000
Loading IBDIAGNET from: /usr/lib64/ibdiagnet1.2
Loading IBDM from: /usr/lib64/ibdm1.2
-W- Topology file is not specified.
      Reports regarding cluster links will use direct routes.
-W- A few ports of local device are up.
      Since port-num was not specified (-p option), port 1 of device 1 will be
      used as the local port.
-I- Discovering the subnet ... 10 nodes (2 Switches & 8 CA-s) discovered.

-----
-I- Bad Guids Info
-----
-I- No bad Guids were found

-----
-I- Links With Logical State = INIT
-----
-I- No bad Links (with logical state = INIT) were found

-----
-I- PM Counters Info
-----
-I- No illegal PM counters values were found

-----
-I- Bad Links Info
-----
-I- No bad link were found

-I- Done. Run time was 8 seconds.
```

## Logging and debugging options

The following information pertains to logging and debugging:

- The Omni-Path Subnet Manager initializes the fabric and facilitates fabric topology management. You can use the `opafmcmd` command for debugging Omni-Path problems.
- The InfiniBand subnet manager is called OpenSM. The `opensm(8)` manpage describes the ranges for the debugging and logging options. When you start a troubleshooting session, HPE recommends that you set the following parameters:
  - `-D 0x7`, which sets a reasonable log verbosity level.
  - `-d 2`, which clears the logs immediately after each log message.

For more information about the OpenSM utility, log into one of the leader nodes and see the `opensm(8)` manpage.

# System monitoring

The following topics explain how to use the system monitoring software on HPE clusters:

- [\*\*Hardware event tracker \(HET\) notifications\*\*](#) on page 253
- [\*\*Ganglia\*\*](#) on page 256
- [\*\*Hardware event logs\*\*](#) on page 258
- [\*\*Heartbeat daemon\*\*](#) on page 259
- [\*\*Nagios\*\*](#) on page 260
- [\*\*Performance Co-Pilot\*\*](#) on page 263

## Hardware event tracker (HET) notifications

The following topics contain information about HET:

- [\*\*About HET\*\*](#) on page 253
- [\*\*Customizing the default HET notification script\*\*](#) on page 254
- [\*\*Using environment variables to create a site-specific HET notification\*\*](#) on page 254
- [\*\*HET example\*\*](#) on page 256

### About HET

The baseboard management controller (BMC) on each cluster system sends SNMP traps to the admin node.

The cluster BMC firmware and the cooling node firmware on a hierarchical systems include threshold values for each component. When a noncritical event occurs, HET logs the event to `/var/log/het/het_trap_processor.log`. If a system condition becomes too low or too high for its threshold, the BMC sends a critical event alert to the default email address of `root@localhost`. The following are examples of critical system events that cause an alert:

- Ambient air temperature higher than recommended
- BMC detects 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

HPE recommends that you complete the procedure in the following topic to configure HET for your site:

[\*\*Customizing the default HET notification script\*\*](#) on page 254

There can be situations in which manual configuration is required. For information about manual configuration, HET defaults, and HET internal processes, see the `het(8)` manpage.

## Customizing the default HET notification script

You can customize the email addresses to which HET sends event information NON-RECOVERABLE events. In addition, you can specify a site-specific email address for less-severe events or for all HET events.

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

1. Log into the admin node as root and open the following file with a text editor:

```
/etc/sysconfig/het
```

2. Search the file for the following string:

```
HET_MAIL_NON_RECOVERABLE_TO
```

3. Change the default recipient, `root`, to be the email address of a person or the email alias of a group.

The person or group you specify must be able to 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_CRITICAL_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.

---

 **CAUTION:** If you configure an email recipient for all HET events, realize that the quantity of email could cause excessive network traffic.

---

## Using environment variables to create a site-specific HET notification

HET logs all events to the following file:

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

If you use Nagios, be aware of the following:

- Nagios monitors the `het_trap_processor.log` file for WARNING and CRITICAL messages.
- Nagios issues an alert for each WARNING or CRITICAL log file entry.

You can set HET environment variables to send the information that resides in the HET log file to an administrator. The following topics explain how to use the HET environment variables:

- [Creating a site-specific HET notification](#) on page 255
- [HET environment variables](#) on page 255

## Creating a site-specific HET notification

You can edit a package-provided sample script to send specific notifications to one or more administrators. Be aware of the following information when you customize this script:

- You can edit the sample file to include the environment variables that you need. The sample script is in the following location:

`/etc/het.action.d/het_user_action.example`

For information about the environment variables that are available, see the following:

[HET environment variables](#) on page 255

- Rename the script to a file name that does not include a period (.) character.

The example script name includes a period character to ensure that the sample file itself does not run. To create a functioning notification, edit and then rename the script. For example, you could rename the script to `het_user_action`.

- The content of the sample script is as follows:

```
#!/bin/sh

# Copyright ...

#=====
#HET - Example Action Program
#=====
# See README in this directory.
# For this script to run on every alert, it needs to be renamed to remove '.':
#   $mv het_user_action.example het_user_action

OUTPUTDIR=/tmp
HET_OUTFILE=het_user_action.out

echo "The following HET(Hardware Environment Tracking) event has been recorded:" >>$OUTPUTDIR/$HET_OUTFILE
# Without some kind of filter, this will run on every alert.
echo "Severity: $HET_ALERTSEVERITY">>$OUTPUTDIR/$HET_OUTFILE
echo "Event Details:">>$OUTPUTDIR/$HET_OUTFILE
printenv | grep HET | sort -t= -k 1 | awk -F= '{printf "\t%-30s %s\n",$1,$2;}'>>$OUTPUTDIR/$HET_OUTFILE
```

## HET environment variables

The HET README file lists all the HET environment variables that you can include in a notification script. This file resides in `/etc/het.action.d/README`.

## HET example

The following is an example of a HET log file that contains critical information:

```
dump    2018-10-23.07.13.21 [het_process_thread:2] # begin -----
dump    2018-10-23.07.13.21 [het_process_thread:2] agentAddr      172.24.0.2
dump    2018-10-23.07.13.21 [het_process_thread:2] het_type       ipmi
dump    2018-10-23.07.13.21 [het_process_thread:2] guid          r1lead
dump    2018-10-23.07.13.21 [het_process_thread:2] sn            X1-----
dump    2018-10-23.07.13.21 [het_process_thread:2] alertSeverity NONE
dump    2018-10-23.07.13.21 [het_process_thread:2] event          uncorrectableECC
dump    2018-10-23.07.13.21 [het_process_thread:2] sensorName     None-memory
dump    2018-10-23.07.13.21 [het_process_thread:2] sensorNumber   0x00
dump    2018-10-23.07.13.21 [het_process_thread:2] sensorTypeName memory
dump    2018-10-23.07.13.21 [het_process_thread:2] eventClassName discrete
dump    2018-10-23.07.13.21 [het_process_thread:2] event1         0x51
dump    2018-10-23.07.13.21 [het_process_thread:2] event2         0xff
dump    2018-10-23.07.13.21 [het_process_thread:2] event3         0x51
dump    2018-10-23.07.13.21 [het_process_thread:2] flap_count     1
dump    2018-10-23.07.13.21 [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 May 2018 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      |
| SENSORMNUMBER  | 0x00             |
| SENSORTRESHOLD | 81               |
| SENSORTYPE     | memory           |
| SENSORVALUE    | 255              |
| SEVERITY       | NON-RECOVERABLE  |
| SN             | X1-----          |
| TYPE           | ipmi             |

## Ganglia

Ganglia is a scalable, distributed monitoring system. It displays web browser-based, real-time (on demand) histograms of system metrics.

Each ICE compute node (blade) is a single monitoring source that sends its statistics to the leader node. After collecting the data, the leader node forwards aggregated rack statistics to the admin node. The leader node also sends its own statistics to the admin node. The admin node is the meta-aggregator for the entire hierarchical cluster system. It collects data from all leader nodes and presents the clusterwide metrics. This model enables HPE to scale out Ganglia to large cluster deployments.

The **Node View** as can aid in system troubleshooting. For every blade in the system, the **Location** field of the **Node View** shows the exact physical location of the blade. This information is useful when trying to locate a blade that is down.

The following topics contain information about Ganglia on HPE cluster systems:

- [\*\*Accessing the Ganglia system monitor\*\*](#) on page 257
- [\*\*Monitoring system metrics\*\*](#) on page 257

Detailed information about the Ganglia monitoring system is available at the following link:

<http://ganglia.info/>.

## Accessing the Ganglia system monitor

To access the Ganglia system monitor from the admin node, specify the following URL:

`http://admin_domain_name/ganglia`

To access the Ganglia system monitor from a leader node, specify the following URL:

`http://admin_domain_name/ganglia/leader_name/`

For example: `http://myadmin/ganglia/r1lead`

## Monitoring system metrics

By default, Ganglia monitors standard operating system metrics such as CPU load and memory usage.

The **Grid Report** view shows an overview of your system. The Grid Report includes the following information:

- The number of CPUs
- The number of hosts (compute nodes) that are up or down
- Compute node information
- Memory usage information

The **Last** pull-down menu allows you to view performance data on an hourly, daily, weekly, or yearly basis.

The **Sorted** pull-down menu allows provides an ascending, descending, or by host view of performance data.

The **Grid** pull-down menu allows you to see performance data for a particular rack or compute node.

The **Get Fresh Data** button allows you to see current data performance.

The following topics explain the metrics that Ganglia gathers:

- [\*\*Default admin node metrics\*\*](#) on page 257
- [\*\*Default leader node metrics\*\*](#) on page 258
- [\*\*Default flat compute node and ICE compute node metrics\*\*](#) on page 258

## Default admin node metrics

By default, the cluster manager configures Ganglia to gather the following categories of metrics for admin nodes:

```
cpu  
disk  
diskstat  
load  
memory  
memory_vm  
network  
process  
procstat  
ssl  
tcp  
tcpext  
udp
```

## Default leader node metrics

By default, the cluster manager configures Ganglia to gather the following categories of metrics for leader nodes:

```
no_group  
cpu  
disk  
diskstat  
load  
memory  
memory_vm  
network  
process  
procstat  
ssl  
tcp  
tcpext  
udp
```

---

**NOTE:** The `no_group` metrics include the temperature-related metrics.

---

## Default flat compute node and ICE compute node metrics

By default, the cluster manager configures Ganglia to gather the following categories of metrics for flat compute nodes and ICE compute nodes:

```
cpu  
disk  
load  
memory  
network  
process
```

## Hardware event logs

All server nodes in a cluster have BMCs. The BMCs provide a broad set of functions as described in the IPMI 2.0 standard. The cluster manager uses the BMCs for the following:

- Remote power management
- Remote system configuration
- Gathering and reporting critical hardware events

Critical hardware events are gathered for the following nodes:

- Leader nodes
- CMCs
- ICE compute nodes

These events are logged in the following locations:

- /var/log/messages through syslog
- /var/log/sel/sel.log

All critical hardware events are summarized under the BMC\_CMC event type. One particular event holds the following useful information:

```
MSG ::= <syslog-prefix> TEMPO:<node> EVENT:<event> APP:<app> Date:<date> VERSION:<version> TEXT <text>
```

The following fields are all type string:

**<node>**

The node name, for example, r1i0n5.

**<event>**

BMC\_CMC

**<app>**

SEL-LOGGER

**<date>**

The date and time of the event.

**<version>**

1.0

**<text>**

The exact copy of the hardware event description from the BMC.

After reading the events from the BMCs, the BMC event logs are cleared on the controller to avoid duplicate events.

## Heartbeat daemon

The availability of each node in a hierarchical cluster is monitored by a lightweight daemon called tempohbc. Each managed compute node, leader node, and ICE compute node runs this daemon. The daemon reports its status to the server that monitors the daemon. The server daemon, which runs on the admin node and leader node, reports if the client is down after approximately 120 seconds. In this event, administrator-derived actions can be triggered, for instance sending an email notification to the system administrator.

The HEARTBEAT event contains the following useful information:

```
MSG ::= <syslog-prefix> TEMPO:<node> EVENT:HEARTBEAT APP:TEMPOHBD Date:<date> VERSION:1.0 TEXT <text>
```

The HEARTBEAT event is created when nodes fail or recover, described by the TEXT field.

The following fields are of type string:

**<node>**

The node name, for example, r1i0n5

**<date>**

The date and time of the event

**<text>**

A description of the event:

'Heartbeat not detected'

'Heartbeat lost'

## Nagios

Nagios is a feature-rich, web-based system monitoring tool for networks and clusters. Among its features are the following:

- Monitoring the health of specified cluster nodes and services
- Notifying a specified audience by email or SMS if a critical event occurs
- Gathering and displaying statistics about specified nodes and services
- Highly customizable

The following topics contain more information about Nagios:

- [Accessing Nagios](#) on page 260
- [Examining the cluster components configured for Nagios monitoring](#) on page 261

## Accessing Nagios

Nagios is installed on admin nodes. On hierarchical clusters, Nagios is also installed on the leader nodes. To monitor the entire cluster, access Nagios on the admin node. To only monitor the nodes subordinate to a leader, access Nagios on that leader node.

To access Nagios, enter one of the following into a browser:

- To access Nagios on the admin node, enter the following URL:

`http://admin_public_domain_name/nagios/`

- To access Nagios on a leader node, enter the following URL:

`http://admin_public_domain_name/leader_name/nagios`

For example: `http://myadmin/r1lead/nagios`

For both the admin node and the leader node, the default username is `nagiosadmin`. For both node types, the default password is `cmdefault`.

The following topics explain how to change the default password:

- [Changing the Nagios password on RHEL](#) on page 261
- [Changing the Nagios password on SLES](#) on page 261

## Changing the Nagios password on RHEL

The following procedure explains how to change the Nagios password on RHEL nodes.

### Procedure

1. Enter the following command:

```
# htpasswd -c /etc/nagios/htpasswd.users nagiosadmin
```

2. At the prompt, supply default password, `cmdefault`.

3. After changing the password, enter the following command to restart HTTP services:

```
# systemctl restart httpd
```

## Changing the Nagios password on SLES

The following procedure explains how to change the Nagios password on SLES nodes.

### Procedure

1. Enter the following command:

```
# htpasswd2 -c /etc/nagios/htpasswd.users nagiosadmin
```

2. At the prompt, supply default password, `cmdefault`.

3. After changing the password, enter the following command to restart Apache services:

```
# systemctl restart apache2
```

## Examining the cluster components configured for Nagios monitoring

By default, the cluster manager has Nagios configured to monitor several cluster system components. To see the list of items that Nagios monitors, open Nagios in a web browser, and click any of the items in the left pane. Observe that the right pane refreshes and displays the aspects of that component that Nagios monitors. For example, if you select the **Services** link for the admin node and the leader nodes, you see entries in the right pane for the current load and the current users.

The following topics explain ways to modify the configuration:

- [Modifying the configuration files](#) on page 261
- [Validating changes and reloading Nagios](#) on page 262

## Modifying the configuration files

Nagios is installed on the admin node. On hierarchical clusters, Nagios also is installed on leader nodes.

The main configuration file, `nagios.cfg`, is located in the `/etc/nagios` directory. This file contains the directives that control how Nagios monitors the system. The directives specify object definition files that target the following:

- Hosts
- Services
- Host groups
- Contacts
- Contact groups
- Commands
- Other entities

Within this directory, you can define all the objects that you want to monitor and how you want to monitor them.

You use the `cfg_file` and/or the `cfg_dir` directives to specify the following object definition files in the main configuration file:

---

#### **Object definition file Content**

---

|                           |                                                                                                                                                                                                                                                                                                                              |
|---------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>hosts.cfg</code>    | The hosts associated with the node where Nagios is installed. You can edit this file to add hosts that are not managed by the HPE Performance Cluster Manager. The <code>update-configs</code> script generates the following Nagios host configuration file:<br><br><code>/opt/clmgr/nagios/etc/objects/cm-hosts.cfg</code> |
| <code>cm-hosts.cfg</code> | The hosts managed by the cluster manager. Each time <code>update-configs</code> is called, the cluster manager updates <code>cm-hosts.cfg</code> .                                                                                                                                                                           |
| <code>services.cfg</code> | The services to be run on the hosts or host groups defined in the <code>services.cfg</code> templates.                                                                                                                                                                                                                       |
| <code>commands.cfg</code> | The commands for the Perl, shell, or Python scripts to be executed by <code>services.cfg</code> .                                                                                                                                                                                                                            |
| <code>contacts.cfg</code> | The contacts or contact groups to be notified by services.                                                                                                                                                                                                                                                                   |

---

There are additional plugins available for use with commands and services in the following directories:

- On RHEL, look in `/usr/lib64/nagios/plugins`.
- On SLES, look in `/usr/lib/nagios/plugins`.

## **Validating changes and reloading Nagios**

After you change one of the object definition files, use the following command to validate the change:

```
# nagios -v /etc/nagios/nagios.cfg
```

After you successfully validate your change, use the following command to reload Nagios:

```
# systemctl reload nagios
```

# Performance Co-Pilot

System metrics are available through the Performance Co-Pilot. The Performance Co-Pilot collection daemon (PMCD) runs on the admin node, leader nodes, and flat compute nodes.

The following topics describe how to use Performance Co-Pilot:

- [Monitoring SDR metrics](#) on page 263
- [Starting the pmgcluster cluster performance monitor \(flat clusters\)](#) on page 264

## Monitoring SDR metrics

The sensor data repository (SDR) metrics are available through Performance Co-Pilot. The SDR provides temperature, voltage, and fan speed information for the following:

- Flat compute nodes
- Leader nodes
- ICE compute nodes
- CMCs

This information is collected from the nodes through their BMC interfaces. As this information is collected out-of-band, it does not affect node performance.

The following metrics are available through the PMCD:

```
admin:~ # pminfo -h rllead sensor
sensor.value.fan
sensor.value.voltage
sensor.value.temperature
```

Each sensor has a separate instance within the domain, with the instance of the form:

```
<nodeName>:<nodeType>:<metricName>
```

```
nodeName ::= Tempo for SGI ICE X node names (rXlead, rXiYc, rXiYnZ)
nodeType ::= "service", "cmc", "blade", "leader"
```

For example, to view voltages for the leader node, enter the following:

```
admin:~ # pminfo -h rllead -f sensor.value.voltage | grep -E '^(\$|^sensor|rllead)'

sensor.value.voltage
inst [0 or "rllead:leader:CPU1_Vcore"] value 1.3
inst [1 or "rllead:leader:CPU2_Vcore"] value 1.3
inst [2 or "rllead:leader:3.3V"] value 3.26
inst [3 or "rllead:leader:5V"] value 4.9
inst [4 or "rllead:leader:12V"] value 11.71
inst [5 or "rllead:leader:-12V"] value -12.3
inst [6 or "rllead:leader:1.5V"] value 1.47
inst [7 or "rllead:leader:5VSB"] value 4.9
inst [8 or "rllead:leader:VBAT"] value 3.31
```

For additional examples, see the following manpages:

- `pmval(1)`. This manpage shows how to retrieve values.
- `pmie(1)`. This manpage shows how to perform trend analyses.

## Starting the `pmgcluster` cluster performance monitor (flat clusters)

You can use the `pmgcluster` tool to monitor a flat cluster. The following procedure enables the `pmgcluster` command.

### Procedure

1. Install the `pcp-sgi` package on the admin node.
2. Enter the following command to create a node list:  
`# cnodes --compute > /etc/nodes`
3. Enter the `pmgcluster` command at the system prompt.

# System maintenance and troubleshooting

The following topics describe system maintenance and troubleshooting:

- [Hardware maintenance procedures](#) on page 265
- [Node replacement process for cold spares](#) on page 270
- [Troubleshooting IRU power up and automatic power down problems \(hierarchical clusters\)](#) on page 274
- [Miscellaneous troubleshooting tools](#) on page 284
- [Retrieving system firmware information](#) on page 289
- [Booting a flat compute node or a leader node on an installed cluster](#) on page 290
- [Overriding installation scripts](#) on page 293

## Hardware maintenance procedures

The following topics describes some common maintenance procedures:

- [Taking one ICE compute node or flat compute node offline for maintenance temporarily](#) on page 265
- [Taking one leader node in a highly available \(HA\) leader node configuration offline for maintenance temporarily](#) on page 266
- [Replacing a failed blade](#) on page 267
- [Replacing a management switch](#) on page 267

### Taking one ICE compute node or flat compute node offline for maintenance temporarily

The following procedure explains how to temporarily take an ICE compute node or a flat compute node offline for maintenance.

#### Procedure

1. Disable the node in the batch scheduler.

See your batch scheduler documentation for this procedure.

2. Power off the node.

For example:

```
# cpower node off r1i0n0
```

3. Mark the node offline.

For example:

```
# cadmin --set-admin-status --node r1i0n0 offline
```

4. Perform maintenance on the blade.

5. Mark the node online, as follows:

For example:

```
# cadmin --set-admin-status --node r1i0n0 online
```

6. Power up the node.

For example:

```
# cpower node on r1i0n0
```

7. Enable the node in the batch scheduler.

See your batch scheduler documentation for this procedure.

## Taking one leader node in a highly available (HA) leader node configuration offline for maintenance temporarily

This topic explains how to shut down one of the two leader nodes. The procedure shuts down the node in an orderly way. The procedure also avoids the unexpected results that can occur when the cluster manager perceives one of the nodes being down as a failure. This procedure works for all nodes, regardless of their status in the cluster as a master node or a slave node. This procedure is best run on two windows.

### Procedure

1. As the root user, log into the leader node that you do not want to take down.

In other words, log into the leader node that you want to remain operational during the maintenance procedure.

2. In one window, enter the following command so you can monitor the leader node transition to the standby node:

```
# crm_mon
```

3. In a second window, enter the following command to set the other node (the target node) to standby status:

```
# crm_standby --update=on --node
```

4. In the first window, monitor the progress of the target node as it transitions to `standby` mode.

Make sure that the node is not listed as `Online`.

5. (Optional) Enter the following command to safely power off the target node at this time:

```
# cpower leader shutdown
```

6. Perform the maintenance activity on the target node.

7. Enter the following command to clear the `standby` status on the node:

```
# crm_standby --update=off --node
```

Monitor the first window to make sure that the status changes from `standby` to `OFFLINE`.

8. (Conditional) Power on the target node.

Complete this step if the node is not powered on.

Enter the following command:

```
# cpower leader on
```

## Replacing a failed blade

On some platforms, beware that some blades include multiple nodes.

---

**NOTE:** See your technical support representative for the physical removal and replacement of ICE compute nodes (blades).

---

The following procedure explains how to permanently replace a failed blade.

### Procedure

1. (Optional) Disable the node in the batch scheduler.

See your batch scheduler documentation for this procedure.

2. Power off the node.

For example:

```
# cpower node off r1i0n0
```

3. Mark the node offline.

For example:

```
# cadmin --set-admin-status --node r1i0n0 offline
```

4. Physically remove and replace the failed blade.

It is not necessary to run `discover-rack` when you replace a blade. The `blademond` daemon performs that task.

5. Power on the node.

For example:

```
# cpower node on r1i0n0
```

6. Set the node to boot the required compute image.

For example:

```
# cimage --set mycomputeimage mykernel r1i0n0
```

For information about this step, see the following:

- Run the `cimage --show-images` command, and observe the output.
- [Using the cimage command to manage ICE compute node images](#) on page 225

7. (Optional) Enable the node in the batch scheduler.

See your batch scheduler documentation for this procedure.

## Replacing a management switch

The following topics explain how to replace a management switch:

- [Backing up the current management switch configuration file](#) on page 268
- [Configuring the new management switch](#) on page 268
- [Management switch replacement example](#) on page 269

## Backing up the current management switch configuration file

The following procedure explains how to back up a management switch configuration file.

### Procedure

1. Save the running configuration file as the startup configuration file.

Use the `switchconfig` command in the following format:

```
switchconfig config --switches hostname --save
```

For *hostname*, specify the switch hostname.

For example:

```
# switchconfig config --switches mgmtsw0 --save
```

2. Save the startup configuration file locally.

Use the `switchconfig` command in the following format:

```
switchconfig config --switches hostname --pull
```

For *hostname*, specify the switch hostname.

The preceding command saves the startup configuration file to the following default location on the admin node:

```
/opt/clmgr/repos/mgmtsw_config_files/hostname/config_file
```

For example, the following command writes the switch configuration file to `/opt/clmgr/repos/mgmtsw_config_files/mgmtsw0/primary.cfg`:

```
# switchconfig config --switches mgmtsw0 --pull
```

3. Proceed to the following:

[\*\*Configuring the new management switch\*\*](#) on page 268

## Configuring the new management switch

The following procedure copies the saved configuration file from the admin node to the new switch.

### Procedure

1. Use the documentation from the switch manufacturer to physically replace the old switch with the new switch.

Make sure that the cabling is identical to the way the old switch cabling was configured.

2. Log into the admin node as the root user.

3. Use the `cadmin` command to update the cluster database with the MAC address of the new switch.

Use the following format:

```
cadmin --set-mac-address --node hostname --eth eth0 mac_address
```

The variables are as follows:

---

| Variable           | Specification                                                                         |
|--------------------|---------------------------------------------------------------------------------------|
| <i>hostname</i>    | The name of the switch that is being replaced.                                        |
| <i>mac_address</i> | The MAC address, on the switch, that the switch can use to obtain an IP address.<br>s |

---

For example:

```
# cadmin --set-mac-address --node mgmtsw0 --eth eth0 02:04:96:98:3c:91
```

#### 4. Use the `switchconfig` command to push the configuration file to the new switch.

The format is as follows:

```
switchconfig config --switches hostname --push --local-file config_file
```

For *config\_file*, specify the file that you want the new switch to load when you boot the switch. To find the name, look in the following directory:

```
/opt/clmgr/repos/mgmtsw_config_files/hostname/startup_config_file
```

For example:

```
# switchconfig config --switches mgmtsw0 --push --local-file primary.cfg
```

#### 5. Use the `switchconfig` command to boot the new management switch.

The format is as follows:

```
# switchconfig reset --switches hostname
```

For example:

```
# switchconfig reset --switches mgmtsw0
```

## Management switch replacement example

The following example shows the process for replacing a management switch. Assume that the switches are as follows:

- The old switch has MAC address 02:04:96:99:88:77.
- The replacement switch has MAC address 02:04:96:98:3c:91.
- The switch hostname is mgmtsw1.
- The cabling for the two switches is identical.
- DHCP is enabled on the replacement switch.

Example:

```
# cadmin --show-mac-address --node mgmtsw1                                # Verify old switch's MAC address
Interface      Ethernet      MAC
mgmtsw1        eth0          02:04:96:99:88:77
# switchconfig config --switches mgmtsw1 --save                            # Save and back up mgmtsw1's config file
# switchconfig config --switches mgmtsw1 --pull                            # Find name of old switch config file
# ls /opt/clmgr/repos/mgmtsw_config_files/mgmtsw1
primary.cfg
# cadmin --set-mac-address --node mgmtsw1 --eth eth0 02:04:96:98:3c:91      # Add new switch's MAC address
# cadmin --show-mac-address --node mgmtsw1                                # Verify new switch's MAC address
Interface      Ethernet      MAC
mgmtsw1        eth0          02:04:96:98:3c:91
# switchconfig config --switches mgmtsw1 --push --local-file primary.cfg
```

```
# switchconfig restart --switches mgmtsw1          # Push config file to new switch
  # Boot the new switch
```

## Node replacement process for cold spares

This topic introduces the process for installing and configuring a spare node. The node can be an admin, leader, or flat compute node. The cold spare can be a shelf spare or a factory-installed cold spare that shipped with your system. The replacement process applies equally to the case where the spare is actually the failed node itself with a motherboard replacement.

The following topics explain the procedures to complete to replace the failed node:

### Procedure

1. [Ensure that the spare is an appropriate replacement](#) on page 270
2. [Identify the failed unit and replace it](#) on page 270

---

**NOTE:** If you are using multiple root slots, repeat the procedures described in this process for each slot.

### Ensure that the spare is an appropriate replacement

A cold spare node is like one of the nodes on your running cluster. The spare sits on a shelf or is a factory preinstalled node. The cold spare is intended to be used in an emergency.

Make sure that HPE supplied your spares. HPE does not support spares not supplied by HPE.

You need the following two spares:

- One spare for the admin node.
- One spare for a leader or flat compute node.

For node replacement, spares for leaders and flat computes are interchangeable.

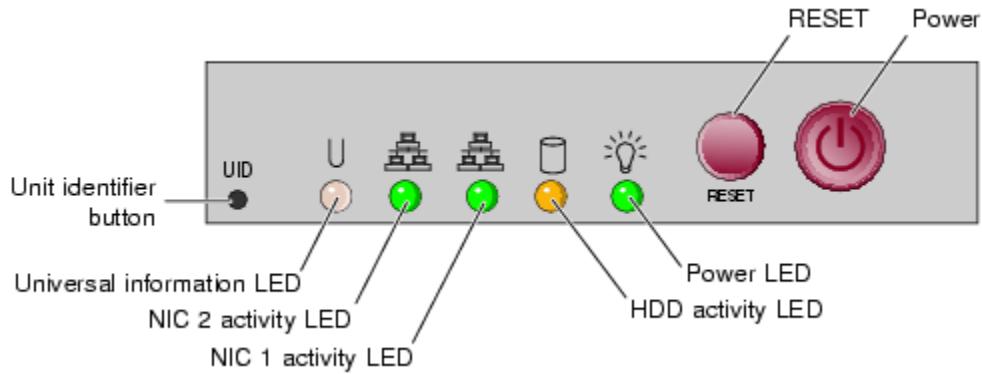
The following are some reasons to have the two types of spares:

- Admin node BIOS settings are different from BIOS settings for leader nodes and flat compute nodes.  
For example, an admin node does not PXE boot by default. However, a leader must PXE boot each time. In addition, the boot order is different for each node type. Attempts to use the `discover` command to configure the node into a cluster will fail.
- The BMCs of leader nodes and flat compute nodes are configured to use DHCP by default. An admin node cannot be configured this way. Attempts to use the `discover` command to configure the node into a cluster will fail.
- The factory programs a cluster serial number into the admin node. The admin node spare must have correct cluster serial number programmed into it. In this situation, parts of the cluster and support software might not work correctly.

### Identify the failed unit and replace it

The procedure in this topic explains how to handle the physical hardware.

If the unit has failed, the front panel lights on the server can indicate the failure. In addition, the front panel includes other failure information. The following figure shows the front panel:



**Figure 54: Server front panel controls and indicator LEDs**

The universal information LED (left side of the panel) shows two types of failure that can bring the server down. This multicolor LED flashes red quickly to indicate a fan failure. The LED flashes red slowly for a power failure. A continuous solid red LED indicates that a CPU is overheating.

If the unit power supply has failed or been disconnected, the power LED (far right) is dark. Check both ends of the power cable for a firm connection prior to switching over to the cold spare.

The admin node stores the systemwide serial number. Order a new admin node shelf spares from the HPE factory. The HPE factory configures admin node shelf spares with the proper serial number.

The following procedure explains how to replace a node that has failed.

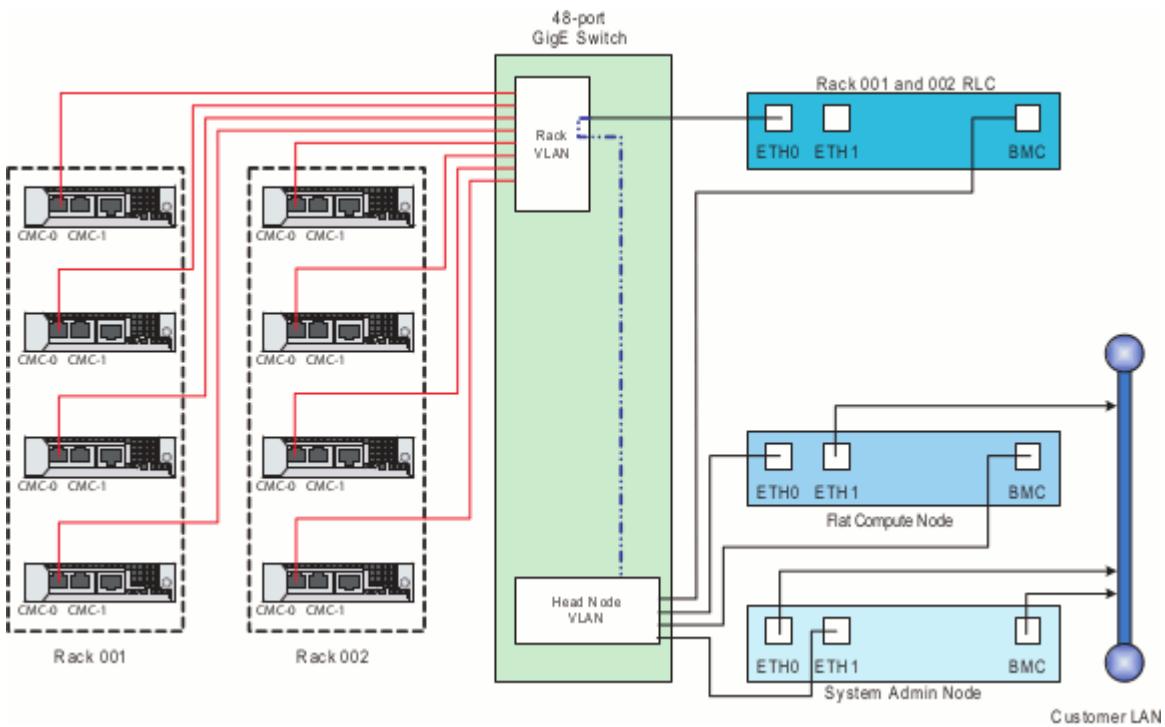
### Prerequisites

Obtain a VGA screen and a keyboard. You access the LSI BIOS tool to import the root volumes. You cannot access the LSI BIOS from an IPMI serial console session because the LSI BIOS tool requires the use of **Alt** characters. Such characters often do not transfer through the serial console properly.

### Procedure

1. If possible, power down the failed node.
2. Disconnect both power cables.

The following figure shows server connection locations:



**Figure 55: Simple CMC LAN (VLAN) cable examples**

3. Remove the two system disks from the failed node and set them aside for later reinstallation.
4. Unplug the Ethernet cable used for system management.

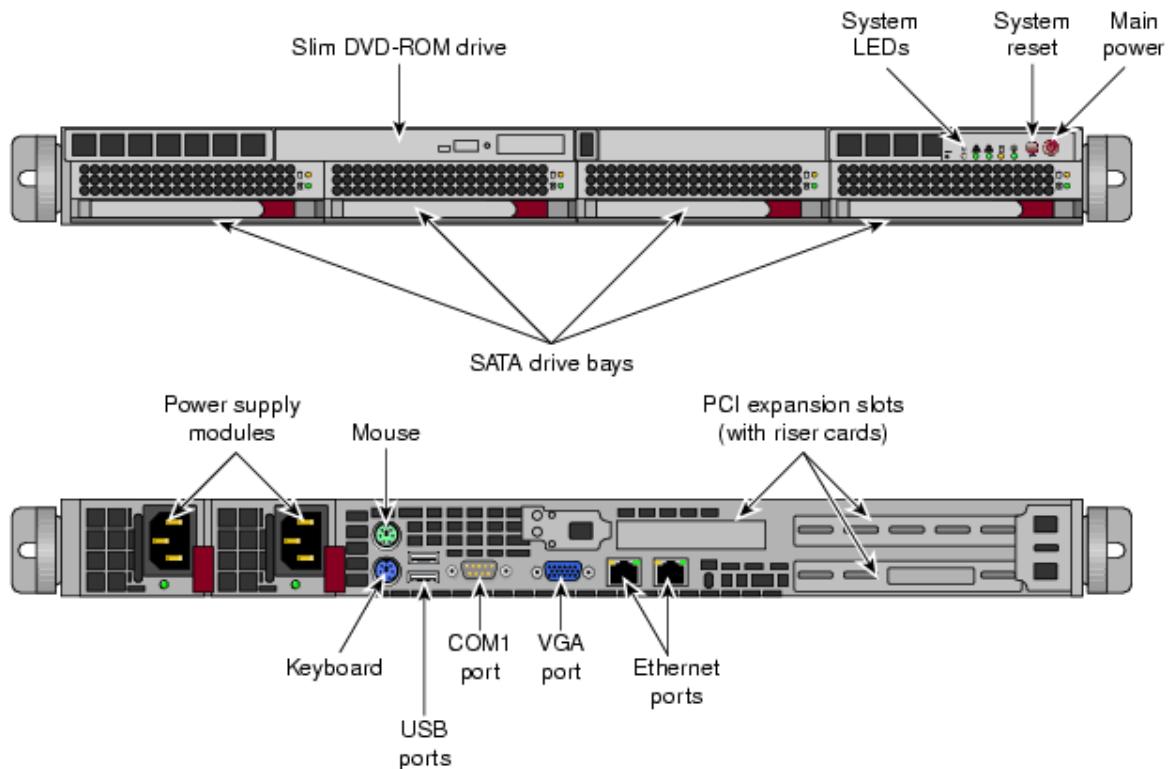
To avoid confusing them, note the plug number and label the cables. It is important that they stay in the same jacks in the new node. This connection is vital to proper system management and communication.

---

**NOTE:** The Ethernet cable must be connected to the same plug on the cold spare unit.

---

5. Remove the keyboard, mouse, and video cables.  
Complete this step if the unit has a system console attached.
6. Remove the system from the rack.
7. From the existing server, transfer any PCI cards and disk drives to the cold spare.  
Use the following figure for guidance:



**Figure 56: Server front features and rear connector locations**

8. Install the shelf spare system into the rack.
9. Install the system disks you set aside earlier in this procedure.  
These disks are from the failed system.
10. Connect the Ethernet cables in the same way they were connected to the replaced node.
11. Connect AC power.
12. Connect a keyboard, VGA monitor, and mouse.
13. Power up the replaced node and restart the cluster.
14. (Conditional) Update the cluster manager database.

Complete this step if the failed unit is a leader node or a flat compute node.

Update the following in the cluster database:

- The MAC address of the spare
- The MAC address of the BMC in the spare

When you update the preceding address information in the database, you ensure that the cold spare can boot and function properly. If necessary, use the BIOS to retrieve the new MAC addresses.

From the admin node, use the `cadmin` command to query and set MAC addresses in the database. The following table shows the command parameters that you can use:

| Parameter                              | Effect                                                       |
|----------------------------------------|--------------------------------------------------------------|
| show-mac-address                       | Displays the MAC address of the node you specify.            |
| set-mac-address <i>mac-address</i>     | Sets the MAC address of the node you specify.                |
| show-bmc-mac-address                   | Displays the MAC address of the BMC of the node you specify. |
| set-bmc-mac-address <i>mac-address</i> | Sets the MAC address of the BMC of the node you specify.     |
| node <i>nodename</i>                   | Specifies the target node.                                   |

Example 1. The following example displays the MAC address of flat compute node `service0`:

```
# cadmin --show-mac-address --node service0
Interface          Ethernet      MAC
service0           eth0         00:25:90:03:4e:02
```

Example 2. The following example sets the MAC address of `service0`:

```
# cadmin --set-mac-address 00:25:90:04:4e:01 --node service0
```

Example 3. The following example shows the MAC address of the BMC on `service0`:

```
# cadmin --show-bmc-mac-address --node service0
Interface          Ethernet      MAC
service0-bmc       bmc0         00:25:90:03:53:6b
```

Example 4. The following example sets the MAC address of the BMC on `service0`:

```
# cadmin --set-bmc-mac-address 00:25:90:03:51:1d --node service0
```

## Troubleshooting IRU power up and automatic power down problems (hierarchical clusters)

The following topics describe how to troubleshoot power up and power down actions on a hierarchical cluster system:

- [About the power on process \(hierarchical clusters\)](#) on page 275
- [CMC monitoring \(hierarchical clusters\)](#) on page 275
- [Power cycling the IRUs \(hierarchical clusters\)](#) on page 275
- [Power supplies and the watchdog timer \(hierarchical clusters\)](#) on page 276
- [Interpreting the power supply LEDs \(hierarchical clusters\)](#) on page 277
- [Troubleshooting the devices on the CAN bus interface \(hierarchical clusters only\)](#) on page 277
- [Flashing the firmware on a power shelf or fan controller \(hierarchical clusters\)](#) on page 278
- [Troubleshooting a missing power shelf \(hierarchical clusters\)](#) on page 279
- [Power consumption log files \(hierarchical clusters\)](#) on page 282

- [Retrieving information about the power supplies \(hierarchical clusters\)](#) on page 282
- [Retrieving information about the PMBus registers \(hierarchical clusters\)](#) on page 284

## About the power on process (hierarchical clusters)

When you issue a power on, the first chassis management controller (CMC) in a power domain performs the power on. When you issue a power off, the last CMC in the power domain performs the power off.

The power on and the power off processes occur in phases. When you use the `cpower` command to power up and power down, the command handles the process for you.

When you enter the `cpower node on "r*i*n"` command on an admin node to power up the ICE compute nodes, the following occurs:

1. Each CMC turns on the power supplies.

At this point the BMCs on the compute blades have power, are booted, and are running.

2. The CMC enables the fans and waits until it determines that air is moving through the IRU.
3. The CMC sends an IPMI command to the BMCs. This command tells the BMCs to enable power to the compute blades.
4. The BMCs enable power to the compute blades.

## CMC monitoring (hierarchical clusters)

During typical operation, the CMC monitors several aspects of the power supply.

Some hierarchical cluster systems have M-Racks. These systems use external cooling. On these systems, the CMC verifies the following:

- That communication between the CMC and its associated CRC and CDU is open. The CMC monitors the CRC and CDU for error conditions and, if needed, can power off the IRU. The rack number of the CMC determines the CRC and CDU that the CMC monitors.
- That the correct number of power shelves can be detected.

Some hierarchical cluster systems have D-Racks. On these systems, the CMC verifies the following:

- That a certain number of fans are present and spinning. Environmental software on the CMC controls the fan speed and reports failures.
- That the correct number of power shelves can be detected.

## Power cycling the IRUs (hierarchical clusters)

The chassis management controllers (CMCs) enable power to the IRUs. If a hierarchical cluster loses power abruptly, power cycle the IRUs. If power supplies are turned off, the LEDs on the power supplies flash green. If one of the power supplies has a fault, the LED is solid amber. Depending on how the software in the system detected the power off, log entries might provide more information. For information about the log entries, see the following:

[Power consumption log files \(hierarchical clusters\)](#) on page 282

In most cases, if you can power cycle the IRUs, the CMCs can restore power.

The following procedure explains how to power cycle the IRUs.

## Procedure

1. Log into the admin node as the root user.
2. (Conditional) Use the `cnodes` command to retrieve the list of leader nodes and CMCs.

Perform this step if you are unsure of the system ID for the affected leader node and CMC.

Enter the following commands:

```
# cnodes --leader
r1lead
r2lead
# cnodes --cmc
r1i0c
r2i1c
```

The preceding commands show the IDs for the leader nodes and CMCs on a two-rack system.

3. Use the `cpower` command to retrieve information about the CMCS in the rack.

For example:

```
# cpower iru status "rli*"
xxxxx
xxxxx
.
.
.
r1i0c: power is On
r2i1c: power is On
```

4. Power off the CMCS in the rack.

For example:

```
# cpower iru off "rli*"
```

5. Power on the CMCS in the rack.

For example:

```
# cpower iru on "rli*"
```

## Power supplies and the watchdog timer (hierarchical clusters)

If all the CMCS in the power domain detect a fault condition, the following occur:

- The watchdog timer expires
- The system powers off all the power supplies

When the system is operating as expected, the CMCS detect no faults. Under these conditions, the CMCS send a watchdog reset every 10 seconds.

The power shelves must receive a watchdog reset once every 45 seconds from each CMC in the power domain. If the watchdog timer expires, the power shelf controller disables the power supplies on that shelf and sets the `WDOG` status bit.

The following conditions can prevent the CMC from sending the watchdog reset to the power shelves:

- The CMC cannot confirm that a minimum number of fans are spinning. This condition pertains to hierarchical clusters with D-Racks.
- The CMC cannot communicate with the external CRC and/or CDU. The CMC detects a fault condition reported from the CRC and/or the CDU. These conditions pertain to hierarchical clusters with M-Racks.

The output from the CMC `pfctl status` command shows the status of the WDOG status bit. You can enter the `pfctl status` from any CMC in the power domain. The command reports power shelf and supply status and reports fan or CRC/CDU status, depending on rack type.

## Interpreting the power supply LEDs (hierarchical clusters)

The following table explains how to read the status indicators on the power supply LEDs:

| If the light is ... | Meaning                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|---------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Solid green         | Power supply is on and OK.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| Blinking green      | <p>Power supply has AC, but it is not on.</p> <p>If all the power supply LEDs are blinking green, the power was turned off. This situation could result from one of the following:</p> <ul style="list-style-type: none"> <li>• The watchdog timer firing.</li> <li>• A power down issued by all CMCS because of a cooling problem. The cooling problem could be due to one of the following: <ul style="list-style-type: none"> <li>◦ The fan controller on a D-Rack system</li> <li>◦ The CRC/CDU unit on an M-Rack system</li> </ul> </li> </ul> |
| Solid amber         | The power supply has failed.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| Blinking amber      | <p>This signal is a power supply warning. The supply is still operating.</p> <p>There is no AC to the power supply, but the power supply is plugged into the system.</p> <p>There is no AC input (under voltage).</p>                                                                                                                                                                                                                                                                                                                               |

## Troubleshooting the devices on the CAN bus interface (hierarchical clusters only)

The CAN bus is the interface that connects all the CMCS, power shelves, and D-Rack fan controllers. You can use the `pfctl ping` command to retrieve the status of each device and then take corrective action. To use this command, log into the CMC and enter the command at the system prompt.

If the `pfctl ping` command reports missing power shelves, see the following:

### Troubleshooting a missing power shelf (hierarchical clusters) on page 279

Example 1. The following output was obtained for an M-Rack, with two IRUs in a power domain.

```
> pfctl ping
PWR-UPPER-CMC1: r1i5c
PWR-UPPER-CMC0: r1i1c
```

```
PWR_SHELF3:      -
PWR_SHELF2:      PRESENT
PWR_SHELF1:      PRESENT
PWR_SHELF0:      PRESENT
PWR-LOWER-CMC1: r1i4c
PWR-LOWER-CMC0: r1i0c
```

#### EXTERNAL FANS

Example 2. The following output was obtained on an M-Rack, with one IRU in a power domain and twin node blades.

```
> pfctl ping
PWR-UPPER-CMC1: -
PWR-UPPER-CMC0: r1i4c
PWR_SHELF3:      -
PWR_SHELF2:      PRESENT
PWR_SHELF1:      PRESENT
PWR_SHELF0:      PRESENT
PWR-LOWER-CMC1: -
PWR-LOWER-CMC0: r1i0c
EXTERNAL FANS
```

Example 3. The output in this example is from a D-Rack. The fan controller hosts two programmable systems on a chip (PSOC) units. The fan controller controls 12 fans. The following output shows the fan controllers that appear in the FAN-CONTROL lines as PRESENT. This output is typical for a system that is operating properly.

```
> pfctl ping
PWR-UPPER-CMC1: -
PWR-UPPER-CMC0: r1i1c
PWR_SHELF3:      -
PWR_SHELF2:      -
PWR_SHELF1:      PRESENT
PWR_SHELF0:      PRESENT
PWR-LOWER-CMC1: -
PWR-LOWER-CMC0: r1i0c

FAN-UPPER-CMC1: -
FAN-UPPER-CMC0: r1i1c
FAN-CONTROL1     PRESENT
FAN-CONTROL0     PRESENT
FAN-LOWER-CMC1: -
FAN-LOWER-CMC0: r1i0c
```

## Flashing the firmware on a power shelf or fan controller (hierarchical clusters)

In rare situations, the power shelf firmware or fan controller firmware can become corrupted. In this situation, the power shelf or the fan controller becomes broken or remains perpetually in bootloader mode. If in bootloader mode, the fan controller firmware can respond to the firmware flashing utility.

The following procedure explains how to flash the firmware.

### Procedure

1. Log in to the power shelf or the fan controller.

For a power shelf, log into the lowest CMC in the power domain.

For information about how to log into a CMC, see the following:

**Power cycling the IRUs (hierarchical clusters)** on page 275

2. Enter the following command to change to the directory that contains the firmware images:

```
> cd /usr/local/firmware/psoc
```

3. Use the following command to flash the firmware:

```
flashcan -f image -p target -r
```

The variables are as follows:

| Variable      | Specification                                                                                                                                                                                                                                                                                                                               |
|---------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>image</i>  | The name of one of the firmware images from the following directory:<br><code>/usr/local/firmware/psoc</code>                                                                                                                                                                                                                               |
| <i>target</i> | One of the following: <ul style="list-style-type: none"><li>• 0, which specifies power shelf 0.</li><li>• 1, which specifies power shelf 1.</li><li>• 2, which specifies power shelf 2.</li><li>• 3, which specifies power shelf 3.</li><li>• 4, which specifies fan controller 0.</li><li>• 5, which specifies fan controller 1.</li></ul> |

## Troubleshooting a missing power shelf (hierarchical clusters)

It is possible for a power shelf to be physically present but not appear in `pfctl ping` command output. In this case, use the information in the following topics to troubleshoot:

- **Booting a power shelf manually (hierarchical clusters)** on page 279
- **Fixing problems related to a newly installed power shelf (hierarchical clusters)** on page 280

### Booting a power shelf manually (hierarchical clusters)

Occasionally, when the AC power breakers are enabled, the power shelf controller or the fan controller might not boot properly. The power shelf is said to be **wedged** in this situation. In this case, complete the following procedure to power cycle the AC power to all the power supplies in the power domain or cooling domain.

#### Procedure

1. Power cycle the system again.

Use the following procedure:

**Power cycling the IRUs (hierarchical clusters)** on page 275

2. Manually flip the power breakers on the power distribution unit (PDU) at the top of the rack.

3. Enter the following command from the CMC:

```
> pfctl ping
```

4. Examine the output.

The output shows the correct number of power shelves as present. The following examples show correct output for their specific systems.

Example 1. The following output was obtained for the D-Rack of a hierarchical cluster:

```
> pfctl ping
PWR-UPPER-CMC1: -
PWR-UPPER-CMC0: r1i1c
PWR_SHELF3: -
PWR_SHELF2: -
PWR_SHELF1: PRESENT
PWR_SHELF0: PRESENT
PWR-LOWER-CMC1: -
PWR-LOWER-CMC0: r1i0c

FAN-UPPER-CMC1: -
FAN-UPPER-CMC0: r1i1c
FAN-CONTROL1 PRESENT
FAN-CONTROL0 PRESENT
FAN-LOWER-CMC1: -
FAN-LOWER-CMC0: r1i0c
```

Example 2. The following output was obtained for the M-Rack of a hierarchical cluster, with one IRU in a power domain and single-node blades:

```
> pfctl ping
PWR-UPPER-CMC1: -
PWR-UPPER-CMC0: -
PWR_SHELF3: -
PWR_SHELF2: PRESENT
PWR_SHELF1: PRESENT
PWR_SHELF0: PRESENT
PWR-LOWER-CMC1: -
PWR-LOWER-CMC0: r1i0c
```

EXTERNAL FANS

## Fixing problems related to a newly installed power shelf (hierarchical clusters)

If you recently added or replaced a power shelf, the following problems might exist:

- The firmware on the new power shelf was not flashed
- A problem might exist with the CAN bus connection

The following procedure explains how to troubleshoot a new power shelf that is not integrated properly.

### Procedure

1. Log into the CMC.

For systems with M-Racks, log into the lower CMC in the IRU.

For systems with D-Racks, log into the lower CMC in the pair.

For information about how to log into the CMC, see the following:

**Power cycling the IRUs (hierarchical clusters)** on page 275

2. Enter the following command to change to the directory that contains the firmware images:

```
> cd /usr/local/firmware/psoc
```

3. Use the following command to flash the firmware:

```
flashcan -f image -p controller -r
```

These variables are as follows:

| Variable          | Specification                                                                                                                                                                                                                                                                                                                               |
|-------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>image</i>      | The name of one of the firmware images from the following directory:<br><code>/usr/local/firmware/psoc</code>                                                                                                                                                                                                                               |
| <i>controller</i> | One of the following: <ul style="list-style-type: none"><li>• 0, which specifies power shelf 0.</li><li>• 1, which specifies power shelf 1.</li><li>• 2, which specifies power shelf 2.</li><li>• 3, which specifies power shelf 3.</li><li>• 4, which specifies fan controller 0.</li><li>• 5, which specifies fan controller 1.</li></ul> |

4. Enter the `pfctl ping` command to retrieve the status of the power shelf.

If the status returns `PRESENT` for the problem power shelf, you are finished.

If the command does not return `PRESENT`, continue with this procedure to troubleshoot other causes of the problem.

5. Perform one or more of the following remedies:

- Reseat the power shelf.
- Visually inspect the connector on the shelf and make sure that it is correct.

Inspecting the blind connector at the rear of the power shelf slot can be difficult to do.

- Visually inspect the LED lights.

If there were power supplies in the shelf with the fail LED light, the failed supply might have damaged the power shelf.

If a power supply turns on immediately when the AC power is applied, the shelf itself might be damaged.

If the following are both true, the power shelf might be bad:

- If the power supplies in all the other shelves are off (flashing green)
- If the power supplies in the missing shelf turn solid green as soon as the breaker is enabled

Find the sticker on the power shelf, and see if it is discolored.

- Inspect the CAN bus cable in the back on the rack.

## Power consumption log files (hierarchical clusters)

On a CMC, the following log files contain information that can help you troubleshoot a power problem:

- The `/tmp/pfctld.log` file contains entries from the power and fan control daemon, `pfctld`. When the `pfctld` daemon powers down an IRU, it records a log entry in the log file. The entry includes the reason for the power down.
- The `/tmp/eric.log` file contains output from an environmental software monitoring application, called ERIC. ERIC runs on the CMC. The actions of ERIC actions are written to this log file. ERIC monitors blade temperatures and adjusts fans speeds appropriately. ERIC also monitors the CMC inlet air temperature and powers down the IRU when appropriate. That is, ERIC powers down the blades associated with that CMC.

If your hierarchical cluster has D-Racks, and the following conditions are all present, the problem might be related to the CMC air inlet temperature:

- The blades of only one IRU are powered down
- The blades in the other IRU are still on
- Power supply LEDs are solid green

In an M-Rack configuration, there could be a problem with the CMC air inlet temperature. In this case, ERIC could power off only the upper or lower board in the blade.

## Retrieving information about the power supplies (hierarchical clusters)

After you log into the CMC, you can use the `pmbus_drack` and `pmbus_mrack` scripts to retrieve information about the power supplies. These scripts dump some of the PMBus data that is available.

The following example shows output from the `pmbus_mrack` script:

```
> pmbus_mrack
Shelf0 PS0 Vout: 12          Iout: 3.5        Temp: 29        Status: 0x0000
Shelf0 PS1 Vout: 11.6875     Iout: 0          Temp: 29.5      Status: 0x0000
Shelf0 PS2 Vout: 12.0312     Iout: 1          Temp: 28.5      Status: 0x0000
Shelf1 PS0 Vout: 11.625      Iout: 0          Temp: 29        Status: 0x0000
Shelf1 PS1 Vout: 11.6562     Iout: 0          Temp: 29.5      Status: 0x0000
Shelf1 PS2 Vout: 11.6875     Iout: 0          Temp: 28.5      Status: 0x0000
Shelf2 PS0 Vout: 11.6875     Iout: 0          Temp: 28.5      Status: 0x0000
Shelf2 PS1 Vout: 11.625      Iout: 0          Temp: 28        Status: 0x0000
Shelf2 PS2 Vout: 11.6562     Iout: 0          Temp: 29        Status: 0x0000
Shelf3 PS0 Vout:             Iout:             Temp:           Status: not available
Shelf3 PS1 Vout:             Iout:             Temp:           Status: not available
Shelf3 PS2 Vout:             Iout:             Temp:           Status: not available
```

The output shows the output voltage ( $V_{out}$ ), the output current ( $I_{out}$ ), and temperature from each of the power supplies.

You can use the  $I_{out}$  values to determine whether power supply load sharing is working correctly within the power domain. All power supply readings ideally report within +/- 10% of the average. The status bits record warnings and faults when they occur. All bits are decoded if present.

The following status messages can appear in the `pmbus_mrack` output:

**VOUT**

Output voltage warning or fault.

**IOUT**

Output current warning or fault.

**INPUT**

Input fault.

**MFR**

Manufacturer fault. Related to the 3.3v auxiliary supply used to power the power shelf controllers, fan shelf controllers, and the CMCS.

**PWRGOOD**

Power output is good (active low).

**FANS**

Internal fan failure.

**OTHER**

Another warning or fault not indicated by other status flags.

**UNKNOWN**

An internal power supply controller condition was detected.

**OFF**

Power supply is off.

**VOUT\_OV**

Output voltage over limit.

**IOUT\_OC**

Output current over limit.

**VIN\_UC**

Input voltage under limit.

**TEMP**

Temperature warning or fault.

**CML**

Communication error. Can be ignored.

**NOTA**

None of the above.

Power supplies shut down on any fault condition and remain off unless the fault is a temperature fault. After the power supply has cooled, it re-enables itself. Generally, if you cycle the AC power to the faulted power supply, it resets all status flags. Hard failures reoccur. If the system is under heavy load and a

power supply fails, the other supplies pick up the load. If yet another supply fails, this failure can cause an overcurrent across all supplies. An overcurrent can power down all compute blades in the power domain.

## **Retrieving information about the PMBus registers (hierarchical clusters)**

After you log into the CMC, you can use the `pfctl pmbus dump` command to retrieve information about the PMBus registers. This command queries all power supplies in the power domain. In the command output, look for nonzero readings to locate possible problems.

The following example shows output from the `pfctl pmbus dump` command:

```
> pfctl pmbus dump
PWR s0s0          VIN:    213.00
PWR s0s0          IIN:     0.78
PWR s0s0          VOUT:   11.97
PWR s0s0          IOUT:   9.00
PWR s0s0          3.3 VOUT: 3.34
PWR s0s0          3.3 IOUT: 1.34
PWR s0s0          TEMP:   23.00
PWR s0s0          FAN1 RPM: 6320
PWR s0s0          FAN2 RPM: 6320
PWR s0s0          STATUS_BYTE: 00
PWR s0s0          STATUS_WORD: 0000
PWR s0s0          STATUS_VOUT: 00
PWR s0s0          STATUS_IOUT: 00
PWR s0s0          STATUS_INPUT: 00
PWR s0s0          STATUS_TEMPERATURE: 00
PWR s0s0          STATUS_CML: 00
PWR s0s0          STATUS_3V3: 00
PWR s0s0          STATUS_FANS_1_2: 00
PWR s0s0          SMB ALERT : 00 NO
PWR s0s0 SOFTWARE REVISION : pri 167 app 169 boot 2
PWR s0s0          PMBUS: I I II I
PWR s0s0          ID: DELTA
PWR s0s0          MODEL: AHF-2DC-2837W-12V-240V
PWR s0s0          REVISION: 1 6 167 169
PWR s0s0          LOCATION: DES
PWR s0s0          DATE: 23/10 (13:58:00 06/08/10)
PWR s0s0          SERIAL: A000379
```

## **Miscellaneous troubleshooting tools**

The following topics describe some troubleshooting tools:

- [cm\\_info\\_gather command](#) on page 284
- [cminfo command](#) on page 285
- [kdump utility](#) on page 285

### **cm\_info\_gather command**

The `cm_info_gather` command collects system data that you can use to troubleshoot problems. The command collects information about the following:

- dmidecode output, system logs, Dynamic Host Configuration Protocol (DHCP), NFS.
- MySQL/MariaDB cluster database.
- Network service configuration files. For example, Ganglia, DHCP, domain name service (DNS) configuration files.
- Installed system images.
- Log files in /var/log/messages.
- Log files in /var/log/dhcp.
- Chassis management controller (CMC) slot table information for each rack.
- Basic input-output system (BIOS), Baseboard Management Controller (BMC), CMC, and InfiniBand fabric software versions from all ICE compute nodes.

## cminfo command

Many cluster scripts use the `cminfo` command internally. In a troubleshooting situation, you can use the command to gather information about your system.

Example 1. To display the BMC IP address of a leader node, enter the following command:

```
r1lead:~ # cminfo --bmc_base_ip
192.168.160.0
```

Example 2. To display the leader node DNS domain, enter the following command:

```
r1lead:~ # cminfo --dns_domain
ice.domain_name.mycompany.com
```

Example 3. To see the IP address of the ib1 InfiniBand fabric, enter the following command:

```
r1lead:~ # cminfo --ib_1_base_ip
10.149.0.0
```

## kdump utility

You can download the kernel RPMs for use with the crash package from either RHEL or SLES. For RHEL, you can download the `debuginfo` package from Red Hat Network (RHN) and incorporate the package into your node images and running nodes. On SLES, the `kdump` facility is enabled and working by default. For information about how to add packages to your RPM lists and your software images, see the following:

### [Using a custom repository for site packages](#) on page 219

The `kdump` utility is a `kexec`-based crash dumping mechanism for the Linux operating system. By default, the `kdump` crash dump capability is enabled on hierarchical systems after installation.

The following topics provide more information about `kdump`:

- [Obtaining a traceback or system dump](#) on page 286
- [Retrieving the current kdump memory allocation setting](#) on page 287
- [Disabling kdump](#) on page 287
- [Setting a site-specific crashkernel value](#) on page 288
- [Resetting the crashkernel value to the system default](#) on page 288

---

**NOTE:** For a simpler mechanism to trigger a crash dump, consider the `nodetrace` tool. For information about `nodetrace`, enter the following on the admin node:

```
# nodetrace -h
```

---

## Obtaining a traceback or system dump

You can obtain a system dump from an ICE compute node, a leader node, or a flat compute node.

For an admin node, system dump information resides on the node in the following locations:

- Traceback information is in the following file:

```
/net/r1lead/var/log/consoles
```

- System dump information is in the following directory:

```
/net/leader_node/var/crash/sgi_kdump/IP-date
```

For example:

```
root@r1lead ~]# cd /var/crash/sgi_kdump/10.159.0.6-2018-04-20-14\:54\:20/
[root@r1lead 10.159.0.6-2018-04-20-14:54:20]# ls -l
total 128488
-rw-r--r-- 1 sgi_kdump sgi_kdump      69408 Apr 20 14:54 vmcore-dmesg.txt
-rw-r--r-- 1 sgi_kdump sgi_kdump 131499356 Apr 20 14:54 vmcore.flat
```

For an ICE compute node, leader node, or flat compute node, access dump information as follows:

1. Log into the admin node.
2. Bring up a console to the node in question.
3. Enter the following to obtain the crash dump for the selected node:

```
^e c l 1 8
^e c l 1 t      #traceback
^e c l 1 c      #dump
```

For example, to obtain information from a leader node, enter the following:

```
console r1i0n0
^e c l 1 8
^e c l 1 t      #traceback
^e c l 1 c      #dump
```

---

**NOTE:** This example shows the letter “c”, a lowercase L “l”, and the number one “1” in all three lines.

## kdump examples

The `kdump` directory listings include the node hostname and the node IP address.

Example 1. On RHEL platforms, the `kdump` crash directories are located in `/var/crash/sgi_kdump`. The following example shows a listing of crash files from diskless flat cluster nodes:

```
admin_node: # cd /var/crash/sgi_kdump
admin_node:/var/crash/sgi_kdump # ls -la
lrwxrwxrwx  1 sgi_kdump sgi_kdump  30 Mar  1 17:24 service0-2017-03-01-17:19:24 -> 172.23.0.4-2017-03-01-17:19:24
lrwxrwxrwx  1 sgi_kdump sgi_kdump  30 Mar  1 17:31 service0-2017-03-01-17:26:35 -> 172.23.0.4-2017-03-01-17:26:35
```

```
lrwxrwxrwx 1 sgi_kdump sgi_kdump 30 Mar 1 19:07 service0-2017-03-01-19:07:21 -> 172.23.0.4-2017-03-01-19:07:21
lrwxrwxrwx 1 sgi_kdump sgi_kdump 30 Mar 1 19:49 service0-2017-03-01-19:48:35 -> 172.23.0.4-2017-03-01-19:48:35
```

When you change to the `service0-2017-03-01-17:19:24` directory and list the files again, the system displays the following files:

- `vmcore-dmesg.txt`
- `vmcore.flat`

Example 2. The following example is from a SLES cluster. The example shows the ICE compute node crash files. These files are on the leader node.

```
r1lead:/var/crash/sgi_kdump # ls -la
drwxr-xr-x 3 sgi_kdump users 30 Mar 28 15:00 10.159.0.2
lrwxrwxrwx 1 sgi_kdump users 10 Mar 28 15:00 r1i0n0 -> 10.159.0.2
r1lead:/var/crash/sgi_kdump # cd r1i0n0
r1lead:/var/crash/sgi_kdump/r1i0n0 # ls -F
2017-03-28-22:00/
r1lead:/var/crash/sgi_kdump/r1i0n0 # cd 2017-03-28-22:00
r1lead:/var/crash/sgi_kdump/r1i0n0/2017-03-28-22:00 # ls
dmesg.txt makedumpfile-R.pl README.txt rearrange.sh vmcore # CRASH FILES ARE HERE
```

## Retrieving the current `kdump` memory allocation setting

The following procedure explains how to retrieve the current `kdump` setting for a specific system image.

### Procedure

1. Log into the admin node as the root user.
2. Use the `cadmin` command in the following format to retrieve the current `kdump` memory allocation:

```
cadmin --show-crashkernel --image image_name
```

For *image\_name*, specify the name of one of the ICE compute node, leader node, or flat compute node operating system images.

## Disabling `kdump`

When `kdump` is enabled, the system reserves some memory for crash dumps. To make this memory available to user programs, you can disable the `kdump` facility. You can also reduce the size of the memory used for the `kdump` facility.

The following procedure explains how to disable `kdump`.

### Procedure

1. Log into the admin node as the root user.
2. Use the `cadmin` command in the following format to disable the `kdump` facility:

```
cadmin --set-crashkernel --image image_name ""
```

For *image\_name*, specify the name of one of the ICE compute node, leader node, or flat compute node operating system images.

Enter two quotation mark characters at the end of the command to represent the empty string.

For example:

```
# cadmin --set-crashkernel --image ice-sles11sp3 ""
```

3. Push the changes to the desired nodes.

For information about how to push changes, see the following:

[Pushing images from the admin node to the targeted nodes](#) on page 217

## Setting a site-specific `crashkernel` value

The following procedure explains how to specify the amount of memory you want to devote to kdump.

### Procedure

1. Log into the admin node as the root user.
2. Use the `cadmin` command in the following format to specify the amount of memory to use for the `kdump` facility:

```
cadmin --set-crashkernel --image image_name "mem_size"
```

The variables are as follows:

| Variable          | Specification                                                                                       |
|-------------------|-----------------------------------------------------------------------------------------------------|
| <i>image_name</i> | The name of one of the ICE compute node, leader node, or flat compute node operating system images. |
| <i>mem_size</i>   | The amount of memory to allocate to kdump.                                                          |

For example:

```
# cadmin --set-crashkernel --image sles11sp3 "512M"
```

For more information about setting the `--set-crashkernel` boot parameter, see the `kdump(7)` manpage.

3. Push the changes to the desired nodes.

For information about how to push changes, see the following:

[Pushing images from the admin node to the targeted nodes](#) on page 217

## Resetting the `crashkernel` value to the system default

The procedure in this topic explains how to reset the `kdump` value to the system default value.

Complete this procedure to revert to the default settings after either of the following:

- After disabling `kdump`
- After resetting the amount of memory for `crashkernel`

### Procedure

1. Log into the admin node as the root user.
2. (Optional) Display a list of images and display the default `crashkernel` value.

Use the following commands:

```
cimage --show-images  
cadmin --show-crashkernel --image image_name
```

For *image\_name*, specify the one of the images that the `cimage` command displayed.

For example:

```
# cimage --show-images
image: ice-rhel7.4
    kernel: 2.6.32-431.el6.x86_64
# cadmin --show-crashkernel --image ice-rhel7.4
crashkernel=256M
```

3. Use the `cadmin` command in the following format to specify the amount of memory to use for the `kdump` facility:

```
cadmin --set-crashkernel --image image_name
```

For *image\_name*, specify the name of one of the ICE compute node, leader node, or flat compute node operating system images.

Notice that in this format, you do not specify the empty string nor do you specify a string that contains a memory size.

For example:

```
# cadmin --set-crashkernel --image sles12sp3
```

4. Push the changes to the desired nodes.

For information about how to push changes, see the following:

[Pushing images from the admin node to the targeted nodes](#) on page 217

## Retrieving system firmware information

Your cluster system comes preinstalled with the appropriate firmware. For information about updates to the BMC, BIOS, and CMC firmware, see your HPE representative.

Use the following methods to retrieve firmware information:

- To identify the BIOS, you need both the version and the release date. You can get these using the `dmidecode` command. Log onto the node from which you want information, and enter the following command:

```
# dmidecode -s bios-version; dmidecode -s bios-release-date
```

- To retrieve the BMC firmware revision, use the `ipmiwrapper` command. For example, from the admin node, the following command displays the BMC firmware revision for `r1i0n0`:

```
# ipmiwrapper r1i0n0 bmc info | grep 'Firmware Revision'
```

- To retrieve the CMC firmware version, use the `version` command. For example, from the `r1lead` leader node, the following command displays the CMC firmware version:

```
# ssh root@r1i0-cmc version
```

- The `ibstat` command retrieves information for the InfiniBand links and includes the firmware version. The following command displays the InfiniBand firmware version:

```
# ibstat | grep Firmware
```

- The `firmware_revs` script on the admin node displays firmware information for all nodes.

# Booting a flat compute node or a leader node on an installed cluster

The information in this topic might be useful to troubleshoot boot problems. The default boot process is identical for flat compute nodes and leader nodes on an installed cluster. The process assumes the following environment for a node:

- The node is a fully configured node. That is, you ran the `discover` command, and the node became configured into the cluster.
- The node is registered in clusterwide DHCP server database.
- At least one node image is available for this node in the admin node image repository. An image is assigned to the node.
- The GRUB2 network configuration files are available.
- The node has been installed.
- The BIOS hardware on the node is configured for network booting. This state is the factory-defined state.
- You have not configured the node to boot from a local disk.

The following topics describe the boot process:

- [\*\*Phase 1 - Initiating the boot\*\*](#) on page 290
- [\*\*Phase 2 - Loading the kernel for the node\*\*](#) on page 291
- [\*\*Phase 3 - Loading the miniroot\*\*](#) on page 292
- [\*\*Phase 4 - Starting the operating system on the node\*\*](#) on page 293

For information about the boot-from-disk feature, see the following:

[\*\*Booting leaders or flat compute nodes from a local disk\*\*](#) on page 182

---

**NOTE:** The node boot process differs for configured clusters versus unconfigured clusters.

---

## Phase 1 - Initiating the boot

The following events occur when you initiate the boot:

1. From the admin node, a user enters the following command:

```
cpower node on node
```

For example:

```
# cpower node on n0
```

2. The node turns on.
3. The factory-defined boot process starts. The boot proceeds as follows:

- The BIOS boots from the network interface card (NIC).
- The ROM on the NIC activates the PXE protocol to boot over the network.
- The ROM sends a DHCP request for node network information. The node IP address is a static address.

In this case, the admin node receives the DHCP request. If this node is an ICE compute node that is managed by a leader node, the leader node receives the request.

The DHCP configuration files reside in the following directory:

`/etc/dhcp/dhcpd.conf.d/ice.conf`

Include files reside in the `dhcpd.conf.d` directory.

**4.** The admin node responds to the DHCP request. It sends DHCP packets that include the following:

- The network configuration of the node.
- The location of the GRUB2 boot loader.

The loader resides on the admin node in one of the following locations:

- On legacy platforms, this location is as follows:

`/opt/clmgr/repos/boot/grub2/i386-pc/core.0`

- On EFI boot systems, this location is as follows:

`/opt/clmgr/repos/boot/grub2/i386-pc/core.efi`

**5.** The NIC loads GRUB2.

**6.** The admin node directs the node to boot GRUB2.

**7.** The NIC starts.

## Phase 2 - Loading the kernel for the node

In Phase 2, GRUB2 starts up and loads the kernel onto the node. All requests and all transfers are done using TFTP.

The final steps in this subprocess are for GRUB2 to load the following:

- The kernel for the node
- The operating system `initrd` files

These files reside on the admin node in the following directories:

`/opt/clmgr/repos/boot/image_name/kernel`

`/opt/clmgr/repos/boot/image_name/initrd`

The steps are as follows:

**1.** GRUB2 sends a TFTP request to the admin node for a configuration file.

**2.** The admin node responds to GRUB2 by sending the following file:

`/opt/clmgr/repos/boot/grub2/grub.cfg.`

The file includes instructions that explain how to load kernel files and `initrd` files.

**3. The `grub.cfg` file requests the node-specific configuration file.**

The configuration file resides on the admin node in the following location, which includes the node IP address:

```
/opt/clmgr/repos/boot/grub2/tempo/ip_addr.cfg
```

For example, the configuration file might reside in the following file:

```
/opt/clmgr/repos/boot/grub2/tempo/172.23.0.2.cfg
```

The node configuration file includes the following information:

- The kernel and the kernel parameters that the node needs.
- The `initrd` file that the node needs.
- The image assigned to the node. The image itself resides in the following location:  
`/opt/clmgr/image/images/images`
- Instructions for booting the node. The file also includes instructions for installing the node. However, GRUB2 uses these installation instructions only when the node is installed.
- Port number identification information. The miniroot image for the node resides on a specific port. The miniroot images reside in the following directory:

```
/opt/clmgr/image/miniroot/squeezed/images
```

The following is an example `ip_addr.cfg` file:

```
#  
# Boot the SGI miniroot with appropriate options (via the distro initrd)  
# kernel="sles12sp3.24104/vmlinuz-3.0.101-0.47.52-default"  
initrd="sles12sp3.24104/initrd-3.0.101-0.47.52-default"  
append="ROOTFS=disk IMAGE_PENDING=0 IMAGE=sles11sp3.24104 SLOT=1 console=ttyS1,115200n8  
MONITOR_SERVER=172.23.0.1 MONITOR_CONSOLE=yes intel_idle.max_cstate=1 processor.max_cstate=1  
TRANSPORT=rsync TTL=1 MCAST_RDV_ADDR=224.0.0.1 FLAMETHROWER_DIRECTORY_PORTBASE=9000  
START_TIMEOUT=30 RECEIVE_TIMEOUT=5 crashkernel=256M"  
gfxpayload=text  
pxechain=0
```

---

**NOTE:** In the preceding example, lines have been wrapped for inclusion in this documentation.

---

**4. GRUB2 loads the kernel with options.**

**5. The kernel requests the `initrd` file from the operating system distribution.**

## Phase 3 - Loading the miniroot

In Phase 3, GRUB2 loads the miniroot. The steps are as follows:

**1. GRUB2 completes the following actions:**

- It runs the `initrd` file from the operating system distribution on the node.
  - It starts `udpcast` to obtain the miniroot.
  - It runs additional internal scripts associated with the node image.
2. The `udpcast` command transfers the miniroot from the port number assigned to the miniroot to the node.
  3. GRUB2 unpacks the miniroot. The miniroot arrives on the node as a `tar` file.
  4. The miniroot begins operating on the node.

## Phase 4 - Starting the operating system on the node

In Phase 4, the miniroot starts running processes on the node, and the operating system takes over. The steps are as follows:

1. The miniroot finds the root and boot file systems.
2. The miniroot mounts the root and boot file systems.
3. The operating system distribution startup scripts start to run.

## Overriding installation scripts

When creating or updating the miniroot, the cluster manager copies files from the `/opt/clmgr/lib/` directory on the admin node into the miniroot. These files are as follows:

```
/opt/clmgr/lib/miniroot-init
/opt/clmgr/lib/miniroot-functions
/opt/clmgr/lib/miniroot-node-install
/opt/clmgr/lib/miniroot-admin-install
```

The preceding scripts drive node installation and booting.

If a version of any of the following files exists in the `/opt/clmgr/lib` directory, it is used in place of the original:

```
/opt/clmgr/lib/miniroot-init-local
/opt/clmgr/lib/miniroot-functions-local
/opt/clmgr/lib/miniroot-node-install-local
/opt/clmgr/lib/miniroot-admin-install-local
```

For example, if `/opt/clmgr/lib/miniroot-node-install-local` exists, then the cluster manager copies `/opt/clmgr/lib/miniroot-node-install-local` into the miniroot as `/opt/sgi/lib/miniroot-node-install`.

When you create a `-local` file, you override the defaults. Rather than creating `-local` files, HPE recommends that you use the following features:

- Custom partitioning. For information about custom partitioning, see the following:

- [\*\*HPE Performance Cluster Manager Installation Guide\*\*](#)
- [\*\*Creating custom partitions\*\*](#) on page 196
- Disk reservations. This feature enables you to reserve space at the end of the disk for a scratch space. For information, see the following:  
[\*\*Configuring scratch disk space on system disks\*\*](#) on page 140
- System imager pre- and post-installation scripts.

---

**⚠ CAUTION:** HPE does not recommend that you override the default files. HPE updates these files with features and fixes in both patches and in releases. The updated content is not reflected in locally managed copies. If you use this feature, make sure to update the code in your version to match the versions from the cluster manager. To avoid customized versions becoming outdated or incompatible between releases, perform merges. Patches do not touch the customized files, and patches might fix important bugs.

The following `cinstallman` operations update the miniroot, including copying the files:

```
cinstallman --update-miniroot  
cinstallman --zypper-image      # as long as --duk not present  
cinstallman --yum-image        # as long as --duk not present
```

---

# Security features

The Linux distributions include security features. The cluster manager provides additional security features, and among those features are the following:

- Secure provisioning over the management network
- Restricted node-to-node login access
- No root user `ssh` files in default images
- Secure environment for the cluster manager database

To provide these features, the cluster manager creates encryption passwords, certificates, `ssh` keys, and other security constructs. The cluster manager documentation refers to these items collectively as **secrets**. The secrets that the cluster manager uses to build the node security infrastructure are called **bootstrap secrets**.

The following topics describe secrets, the security features, and how you can enhance the security of your cluster:

- [\*\*Secret creation\*\*](#) on page 295
- [\*\*Secure provisioning\*\*](#) on page 296
- [\*\*Restricted node-to-node login access\*\*](#) on page 297
- [\*\*Cluster manager database security\*\*](#) on page 299

## Secret creation

The cluster manager creates security secrets when the cluster manager is installed and configured. The installation step that creates security secrets is the `configure-cluster` step.

Using the newly created secrets and bootstrap secrets, the installation process creates the necessary security framework on the admin node. The installer creates the complementary security infrastructure on the other nodes as they are installed.

The following topics explain how the cluster manager creates secrets:

- [\*\*Packaging and file residence\*\*](#) on page 295
- [\*\*Recreating secrets\*\*](#) on page 296

## Packaging and file residence

The cluster manager manages the secrets files. You do not need to manually manage the files.

The secrets files reside in the following directory on the admin node:

`/opt/sgi/secrets`

The bootstrap secrets reside in compressed, encrypted files in the following directory:

`/opt/sgi/secrets/bootstrap-secrets`

The following command output shows the two secrets files:

```
[root@myadmin ~]# ls /opt/sgi/secrets/bootstrap-secrets
compute.tar.xz.aes    leader.tar.xz.aes
```

File `compute.tar.xz.aes` is the secrets file for flat compute nodes and ICE compute nodes. File `leader.tar.xz.aes` is the secrets file for leader nodes.

For added security, the encrypted bootstrap secrets can be transferred to a node only if the node is marked for installation.

## Recreating secrets

For added security, you can recreate the set of secrets. Doing so, however, requires that you reinstall the cluster nodes afterwards, except for the admin node.

The following procedure explains how to recreate secrets:

### Procedure

1. Run the following script:

```
/opt/sgi/lib/create-secrets
```

2. Restart the `smc-adminmd` service.

3. Mark all leader and flat compute nodes for reinstallation.

Example 1. On flat clusters, run the following command:

```
# cinstallman --next-boot image --node "n*"
```

Example 2. On hierarchical clusters, run the following command:

```
# cinstallman --next-boot image --node "service*","r*lead"
```

You do not need to mark all the ICE compute nodes for reinstallation. The cluster manager reinstalls them automatically after the leader nodes reboot.

4. Reboot all the nodes.

Example 1. For flat clusters, enter the following command:

```
# cpower node reboot "n*"
```

Example 2. On hierarchical clusters, enter the following command:

```
# cpower node reboot "service*","r*lead"
```

## Secure provisioning

The following topics explain secure provisioning:

- [\*\*Safeguards against unauthorized requests\*\*](#) on page 296
- [\*\*Image encryption and authentication\*\*](#) on page 297

## Safeguards against unauthorized requests

To guard against unauthorized requests for images, the cluster manager does a series of checks. The following are some of the checks:

- The cluster manager ensures that the request is coming from a node that has been marked for provisioning. The `cinstallman` command with the `next-boot` option marks a node for provisioning.

If the node has a disk and is not marked for provisioning, the cluster manager does not send the image.

- The password that decrypts the bootstrap secrets is transferred. As part of this activity, the cluster manager sends the image only to a root-specific path on the node being installed.

## Image encryption and authentication

By default, the cluster manager uses UDPcast to transfer images from the admin node to the other nodes. The cluster manager implements secure provisioning when UDPcast is used. During provisioning, the cluster manager authenticates, encrypts, and decrypts the images at the appropriate transfer points. The bootstrap secrets provide the resources for authentication, encryption, and decryption.

In some cases, factors such as node type and performance might cause you to choose another transport method. If you choose to transport images from the admin node to other nodes using BitTorrent, the cluster manager does not authenticate or encrypt the images. The following table is a support matrix of image transfer methods. The table shows whether secure provisioning is provided:

| Transfer method | Secure mode provided? |
|-----------------|-----------------------|
| rsync           | Yes                   |
| BitTorrent      | No                    |
| UDPcast         | Yes                   |

For information about how to choose the appropriate transfer method for your site, see the following:

### [HPE Performance Cluster Manager Installation Guide](#)

## Restricted node-to-node login access

The action of logging into a node without a password is called a **passive login**.

The cluster manager imposes security constraints on the root user. The cluster manager constrains the root user because many cluster manager interfaces run under the auspices of the root user, passively logging into nodes.

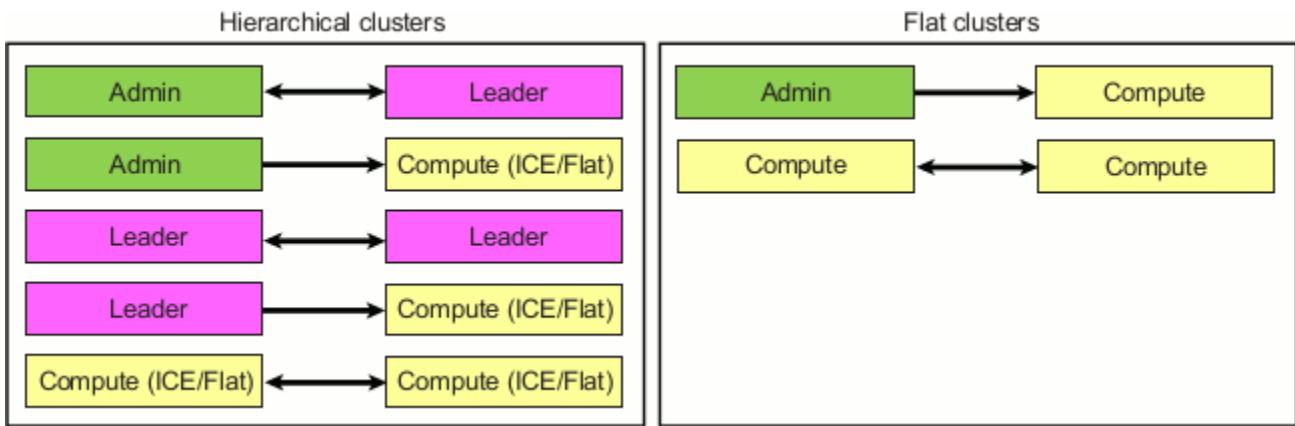
The cluster manager allows the root user to passively log into any node from the admin node. It permits other passive logins. By default, the root user is allowed to passively log in as follows:

- From the admin node to a leader node
- From the admin node to an ICE compute node
- From the admin node to a flat compute node
- From a leader node to another leader node
- From a flat compute node to another flat compute node

The cluster manager prevents the root user from passively logging into some nodes from other nodes. By default, the root user is not allowed to passively log in as follows:

- From a flat compute node to a leader node
- From an ICE compute node to a leader node
- From a flat compute node to the admin node

The following figure illustrates the login flows allowed for the root user.



**Figure 57: Root user passive logins**

The following topics contain more information about login access:

- [Cluster manager ssh zones](#) on page 298
- [Default images and customizing your ssh configuration](#) on page 299
- [Security recommendations](#) on page 299

## Cluster manager ssh zones

The cluster manager uses the `ssh` security scheme (RSA 2 private and public keys) to restrict the node-to-node login access for the root user.

The root-user-authorized key files on the admin and leader nodes are configured such that root users from flat compute nodes or ICE compute nodes cannot passively log in. The root-user-authorized key files for the compute nodes contain the public keys for the admin nodes, leader nodes, and other compute nodes. The root-user-authorized key files do not contain public keys for the root users from the compute nodes.

To document this login flow in the authorized key files, the cluster manager does the following:

- The cluster manager uses the notion of zones.
- The cluster manager labels each public key it places in the file according to the login flow it allows. The cluster manager places the zone label as a comment at the end of the key.

For example, the following command returns the labels and zones for the public keys:

```
r1lead:~/ssh # cat authorized_keys
```

The following table summarizes the zones that the cluster manager places in the authorized key file of the root user:

| Node type                                 | Zones in authorized key file for the root user                |
|-------------------------------------------|---------------------------------------------------------------|
| Admin                                     | Admin node, leader node                                       |
| Leader                                    | Admin node, leader node                                       |
| ICE compute nodes and flat computes nodes | Admin node, leader node, ICE compute nodes, flat compute node |

With such zones in the authorized key file, you can quickly ascertain the types of hosts that are allowed passive root access.

For information about how the cluster manager restricts node-to-node login access for the root user, see the following:

[Restricted node-to-node login access](#) on page 297

## Default images and customizing your ssh configuration

For added security, the default images on the admin node do not contain ssh keys. When the cluster manager installs a node, it populates the ssh files for the root user of that node.

After a node is installed, you can add your own root user ssh keys and configuration to the image on the admin node. Then, when you install that image, the cluster manager uses your ssh keys and configuration instead of its own.

 **CAUTION:** If you modify the ssh configuration for the root user on the admin or a leader node, do not disrupt the login flows for the root user on those nodes. For more information about logging into nodes, see the following:

[Restricted node-to-node login access](#) on page 297

---

The following topic explains how the cluster manager creates the ssh keys (RSA 2) and other bootstrap secrets:

[Secret creation](#) on page 295

## Security recommendations

To maintain a secure cluster, you must go beyond the protections provided by the ssh zoning feature for the root user. Your site must restrict the running of user code and user jobs on admin and leader nodes. The cluster manager does not prevent such jobs and processes from running. Your site must assess the risks and benefits.

## Cluster manager database security

The primary data repository for the cluster manager resides in a relational database. Many cluster management functions read from and write to this database. For increased security, the cluster manager provides the following database safeguards:

- Centralized database

The database resides only on the admin node with no replication to other cluster nodes. There is only one password to protect. All client requests are ultimately served through the daemons on the admin node. On leader nodes, a daemon acts as a proxy for such requests.

- No database network interfaces

The cluster manager shields the interfaces that read from and write to this database. The cluster manager interfaces use only UNIX sockets for database access, no network interfaces.

- Secure connections for database interfaces

The cluster manager encrypts the requests and data used by the cluster daemons to access the database. Further, it provides two levels of database access: one for the root user and one for other users. The nonroot users have highly restricted access to database content.

# Using Singularity containers

A Singularity container is a portable and self-contained computing environment.

The following topics explain how to create and run Singularity containers:

- [Installing the container software](#) on page 300
- [Building a container](#) on page 301
- [Running containers](#) on page 304

## Installing the container software

### Procedure

1. Download and compile Singularity RPMs from source or obtain pre-build RPMs from your operating system distribution.

For example, you can obtain downloads from the following website:

<http://singularity.lbl.gov/install-linux>

2. Create and copy the RPMs to the following directory on the admin node:

/opt/clmgr/repos/other/singularity-2.X

3. Use the `crepo` command to add the RPMs to the repository.

For example:

```
[ admin ~]# mkdir -p /opt/clmgr/repos/other/singularity-2.X
[ admin ~]# cp singularity*.rpm /opt/clmgr/repos/other/singularity-2.X
[ admin ~]# crepo --add /opt/clmgr/repos/other/singularity-2.X \
[ admin ~]# --custom singularity-2.X
Creating rpm-md metadata for repo...
Creating repodata cache for /opt/clmgr/repos/other/singularity-2.X
[ admin ~]# crepo --select "singularity-2.X"
Selecting: singularity-2.X
```

The preceding example creates and selects the Singularity 2.X repository.

4. Install the Singularity RPMs into the ICE compute node image and onto the flat compute node that you want to use as a container build server.

Use the `cinstallman` command to install the RPMs into the ICE compute node image. For example:

```
[ admin ~]# crepo --show |sort
* CentOS-7.4 : /opt/clmgr/repos/distro/centos7.4
* HPE-MPI-1.2-rhel73 : /opt/clmgr/repos/cm/HPE-MPI-1.2-rhel73
  MOFED-4.0-2.0.0.1-rhel73 : /opt/clmgr/repos/other/MLNX_OFED_LINUX-4.0-2.0.0.1-rhel7.4-x86_64
* patch11431-centos73 : /opt/clmgr/repos/other/patch11431/centos73/
* patch11432-centos73 : /opt/clmgr/repos/other/patch11432/centos73/
* patch11433-centos73 : /opt/clmgr/repos/other/patch11433/centos73/
* Cluster-Manager-1.0.0-rhel73 : /opt/clmgr/repos/cm/Cluster-Manager-1.0.0-rhel73
* singularity-2.4 : /opt/clmgr/repos/other/singularity-2.4
* slurm-17.02.8 : /opt/clmgr/repos/other/slurm-17.02.8
  smee-11.20 : /opt/clmgr/repos/other/smee-11.20/
[ admin ~]# cinstallman --yum-image --image ice-centos7.4 install \
[ admin ~]# singularity
Resolving Dependencies
```

```
--> Running transaction check
--> Package singularity.x86_64 0:2.4-1 will be installed
--> Processing Dependency: singularity-runtime = 2.4-1 for package: singularity-2.4-1.x86_64
--> Processing Dependency: squashfs-tools for package: singularity-2.4-1.x86_64
--> Processing Dependency: libssingularity-runtime.so.1() (64bit) for package: singularity-2.4-1.x86_64
--> Processing Dependency: libssingularity-image.so.1() (64bit) for package: singularity-2.4-1.x86_64
--> Running transaction check
--> Package singularity-runtime.x86_64 0:2.4-1 will be installed
--> Package squashfs-tools.x86_64 0:4.3-0.21.gitaae0aff4.el7 will be installed
--> Finished Dependency Resolution
```

5. Push the modified image to the ICE compute racks, power off the nodes, push the image, and boot the nodes to the new image.

For example:

```
[ admin ~]# cpower node off r1i*n*
[ admin ~]# cimage --push-rack ice-centos7.4 r1
[ admin ~]# cpower node on r1i*n*
```

6. Use the `cinstallman` command to install the image on a service node.

For example:

```
[ admin ~]# cinstallman --yum-node --node service0 install singularity
Resolving Dependencies
--> Running transaction check
--> Package singularity.x86_64 0:2.4-1 will be installed
--> Processing Dependency: singularity-runtime = 2.4-1 for package: singularity-2.4-1.x86_64
--> Processing Dependency: squashfs-tools for package: singularity-2.4-1.x86_64
--> Processing Dependency: libssingularity-runtime.so.1() (64bit) for package: singularity-2.4-1.x86_64
--> Processing Dependency: libssingularity-image.so.1() (64bit) for package: singularity-2.4-1.x86_64
--> Running transaction check
--> Package singularity-runtime.x86_64 0:2.4-1 will be installed
--> Package squashfs-tools.x86_64 0:4.3-0.21.gitaae0aff4.el7 will be installed
--> Finished Dependency Resolution
```

7. Add an option to the following file on the service node to facilitate the command execution of the Singularity binary:

`/etc/sudoers`

---

**NOTE:** Exercise caution. Adding to `/etc/sudoers` gives complete root access to the system. The Singularity build node should be expected as compromised. Do not store sensitive data here. Delete cluster root private keys on this node.

---

For example:

```
[ service0 ~]# cat /etc/sudoers | grep dmk
dmk ALL=(ALL) NOPASSWD: /usr/bin/singularity
```

Without this edit, only the root user can create containers; users need to import already created containers to the system themselves. For a large amount of users, create a `singularity` group. The example in this step adds a single line at the end of the `/etc/sudoers` file on node `service0` node for user `dmk`.

8. Proceed to the following:

[Building a container](#) on page 301

## Building a container

The procedure in this topic creates a custom repository file for the container to use. If you do not use a custom repository file for the container, the container downloads source files from the Internet. The procedure explains how to direct the container to use the installation source files from the admin node.

When you run a container on an InfiniBand cluster, make sure that the same OFED version resides in both the container and the environment. Otherwise, the container cannot use the full InfiniBand cluster.

The following websites provide more information about building containers:

<http://singularity.lbl.gov/docs-build-container>

<http://singularity.lbl.gov/docs-recipes>

Creating a container that works can involve trial and error.

## Procedure

### 1. Create two text files in your home directory.

You can create the container files in your user home directory.

Subsequent steps in this procedure explain what to include in these files. These files have the following purposes:

- The first file is for the container recipe. For example, you can name this file *distroname.recipe*. You can copy the text from the example repository file in this procedure and modify it to reflect your environment. For example, change *MirrorURL* to your distribution name.
- The second file is for a custom repository file that can pull RPMs from the admin node. For example, you can name this file *cluster.repo*. The required file suffix is *.repo*.

If necessary, enter the following command on the admin node to display all repository locations:

```
crepo -show
```

### 2. Edit the container recipe file.

At a minimum, prepend `admin/repo` to `/opt/clmgr/repos/`, and supply text for the *MirrorURL* string.

In the following example recipe file, notice the following:

- In the `%setup` section, there are two lines. The first line deletes the repositories from the container. The second line copies the custom repository file.
- The `%post` section installs the HPE Message Passing Interface (MPI), OFED (InfiniBand), and Slurm RPMs.
- The `%environment` section includes the variables required for proper execution of HPE Message Passing Interface (MPI) and includes the Slurm variable of `pmi2`. These variables are specific to the message passing interface installation and to Slurm.

```
[ dmk@service0 ~]$ cat centos7.recipe
Bootstrap: yum
OSVersion: 7
MirrorURL: http://admin/repo/tftpboot/distro/centos7.4
Include: yum

%setup
    rm ${SINGULARITY_ROOTFS}/etc/yum.repos.d/*
    cp smc-yum.repo ${SINGULARITY_ROOTFS}/etc/yum.repos.d
%post
```

```

yum -y install vim sgi-mpt libmlx4 libmlx5 librdmacm slurm

%environment
  CPATH="/opt/hpe/hpc/mpt/mpt-2.18/include"
  FPATH="/opt/hpe/hpc/mpt/mpt-2.18/include"
  LIBRARY_PATH="/opt/hpe/hpc/mpt/mpt-2.18/lib"
  LD_LIBRARY_PATH="/opt/hpe/hpc/mpt/mpt-2.18/lib:$LD_LIBRARY_PATH"
  PATH="/opt/hpe/hpc/mpt/mpt-2.18/bin:$PATH"
  MPI_ROOT="/opt/hpe/hpc/mpt/mpt-2.18"
  MPI_VERBOSE="1"
  MPT_VERSION="2.18"
  SLURM_MPI_TYPE="pmi2"

```

### 3. Edit the custom repository file.

The following `smc-yum.repo` file points to the admin node to gather the RPMs. Remember, if you do not use a repository file, the container downloads RPMs from the Internet. If you have a specific version of OFED that is different from the version from your software distribution, include the repository for that specific version. The example in this step shows how to include the repository for Mellanox OFED.

```

[ drmk@service0 ~ ]$ cat smc-yum.repo
[main]
cachedir=/var/cache/yum/$basearch/$releasever
keepcache=0
logfile=/var/log/yum.log
exactarch=1
obsoletes=1
gpgcheck=0
plugins=0
distroverpkg=centos-release

[Base]
name=CentOS73
baseurl=http://admin/repo/tftpboot/distro/centos7.3
enable=1
gpgcheck=0

[HPE-MPI-1.2]
name=HPE-MPI-1.2
baseurl=http://admin/repo/tftpboot/sgi/HPE-MPI-1.2-rhel73/RPMS
enabled=1
gpgcheck=0

[MLNX_OFED]
name=MLNX-OFED
baseurl=http://admin/repo/tftpboot/other/MLNX_OFED_LINUX-4.0-2.0.0.1-rhel7.3-x86_64/RPMS
enabled=1
gpgcheck=0

[Slurm]
name=slurm
baseurl=http://admin/repo/tftpboot/other/slurm-17.02.8
enabled=1
gpgcheck=0

```

### 4. Use the `ls` command to verify that the container files reside in your home directory.

For example:

```
[ drmk@service0 ~ ]$ ls
.
```

```
 .
 .
centos7.recipe
smc-yum.repo
```

## 5. Build the container.

For example:

```
[ dmk@service0 ~]$ sudo singularity build centos74-container centos7.recipe
Using container recipe deffile: centos7.recipe
Sanitizing environment
Adding base Singularity environment to container
Found YUM at: /bin/yum
Skipping GPG key import.
base
| 3.6
kB 00:00:00
(1/2): base/
group_gz
| 155 kB 00:00:00
(2/2): base/
primary_db
| 5.6 MB 00:00:00
Resolving Dependencies

Complete!
Finalizing Singularity container
Calculating final size for metadata...
Skipping checks
Building Singularity image...
Singularity container built: centos74-container
Cleaning up...
```

If the command in this step does not succeed, check the following:

- Was the `sudo` configured properly?
- Does the repository file point to the admin node correctly?

If you want to be able to modify the container after it is built, include the `--writeable` parameter on the `sudo` command line. If you include the `--writeable` parameter, the resulting container is at least four times larger.

# Running containers

Example 1. This example runs `hello_world.mpt`, which is included in the field diagnostics executed through four nodes using Slurm. There is no need to use `sudo` to execute a Singularity container. The commands are as follows:

```
[ dmk@service0 ~]$ rpm -qf /usr/diags/bin/hello_world.mpt
field_diags_licensed_ice_x86-3.25-146.x86_64
[ dmk@service0 ~]$ cp /usr/diags/bin/hello_world.mpt ~/. 
[ dmk@service0 ~]$ srun -N 4 singularity exec centos73-container \
[ dmk@service0 ~]$ /home/dmk/hello_world.mpt
MPT: libmpi_mt.so 'HPE MPT 2.18 05/03/18 01:21:48'
      MPT Environmental Settings
MPT: MPI_VERBOSE (default: disabled) : enabled
```

```
MPT: Using the InfiniBand RC interconnect on fabric 0x0
MPT: Using the InfiniBand UD interconnect on fabric 0x0
Hello world from process 0 of 4
Hello world from process 1 of 4
Hello world from process 2 of 4
Hello world from process 3 of 4
```

Example 2. The following command runs a container using the `mpirun` command on three nodes with two cores on each node:

```
[ dmk@service0 ~]$ mpirun r1i0n0,r1i0n1,r1i0n2 -np 2 singularity exec \
[ dmk@service0 ~]$ CentOS7-v23 /home/dmk/hello_world.mpt
Hello world from process 0 of 6
Hello world from process 4 of 6
Hello world from process 2 of 6
Hello world from process 1 of 6
Hello world from process 5 of 6
Hello world from process 3 of 6
```

# Hierarchical cluster system configuration framework information

You can use the configuration framework to adjust the settings of nodes in a hierarchical cluster. There is some overlap between the per-host customization instructions and the configuration framework instructions. Each approach plays a role in configuring your system. The major differences between the two methods are as follows:

- Per-host customization runs at one of the following times:
  - When an admin node pushes an image to the leader nodes
  - Upon demand. Use the `--customizations-only` option to the `cimage` command.
- Per-host customization applies only to ICE compute node images.
- The hierarchical cluster configuration framework information can be used with all node types.

The configuration framework exists to make it easy to adjust configuration items. The cluster manager installs some scripts as part of the cluster manager software. You can add scripts as needed. If you want to stop a script from running, you can exclude the script from running without purging the script.

The following topics describe the hierarchical cluster system configuration framework:

- [\*\*About the hierarchical cluster system configuration framework\*\*](#) on page 306
- [\*\*About the cluster configuration repository\*\*](#) on page 308

## About the hierarchical cluster system configuration framework

This topic contains an FAQ list that addresses the system configuration framework.

### How does the system configuration framework operate?

Files can be added to a running compute node, or to an already created service or compute image, as follows:

- A `/opt/clmgr/lib/cluster-configuration` script is called, from where it is called is described below.
- That script iterates through scripts residing in `/etc/opt/sgi/conf.d`.
- Any scripts listed in `/etc/opt/sgi/conf.d/exclude` are skipped, as are scripts, that are not executable.
- Scripts in system configuration framework **must** be tolerant of files that do not exist yet, as described below. For example, check that a `syslog` configuration file exists before trying to adjust it.
- Scripts ending in a distro name, or a distro name with a specific distro version are run only if the node in question is running that distro. For example, `/etc/opt/sgi/conf.d/99-foo.sles` runs only when the node is running `sles`. This example shows the order of operations.

If you had `88-myscript.sles11`, `88-myscript.sles`, and `88-myscript`:

- On a `sles11` system, `88-myscript.sles11` runs.
  - On a `sles` system that is not `sles11`, `88-myscript.sles` runs.
  - On all other distros, `88-myscript` runs.
- If you want to make a custom version of a script supplied by HPE, change the script's suffix from the distro name to `.local`. The `.local` suffix indicates that you want the local version to run in place of the one supplied by HPE. This naming convention allows you to customize the scripts provided by HPE while preserving the original, default script supplied by HPE. Scripts that end in `.local` have the highest precedence. In other words, if you had `88-myscript.sles` and `88-myscript.local`, then `88-myscript.local` runs in all cases and any other `88-myscript.suffix` scripts never run.

Images destined for ICE compute nodes need to be pushed with the `cimage` command after being altered. For more information, see the following:

#### **Using the cimage command to manage ICE compute node images** on page 225

#### **From where is the framework called?**

- The callout for `/opt/clmgr/lib/cluster-configuration` is implemented as a `yum` plugin that executes after packages have been installed and cleaned.
- On SLES only, there is also a configuration script in the `/sbin/conf.d` directory, called `SuSEconfig.00cluster-configuration`, that calls the framework. This is in case of you are using YaST to install or upgrade packages.
- On SLES only, one of the scripts called by the framework calls `SuSEconfig`. A check is made to avoid a callout loop.

#### **When is the framework called?**

- The framework is called when an image is created.
- The framework is also called when the admin node, leader node, or compute nodes start up. The call is made just after networking is configured. As a site administrator, you could create custom scripts here that check on or perform certain configuration operations.
- When using the `cimage` command to push an ICE compute node root image to leader nodes, the configuration framework executes within the `chroot` of the ICE compute node image after it is pulled from the admin node to the leader node.
- The framework is called when a compute node or a leader node is installed.

#### **How do I adjust my system configuration?**

Create a small script in `/etc/opt/sgi/conf.d` to do the adjustment.

Be sure that you test for existence of files and do not assume they are there. Also see **Why do scripts need to tolerate files that do not exist but should?** in this topic.

#### **Why do scripts need to tolerate files that do not exist but should?**

This is because the `mksiimage` command runs `yume` and `yum` in two steps. The first step only installs 40 or so RPMs but the framework is called then, too. The second pass installs the other hundreds of RPMs. The framework is called for the first time before some packages are installed, and the framework is called

again after everything is in place. So, not all files you expect might be available when your small script is called.

### How does the yum plugin work?

In order for the `yum` plugin to work, the `/etc/yum.conf` file has to have `plugins=1` set in its configuration file. The `sgi-cluster` package ensures that this setting is correct. Anytime `yum` is installed or updated, it verifies that `plugins=1` is set.

### How does yume work?

`yume`, an oscar wrapper for `yum`, works by creating a temporary `yum` configuration file in `/tmp` and then points `yum` at it. This temporary configuration file needs to have plugins enabled. A tiny patch to `yume` makes this happen. This fixes it for `yume` and also `mksiimage`, which calls `yume` as part of its operation.

## About the cluster configuration repository

A hierarchical cluster system includes a cluster configuration repository/update framework. This framework generates and distributes configuration updates to admin node, leader node, and compute nodes in the cluster. Some of the configuration files managed by this framework include `conserver`, `DNS`, `Ganglia`, `hosts` files, and `NTP`.

The following topics contain more information about the cluster configuration repository:

- [Automatic updates](#) on page 308
- [Custom configuration scripts](#) on page 308
- [Preserving custom configuration changes](#) on page 309

### Automatic updates

You can use the `cattr` and `cadmin` commands to change system attributes. Unlike the `cattr` command, the `cadmin` command regenerates the configuration and eliminates the need for you to issue an `update-configs` command to effect the change.

When an event occurs that requires these files to be updated, the framework executes on the admin node. The admin node stores the updated configuration framework in a special cached location and updates the appropriate nodes with their new configuration files.

In addition to the updates happening as required, the configuration file repository is consulted when an admin node, leader node, ICE compute node, or flat compute node boots. This happens shortly after networking is started. Any configuration files that are new or updated are transferred at this early stage so that the node is fully configured by the time the node is fully operational.

### Custom configuration scripts

This update framework is tied in with the `/etc/opt/sgi/conf.d` configuration framework to provide a full configuration solution. You can create configuration scripts and place them in the `/etc/opt/sgi/conf.d` directory of a node or an image on the admin node.

If you want to suppress the running one of the scripts that live in `/etc/opt/sgi/conf.d`, put the name of the script in `/etc/opt/sgi/conf.d/exclude`. By default, the content of the file is two comments, as shown in the following:

```
# In this file, list the names of any scripts here that you wish to skip.  
# That will have the effect of the associated customization not being made.
```

For example, if you do not want the 80-increase-arp-cache-sizes script to run, put that script name in file /etc/opt/sgi/conf.d/exclude, as shown in the following:

```
# In this file, list the names of any scripts here that you wish to skip.  
# That will have the effect of the associated customization not being made.  
80-increase-arp-cache-sizes
```

## Preserving custom configuration changes

To prevent the update-configs command from overwriting configuration files you want to preserve, you can protect those files by adding the desired files or directories to file /etc/opt/sgi/conf.d/exclude-update-configs on the node itself or in an image on the admin node. The format of the file is one filename or directory name per line. Commented lines are ignored. The following is the default content of the file:

```
#please enter your directories or files that you don't want to be overridden.  
#Entries should be in below format without the comment.  
#see rsync man in section --exclude= PATTERN  
#/etc/ganglia  
#/etc/ganglia/gmond.conf
```

# YaST navigation

The following table shows SLES YaST navigation key sequences.

| Key                        | Action                                                                                                                                                     |
|----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Tab</b>                 | Moves you from label to label or from list to list.                                                                                                        |
| <b>Alt + Tab</b>           |                                                                                                                                                            |
| <b>Esc + Tab</b>           |                                                                                                                                                            |
| <b>Shift + Tab</b>         |                                                                                                                                                            |
| <b>Ctrl + L</b>            | Refreshes the screen.                                                                                                                                      |
| <b>Enter</b>               | Starts a module from a selected category, runs an action, or activates a menu item.                                                                        |
| <b>Up arrow</b>            | Changes the category. Selects the next category up.                                                                                                        |
| <b>Down arrow</b>          | Changes the category. Selects the next category down.                                                                                                      |
| <b>Right arrow</b>         | Starts a module from the selected category.                                                                                                                |
| <b>Shift + right arrow</b> | Scrolls horizontally to the right. Useful in screens if use of the <b>left arrow</b> key would otherwise change the active pane or current selection list. |
| <b>Ctrl + A</b>            |                                                                                                                                                            |
| <b>Alt + <i>letter</i></b> | Selects the label or action that begins with the <i>letter</i> you select. Labels and selected fields in the display contain a highlighted <i>letter</i> . |
| <b>Esc + <i>letter</i></b> |                                                                                                                                                            |
| <b>Exit</b>                | Quits the YaST interface.                                                                                                                                  |

# Support and other resources

## Accessing Hewlett Packard Enterprise Support

- For live assistance, go to the Contact Hewlett Packard Enterprise Worldwide website:  
**<http://www.hpe.com/assistance>**
- To access documentation and support services, go to the Hewlett Packard Enterprise Support Center website:  
**<http://www.hpe.com/support/hpesc>**

### Information to collect

- Technical support registration number (if applicable)
- Product name, model or version, and serial number
- Operating system name and version
- Firmware version
- Error messages
- Product-specific reports and logs
- Add-on products or components
- Third-party products or components

## Accessing updates

- Some software products provide a mechanism for accessing software updates through the product interface. Review your product documentation to identify the recommended software update method.
- To download product updates:

### Hewlett Packard Enterprise Support Center

**[www.hpe.com/support/hpesc](http://www.hpe.com/support/hpesc)**

### Hewlett Packard Enterprise Support Center: Software downloads

**[www.hpe.com/support/downloads](http://www.hpe.com/support/downloads)**

### Software Depot

**[www.hpe.com/support/softwaredepot](http://www.hpe.com/support/softwaredepot)**

- To subscribe to eNewsletters and alerts:

**[www.hpe.com/support/e-updates](http://www.hpe.com/support/e-updates)**

- To view and update your entitlements, and to link your contracts and warranties with your profile, go to the Hewlett Packard Enterprise Support Center **More Information on Access to Support Materials** page:

- 
- (!) **IMPORTANT:** Access to some updates might require product entitlement when accessed through the Hewlett Packard Enterprise Support Center. You must have an HPE Passport set up with relevant entitlements.
- 

## Customer self repair

Hewlett Packard Enterprise customer self repair (CSR) programs allow you to repair your product. If a CSR part needs to be replaced, it will be shipped directly to you so that you can install it at your convenience. Some parts do not qualify for CSR. Your Hewlett Packard Enterprise authorized service provider will determine whether a repair can be accomplished by CSR.

For more information about CSR, contact your local service provider or go to the CSR website:

<http://www.hpe.com/support/selfrepair>

## Remote support

Remote support is available with supported devices as part of your warranty or contractual support agreement. It provides intelligent event diagnosis, and automatic, secure submission of hardware event notifications to Hewlett Packard Enterprise, which will initiate a fast and accurate resolution based on your product's service level. Hewlett Packard Enterprise strongly recommends that you register your device for remote support.

If your product includes additional remote support details, use search to locate that information.

### Remote support and Proactive Care information

#### HPE Get Connected

[www.hpe.com/services/getconnected](http://www.hpe.com/services/getconnected)

#### HPE Proactive Care services

[www.hpe.com/services/proactivecare](http://www.hpe.com/services/proactivecare)

#### HPE Proactive Care service: Supported products list

[www.hpe.com/services/proactivecaresupportedproducts](http://www.hpe.com/services/proactivecaresupportedproducts)

#### HPE Proactive Care advanced service: Supported products list

[www.hpe.com/services/proactivecareadvancedsupportedproducts](http://www.hpe.com/services/proactivecareadvancedsupportedproducts)

### Proactive Care customer information

#### Proactive Care central

[www.hpe.com/services/proactivecarecentral](http://www.hpe.com/services/proactivecarecentral)

#### Proactive Care service activation

[www.hpe.com/services/proactivecarecentralgetstarted](http://www.hpe.com/services/proactivecarecentralgetstarted)

## Warranty information

To view the warranty for your product or to view the *Safety and Compliance Information for Server, Storage, Power, Networking, and Rack Products* reference document, go to the Enterprise Safety and Compliance website:

[www.hpe.com/support/Safety-Compliance-EnterpriseProducts](http://www.hpe.com/support/Safety-Compliance-EnterpriseProducts)

**Additional warranty information**

**HPE ProLiant and x86 Servers and Options**

[www.hpe.com/support/ProLiantServers-Warranties](http://www.hpe.com/support/ProLiantServers-Warranties)

**HPE Enterprise Servers**

[www.hpe.com/support/EnterpriseServers-Warranties](http://www.hpe.com/support/EnterpriseServers-Warranties)

**HPE Storage Products**

[www.hpe.com/support/Storage-Warranties](http://www.hpe.com/support/Storage-Warranties)

**HPE Networking Products**

[www.hpe.com/support/Networking-Warranties](http://www.hpe.com/support/Networking-Warranties)

## Regulatory information

To view the regulatory information for your product, view the *Safety and Compliance Information for Server, Storage, Power, Networking, and Rack Products*, available at the Hewlett Packard Enterprise Support Center:

[www.hpe.com/support/Safety-Compliance-EnterpriseProducts](http://www.hpe.com/support/Safety-Compliance-EnterpriseProducts)

### Additional regulatory information

Hewlett Packard Enterprise is committed to providing our customers with information about the chemical substances in our products as needed to comply with legal requirements such as REACH (Regulation EC No 1907/2006 of the European Parliament and the Council). A chemical information report for this product can be found at:

[www.hpe.com/info/reach](http://www.hpe.com/info/reach)

For Hewlett Packard Enterprise product environmental and safety information and compliance data, including RoHS and REACH, see:

[www.hpe.com/info/ecodata](http://www.hpe.com/info/ecodata)

For Hewlett Packard Enterprise environmental information, including company programs, product recycling, and energy efficiency, see:

[www.hpe.com/info/environment](http://www.hpe.com/info/environment)

## Documentation feedback

Hewlett Packard Enterprise is committed to providing documentation that meets your needs. To help us improve the documentation, send any errors, suggestions, or comments to Documentation Feedback ([docsfeedback@hpe.com](mailto:docsfeedback@hpe.com)). When submitting your feedback, include the document title, part number, edition, and publication date located on the front cover of the document. For online help content, include the product name, product version, help edition, and publication date located on the legal notices page.

# Websites

## General websites

Hewlett Packard Enterprise Information Library

[www.hpe.com/info/EIL](http://www.hpe.com/info/EIL)

Single Point of Connectivity Knowledge (SPOCK) Storage compatibility matrix

[www.hpe.com/storage/spock](http://www.hpe.com/storage/spock)

Storage white papers and analyst reports

[www.hpe.com/storage/whitepapers](http://www.hpe.com/storage/whitepapers)

For additional websites, see [Support and other resources](#).